Econometrics IV

Štěpán Jurajda

First Draft 1997

April 8, 2010

Contents

Preamble

These lecture notes were originally written (before the Wooldridge textbook became available) for a 2nd-year Ph.D. course in cross-sectional econometrics. The goal of the course is to introduce tools necessary to understand and implement empirical studies in economics focusing on other than time-series issues. The main emphasis of the course is twofold: (i) to extend regression models in the context of panel data analysis, (ii) to focus on situations where linear regression models are not appropriate and to study alternative methods. Due to time constraints, I am not able to cover dynamic panel data models. Examples from applied work will be used to illustrate the discussed methods. Note that the course covers much of the work of the Nobel prize laureates for 2000. The main **reference textbook** for the course is *Econometric Analysis of Cross Section and Panel Data*, [W], Jeffrey M. Wooldridge, MIT Press 2002. Other useful references are:

1. *Econometric Analysis*, [G], William H. Greene. 5th edition, Prentice Hall, 2005.

2. *Analysis of Panel Data*, [H], Cheng Hsiao, Cambridge U. Press, 1986.

3. *Limited-dependent and Qualitative Variables in Econometrics*, [M], G.S. Maddala, Cambridge U. Press, 1983.

4. *Structural Analysis of Discrete Data and Econometric Applications* [MF], Manski & McFadden <elsa.berkeley.edu/users/mcfadden/discrete.html>

5. *Panel Data Models: Some Recent Developments*, [AH] Manuel Arellano and Bo Honoré <ftp://ftp.cemfi.es/wp/00/0016.pdf>

I provide suggestions for reading specific parts of these additional references throughout the lecture notes, but these suggestions are always additional to already having read the relevant part of the Wooldridge textbook.

Below find a *simplified* course outline including *selected* suggested readings.[1]

1. Causal parameters and policy analysis in econometrics

    - Heckman, J.J. (2000) "Causal parameters and policy analysis in econometrics: A twentieth century perspective" *QJE* February 2000.

2. Cases where residuals are correlated

    - GLS
      Deaton A. (1997) *Analysis of Household Surveys*, Chapter 2.[2]
    - Panel data analysis
      Hsiao C. and M.H. Pesaran (2004) "Random Coefficient Panel Data Models,' IZA Discussion Paper no. 1236.

[1] You can find most of the listed papers as well as additional readings at \\ftp\LECTURES\YEAR_2\EconometricsIV.

[2] http://www.worldbank.com/lsms/tools/deaton/index.htm

3. Cases where residuals and regressors are correlated

   - Unobserved fixed effect in panel data analysis ([H] 3)

     Ashenfelter O. and A. Kruger (1994) "Estimates of the Economic Return to Schooling from a New Sample of Twins," *American Economic Review* 84: 1157-1173.

   - Errors in variables ([H] 3.9, [G] 9)

     Griliches Z. and J. Hausman (1986) "Errors in Variables in Panel Data," *Journal of Econometrics* 31:93-118.

   - Simultaneity

     Kling, J. (1999) "Interpreting IV Estimates of the Returns to Schooling," Princeton University, Industrial Relations Section WP 415.

     Hahn J. and J. Hausman (2002) "A New Specification Test for the Validity of Instrumental Variables," *Econometrica* 70(1)163-189.

4. Cases where linear regression models are not appropriate

   - Qualitative response models
   - Censored models and self-selection models

     Heckman, J.J. (1979) "Sample Selection Bias as a Specification Error," *Econometrica* 47:153-161.[3]

   - Duration analysis

     Kiefer N. (1988) "Economic Duration Data and Hazard Functions," *Journal of Economic Literature* 25(3): 646-679.

5. Introduction to nonparametric and semiparametric methods

   - Kernel estimation and Local Linear Regression

     Applied Nonparametric Regression, Härdle, Cambridge U Press, 1989.

   - Censored and sample-selection models

     Chay & Powell (2001) "Semiparametric Censored Regression Models"

[3] See also Lewbel A. (2004) "Simple Estimators For Hard Problems: Endogeneity in Discrete Choice Related Models"

Manning, A. (2003) "Instrumental Variables for Binary Treatments with Heterogeneous Treatment Effects: A Simple Exposition,"[4]

Carneiro, Hansen, Heckman (2003)[5]

Imbens, G.W. (2003) "Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review," UC Berkeley and NBER

[4] http://econ.lse.ac.uk/staff/amanning/work/econometrics.html

[5] http://www.nber.org/papers/w9546

Part I
Introduction

What do economists mean by "x affects y"?

We ask about effects using the conditional expectation function of y given x. Why? Because it is the best predictor of y in the sense of minimizing the mean of squared errors. We also like expectation as a concept that speaks to the typical values of y (given a particular x).

We typically quantify effects using regression analysis, which, originally, was a descriptive statistical tool.[6] *The* regression function, i.e., a conditional mean $E[y|x] = \int_{-\infty}^{\infty} y dF(y|x)$, has no behavioral meaning.

We can comfortably speak of x as being a *causal determinant* of y when we have (i) a theoretical model suggesting x causes y, and (ii) exogenous variation in x identifying the model (with identification being part of the model, ideally). When we then regress $y = x\beta + \varepsilon$ to estimate $\widehat{\beta}$, we can measure how much x causes y.[7] A causal variable x typically captures some treatment (policy) and we wish to answer policy-relevant "what if" questions.

1. Causal Parameters and Policy Analysis in Econometrics

Econometrics[8] differs from statistics in defining the identification problem (in terms of structural versus reduced form equations). "Cross-sectional" econometrics (as opposed to time-series) operationalizes Marshall's comparative statics idea (ceteris paribus) into its main notion of causality (compare to time-series analysis and its statistical Granger causality definition). The ultimate goal of econometrics is to provide policy evaluation.

[6] See, e.g., Deaton (1997) p. 63.

[7] Often, we focus on the effect of one *causal* variable (for which we have an exogenous source of variation) and use other regressors as *control* variables.

[8] This introductory class is based on a recent survey by J.J. Heckman (2000). By the end of the couse, make sure to come back to this introduction. By then, you could also read A. Deaton's 2008 Keynes Lecture (NBER WP no. 14690) and the reply by G. Imbens (NBER WP no. 14896).

In the classical paradigm of econometrics, economic models based on clearly stated axioms allow for a definition of well-defined structural "policy invariant" parameters. Recovery of the structural models allows for induction of causal parameters and policy-relevant analysis.

This paradigm was built within the work of the Cowless Commission starting in the 1930s. The Commission's agenda concerned macroeconomic Simultaneous Equation Models and was considered an intellectual success, but empirical failure due to incredible identifying assumptions.

A number of responses to the empirical failure of SEM developed, including first VAR and structural estimation methodology and later calibration, non-parametrics (sensitivity analysis), and the "natural experiment" approach. Let us in brief survey the advantages (+) and disadvantages (−) of each approach:

- VAR is "innovation accounting" time-series econometrics, which is not rooted in theory.

  (+) accurate data description

  (−) black box; may also suffer from non-credible identifying restrictions (as macro SEM); most importantly, results hard to interpret in terms of models.

- Structural estimation is based on explicit parametrization of preferences and technology. Here we take the economic theory as the correct full description of the data. The arguments of utility or technology are expressed as functions of explanatory variables. Given these $i-$specific arguments and an initial value of structural parameters, the optimization within the economic model (e.g., a nonlinear dynamic optimization problem) is carried out for each decision unit (e.g., unemployed worker). The predicted behavior is then compared with the observed decisions which leads to an adjustment of the parameters. Iteration on this algorithm (e.g., within MLE framework) provides the final estimates.

  (+) ambitious

  (−) computer hungry; empirically questionable: based on many specific functional form and distributional assumptions, but little sensitivity analysis is carried out given the computational demands, so estimates are not credible.[9]

[9] This criticism applies less in recent years, see Example 15.3.

- Calibration: explicitly rely on theory, but reject "fit" as the desired main outcome, focus on general equilibrium issues.

  (+) transparency in conditional nature of causal knowledge

  (−) casual in use of micro estimates, poor fit.

- Non-parametrics (as an extension of sensitivity analysis): do not specify any functional form of the "regression" in fear of biasing the results by too much unjustified structure.

  (+) transparency: clarify the role of distributional and functional form assumptions.

  (−) non-parametrics is very data hungry.

- Natural experiment: search for situations in the real world that remind us of an experimental setup. Use such experiments of nature (as instrumental variables) to identify causal effects; keep theory at an intuitive level. Ultimately, argue that randomized controlled trials (RCT) are the gold standard of science (as in evidence-based medicine[10]) and where possible abandon work with non-randomized data.[11]

  (+) internal validity and transparency: clear and mostly credible identification.

  (−) often we focus on *whether* treatment works but do not learn *why* it works (mechanism); causal parameters are relative to IV (LATE[12]); it may be hard to cumulate knowledge and the estimates do not render counterfactual policy predictions.

We return to the issue of economic-theory-based structural-model estimation, which allows for the estimation of well-defined policy-relevant parameters within

[10] It is hard to get acceptance on causality without experiments. The overwhelming correlation between smoking and lung cancer has been accepted as evidence of causality in absence of direct experimental evidence, but it took a very very long time. In medicine, only about a third of distinct *treatments* are supported by RCT evidence (as beneficial or likely to be beneficial) although most of medical *activity* (treatments used) is likely to be evidence-based. See http://clinicalevidence.bmj.com/ceweb/about/knowledge.jsp

[11] The RCT approach may lead one to abandon the use of economic theory. Further, Deaton (2009) argues that in some cases "...instrumental variables have moved from being solutions to a well-defined problem of inference to being devices that induce quasi-randomization."

[12] See Section 13.

an ex ante policy evaluation, versus IV- or randomization-based 'theory-free' evaluation in Section 13. Of course, the best work uses experimental identification to estimate structural models. To make the best choice, consider the Marschak's Maxim: estimators should answer *well-posed economic* problems with *minimal assumptions.*

Also, the fundamental problem of econometric policy evaluation is that to predict the future, it must be like the past, but the goal is to predict effects of a new policy, i.e. to predict a future that will not be like the past. Here, Marschak (1953) argues that predicting effects of future policy may be possible by finding past variation related to variation induced by new policy. The relationship between past and future variation is made using an economic model. Using this approach we may not need to know the full structural model to evaluate a particular policy. See Ichimura and Taber (2000).[13]

In this course, we will mostly remain within the classical paradigm and discuss parametric reduced-form econometric models. We will also occasionally touch on non-parametric and natural-experiment research and return to discussing causal inference when introducing the program evaluation literature in Section 13.

2. Reminder

This section aims at reminding ourselves with some basic econometrics. We started the introduction with the conditional expectation function $E[Y \mid X]$. The law of iterated expectations decomposes a random variable Y into the conditional expectation function of Y given X and a residual that is mean independent of X (i.e., $E[\varepsilon|x] = 0$) and also uncorrelated with (orthogonal to) any function of X.

[13] Marschak (1953) considers a monopolistic firm that experiments with output levels and observes profit as the outcome (this reduced form can be tabulated without knowledge of any structural parameters). Now, consider the government changing the tax rate on the monopolist's output. This changes the reduced form relationship of profit and output so it would seem that a new round of experimentation (tabulation) is necessary under the new policy for the government to predict the effects on profits. But one can estimate the demand function using the "old" data and then use an economic model (demand only depends on taxes through price) to predict the new profits. By using some aspect of a behavioral model one can exploit other types of variation to mimic a policy change.

Ichimura and Taber (2000) use this insight to provide a framework to link variation from a "natural experiment" to policy-equivalent variation. This requires stronger assumptions than just IV, but weaker compared to a fully structural model. For a related discussion see Deaton (2009).

Exercise 2.1. *Prove or provide a counterexample for the following statements:*

(a) $Y \perp X \iff COV(X,Y) = 0$. See also Exercise 2.2.

(b) $E[X \mid Y] = 0 \iff E[XY] = 0 \iff COV(X,Y) = 0$

(c) $E[X \mid Y] = 0 \implies E[Xg(Y)] = 0 \ \forall g(\cdot)$. Is $COV(X,Y) = 0$?

(d) $E[Y] = E_X[E_Y(Y \mid X)]$ and $V[Y] = \underbrace{E_X[V_Y(Y \mid X)]}_{\text{residual variation}} + \underbrace{V_X[E(Y \mid X)]}_{\text{explained variation}}$.

Why do we often run linear regressions (OLS)? Because when X and Y are jointly normal (see subsection 2.1) or when we work with a fully saturated model (with parameters for every combination of X values), then $E[Y \mid X]$ is linear and the linear regression function is it. More importantly, OLS is also the best linear predictor (best approximation of $E[Y \mid X]$ within the class of linear functions in terms of the minimum mean square error criterion). See also [W]1.

2.1. Note on Properties of Joint Normal pdf

In this note we show that the "true" regression function is linear if the variables we analyze are jointly Normal.[14] Let

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \ \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \ \text{ and } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Exercise 2.2. *Show that*

$$\Sigma_{12} = 0 \iff f(x \mid -) = f(x_1 \mid \mu_1, \Sigma_{11}) f(x_2 \mid \mu_2, \Sigma_{22})$$

i.e., under normality, linear independence is equivalent to independence in probability.

Theorem 2.1. *$E[X_2 \mid X_1]$ is linear in X_1.*

[14] Galton (1886), the fore-father of econometrics, studied height of parents and their children, two normally-distributed variables, and ran the first (linear) regression; he found "regression teoward mediocrity in hereditary stature," what we call today regression to the mean.

Proof. To get the conditional distribution of $X_2 \mid X_1$ first find a linear transformation of X which block-diagonalizes Σ :

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} I_1 & 0 \\ -\Sigma_{21}\Sigma_{11}^{-1} & I_2 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

$$\Longrightarrow VAR \begin{pmatrix} X_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22.1} \end{pmatrix}$$

and $X_1$and Y_2 are independent i.e., $Y_2 \equiv Y_2 \mid X_1 \sim N(\mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1, \Sigma_{22.1})$. Now note that $X_2 = Y_2 + \Sigma_{21}\Sigma_{11}^{-1}X_1$ and conditioning on $X_1$the last term is a constant$\Longrightarrow$ $X_2 \mid X_1 \sim N(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1), \Sigma_{22.1})$ or equivalently $X_2 \mid X_1 \sim N(\mu_{2.1} + \Delta_{2.1}X_1, \Sigma_{22.1})$. ■

Remark 1. $\mu_{2.1} = \mu_2 - \Delta_{2.1}\mu_1$ *is the intercept,* $\Delta_{2.1} = \Sigma_{21}\Sigma_{11}^{-1}$ *is the regression coefficient, and* $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ *is the conditional covariance matrix which is constant i.e., does not depend on* X_1 *(homoscedasticity).*

Remark 2. *OLS is attractive because it gives the minimum mean square error linear approximation to the conditional expectation function even when the linear model is misspecified.*

2.2. Testing Issues

Basic Principles

- Wald, Lagrange Multiplier, Likelihood Ratio. In class we provide a visualization of these in a graph. Note that they are asymptotically equivalent. So, obtaining different answers from each test principle may signal missspecification.[15]
- Specification tests: preview of Hansen and Hausman.
- Non-nested testing.

[15] Also, using the same level of significance for $N = 100$ and $N = 1,000,000$ is not right. With one million observations, you will reject any H0. The Leamer's (1978) rule for an F-test is to reject if $F > \frac{N-k}{r}\left(N^{r/N} - 1\right)$, where r is the number of restrictions and $N - k$ is the number of degrees of freedom of the unrestricted error sum of squares. (See Kmenta, 2nd edition, p. 422). For example, consider $k = 3$ and $r = 2$; when $N = 30$, reject if $F > 3.44$ (5% at 3.32) but with $N = 1,000$, reject if $F > 6.93$ (5% at 3.00).

Data Mining and Inference The validity of empirical work is often questioned because researchers do not test their theory by running one regression, but they "data mine"—i.e., search for the regression confirming their prior.

- Today, the Cowless commission paradigm (Haavelmo, 1944; Popper, 1959) is abandoned in favor of more *interaction with data* (learning) so that the paradigm is merely used as a reporting style (Leamer, 1978).[16] Even our initial theory comes in part from previous empirical work. Because it's clear that we all try different specifications, we report many alternative ones in our papers (sensitivity analysis) in order to convince the audience of the validity of our story.[17]

- A closely related problem of inference concerns *sequential testing*: While test properties are derived based on a one-shot reasoning, in practice we carry out a sequence of such tests, where the outcome of one test affects the next test (where both are based on the same data), invalidating the test properties. These concerns may be dealt with by setting aside a portion of the data before the start of the analysis and verifying the 'final' regression on this subset at the end of the analysis by means of a one-shot specification test. Another response is that you first have to make your model "fly" (i.e. achieve Durbin Watson =2) and only later can you go about testing it by, e.g., predicting out of sample.

- A more serious problem is perhaps the fact that we tend look for mistakes in data and programs as long as we don't like the results – as long as we do not confirm our expectation. The underlying problem is that there is no journal that would publish papers that fail to reject H_0 (a paper with good

[16] Let neither measurement without theory // Nor theory without measurement dominate // Your mind but rather contemplate // A two-way interaction between the two // Which will your thought processes stimulate // To attain syntheses beyond a rational expectation! Arnold Zellner [Zel96]

[17] But this will never be perfect. In Randomized controlled trials (RCT), one should post the research design before conducting the trial. Otherwise, you stop measuring when you reach a significant effect, or you look (data-mine) for the sub-group where treatment works, etc. If you measure the effect of x (treatment) on multiple outcomes (various ys), one of the ys may seem affected by accident. Look up the Bonferroni's correction for multiple comparisons. Together with clustering (see Remarks 10 and 11), this correction is brutal to statistical significance of treatment effects.

motivation, good data and regressions, and insignificant estimates). Meta analysis drives this point home in most fields of science.[18]

- Note that in econometrics we either test theory by means of estimation or use theory to identify our models (e.g., by invoking the Rational Expectations hypothesis in estimation of dynamic models in order to identify valid instruments).

3. Deviations from the Basic Linear Regression Model

Here, we consider 3 main departures from the basic classical linear model: (a) when they occur, (b) what the consequences are, and (c) how to remedy them. This preview sets the stage for our subsequent work in panel-data and limited-dependent-variable (LIMDEP) estimation techniques.

(i) $V[\varepsilon_i|x_i] = \sigma_i^2 \neq \sigma_\epsilon^2$, i.e. the diagonal of the middle part of the variance-covariance matrix is not full of 1s: (a) e.g., linear prediction vs. $E[y \mid x]$ or heteroscedasticity,[19] (b) the inference problem of having underestimated standard errors and hence invalidating tests, (c) GLS based on assumed form of heteroscedasticity or the heteroscedasticity-consistent standard errors (White, 1980). The Huber-White idea is that you don't need to specify the usually unknown form of how $V[\varepsilon_i \mid x_i]$ depends on x_i. The method ingeniously avoids having to estimate N of $\sigma_i^2(x_i)$ by pointing out that the k by k matrix $\sum_{i=1}^N x_i x_i' \widehat{\epsilon_i}^2$, where $\widehat{\epsilon_i}$ is the OLS predicted residual[20], converges to the true matrix with all of the $V[\varepsilon|x]$ so that

$$\widehat{V}(\widehat{\beta}_{OLS}) = \left(\sum_{i=1}^N x_i x_i'\right)^{-1} \sum_{i=1}^N x_i x_i' \widehat{\epsilon_i}^2 \left(\sum_{i=1}^N x_i x_i'\right)^{-1} .$$

(Here we also preview the Hausman test by comparing the OLS and Huber/White variance-covariance matrix. See [G]11.2,11.4, [W]4.2.3.

[18] See the Ashenfelter, Harmon and Oosterbeek (1999) test for publication bias where they run $\widehat{\beta_{IV}}$ estimates from several studies on their standard error. In the medical RCT literature, people use a funnel plot of treatment effect against trial size for the same purpose.

[19] Arises all the time. For example when working with regional averages $y_r = \frac{1}{N_r}\sum_{i=1}^{N_r} y_{ir}$ we have $V(y_r) = \frac{1}{N_r}V(y_{ir})$.

[20] Remember that with heteroscedasticity OLS still provides unbiased estimates of βs, so that $\widehat{\varepsilon} = y - x'\widehat{\beta}_{OLS}$ is also unbiased.

(ii) $COV[\varepsilon_i, \varepsilon_j \mid x_i, x_j] \neq 0$: (a) time series or unobserved random effect (family effects), (b) possible inconsistency of β (for example when estimating $y = \alpha + \epsilon$, the asymptotic variance of $\widehat{\alpha}$ does not converge to 0) , (c) GLS, Chamberlin's trick (see below).

(iii) $E[\varepsilon_i \mid x_i] \neq 0$: (a) Misspecification, Simultaneity, Lagged dependent variables and serial correlation in errors, Fixed effect model, Measurement error, Limited dependent variables; (b) inconsistency of β, (c) GMM/IV, non-parametrics, MLE.

In the first part of the course on panel data, we will first deal with (i) and (ii) by running various GLS estimators. Second we will also explore panel data techniques of dealing with (iii). The second part of the course on LIMDEP techniques will all address (iii).

Example 3.1. *GLS in spacial econometrics (see p.526 in Anselin, 1988) Here we present a way of parametrizing cross-regional correlation in* ϵ*s (using analogy between time correlation coefficient and spacial correlation) and provide an example of how non-nested testing arises (e.g., with respect to how we specify the contiguity matrix summarizing prior beliefs about the spacial correlation) and what it means to concentrate the likelihood. Most importantly, we remind ourselves of how FGLS works in two steps. The first part of the panel data analysis (Section 4) will all be FGLS.*

Part II
Panel Data Regression Analysis

Reading assignment: [H] 1.2, 2, 3.2 - 3.6, 3.8, 3.9.

4. GLS with Panel Data

So far we talked about cases when OLS fails to do its job and GLS fixes the problem, i.e. cases where the variance assumption is violated. Now, we are going to apply that reasoning in the panel data context.

The model we have in mind is

$$\begin{aligned} y_{it} &= x_{it}'\beta_i + \epsilon_{it} \text{ with } i = 1, ..., N \text{ and } t = 1, ..., T \text{ , or} \\ \underset{T\times 1}{y_i} &= \underset{T\times k}{X_i}\,\underset{k\times 1}{\beta_i} + \epsilon_i \text{ with } i = 1, ..., N \text{ or} \\ \underset{NT\times 1}{y} &= \underset{NT\times kN}{\begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & X_N \end{bmatrix}} \underset{kN\times 1}{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{pmatrix}} + \epsilon \text{ or } y = \underset{NT\times k}{\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}} \beta + \epsilon \end{aligned} \tag{4.1}$$

where the covariance structure of ϵ_{it} will again be of interest to us. In a panel model we can allow for much more flexible assumptions then in a time series or a cross-section.

Remark 3. *N and T do not necessarily refer to number of individuals and time periods respectively. Other examples include families and family members, firms and industries, etc.*

Remark 4. *The number of time periods T may differ for each person. This is often referred to as unbalanced panel.*

Remark 5. *T is usually smaller than N and most asymptotic results rely on $N \to \infty$ with T fixed.*

The first question is whether we constrain β to be the same across either dimension. We cannot estimate β_{it} as there is only NT observations.

4.1. SURE

Suppose we assume $\beta_{it} = \beta_i \ \forall t$, that is for some economic reason we want to know how βs differ across cross-sectional units or F test rejects $\beta_{it} = \beta \ \forall i, t$.

If $E[\varepsilon_{it} \mid x_{it}] = 0 \ \forall t$ and $V[\varepsilon_{it} \mid x_{it}] = \sigma^2_{ii}$ and $(x_{it}, \varepsilon_{it})$ is *iid* $\forall t$ then we estimate β_i by running N separate OLS regressions. (Alternatively we can estimate $y_{it} = x'_{it}\beta_t + \epsilon_{it}$.)

Now, if the covariance takes on a simple structure in that $E(\epsilon_{it}\epsilon_{jt}) = \sigma^2_{ij}$ and $E(\epsilon_{it}\epsilon_{js}) = 0$ there is cross-equation information available that we can use to improve the efficiency of our equation-specific β_is. We have $V[\varepsilon] = E[\varepsilon\varepsilon'] = \Sigma \otimes I_T \neq \sigma^2 I_{NT}$, i.e. the ϵ 's are correlated across equations and we gain efficiency by running GLS (if $X_i \neq X_j$) with $\widehat{\sigma^2_{ij}} = \frac{1}{T}\,\widehat{\epsilon_i}'\,\widehat{\epsilon_j}$ where the $\widehat{\varepsilon}$ first comes from OLS as usual. Iterated FGLS results in MLE in asymptotic theory. In class we demonstrate the GLS formula for SURE and get used to having two dimensions in our data (formulas) and variance-covariance matrices.

4.2. Random Coefficients Model

What if we still want to allow *parameter* flexibility across cross-sectional units, but some of the β_is are very uninformative. Then one solution may be to combine the estimate of β_i from each time series regression 4.2 with the 'composite' estimate of β from the pooled data in order to improve upon an imprecise $\widehat{\beta}_i$ using information from other equations.[21] In constructing β, each β_i should then be given a weight depending on how informative it is.

To operationalize this idea, the RCM model allows the coefficients to have a random component (something typical for Bayesians, see [H 6.2.2]), i.e. we assume

$$\underset{T\times 1}{y_i} = X_i\beta_i + \epsilon_i \tag{4.2}$$

where the error terms are well behaved, but

$$\underset{K\times 1}{\beta_i} = \underset{\text{nonstochastic}}{\beta} + \nu_i \text{ with } E[\nu_i] = 0 \text{ and } E[\nu_i\nu_i'] = \Gamma.$$

OLS on 4.2 will produce $\widehat{\beta}_i$ with $V[\widehat{\beta}_i] = \sigma^2_i(X_i'X_i)^{-1} + \Gamma = V_i + \Gamma$

Exercise 4.1. *Show that the variance-covariance matrix of the residuals in the pooled data is* $\Pi = diag(\Pi_i)$, *where* $\Pi_i = \sigma^2_i I + X_i'\Gamma X_i$.

[21] Note that in a SURE system, each $\widehat{\beta}_i$ is coming from equation by equation OLS.

V_i tells us how much variance around β is in $\widehat{\beta}_i$. Large V_i means the estimate is imprecise.

Let $\widehat{\beta} = \sum_{i=1}^{N} w_i \widehat{\beta}_i$, where $\sum_{i=1}^{N} w_i = I$. The optimal choice of weights is

$$w_i = \left[\sum_{j=1}^{N} (V_j + \Gamma)^{-1} \right]^{-1} (V_i + \Gamma)^{-1} \tag{4.3}$$

Γ can be estimated from the sample variance in $\widehat{\beta}_i$'s ([G] p318). Note that $\widehat{\beta}$ is really a matrix weighted average of OLS.

Exercise 4.2. *Show that $\widehat{\beta}$ is the GLS estimator in the pooled sample.*

Remark 6. *As usual we need asymptotics to analyze the behavior of $\widehat{\beta}$ since weights are nonlinear. Also note that $\widehat{\Gamma}$ is coming from the cross-sectional dimension, while $\widehat{\beta}_i$ is estimated off time series variation.*

Finally, we recombine $\widehat{\widehat{\beta}}_i = A_i \widehat{\beta} + (I - A_i)\widehat{\beta}_i$ with optimal[22] $A_i = (\Gamma^{-1} + V_i^{-1})^{-1}\Gamma^{-1}$.

Remark 7. *If $E[\nu_i] = f(X_i) \Longrightarrow E[\nu_i \mid X_i] \neq 0 \Longrightarrow \widehat{\beta}_i$ is not consistent for β_i.*

Remark 8. *As a digression, consider a situation when simple cross-sectional data are not representative across sampling strata, but weights are available to re-establish population moments.*[23] *First consider calculating the expectation of y (weighted mean). Then consider weighting in a regression. Under the assumption that regression coefficients are identical across strata, both OLS and WLS (weighted least squares) estimators are consistent, and OLS is efficient.*[24] *If the parameter vectors differ for each sampling strata $s = 1, ..., S$ so that $\beta_s \neq \beta$, a*

[22] See [H] p.134 if you are interested in the optimality of A_i.

[23] For source see Deaton's *Analysis of Household Surveys* (1997, pp. 67-72).

[24] Although, see the Imbens and Hellerstein (1993/1999) study mentioned at the end of Section 8. Also, when data are grouped, running the average Y for each value of X on X will replicate the microdata regression of Y on the grouped X when weighting by the number of Y observations for each grouped X. On the other hand, weighting using heteroscedasticity weigths (including those used in the linear probability model of Section 10.1.1) is also questionable since the conditional variance model may be poorly estimated, thus messing up the hoped-for efficiency improvements. Unweighted regressions will always be the minimum mean square error approximations of the population conditional expectation.

regression slope estimator analogous to the mean estimator is a weighted average of strata-specific regression estimates:

$$\widehat{\beta} = \sum_{s=1}^{S} W_s \widehat{\beta}_s, \qquad \widehat{V}(\widehat{\beta}) = \sum_{s=1}^{S} W_s^2 \widehat{V}(\widehat{\beta}_s), \tag{4.4}$$

where W_s are scalar strata-specific weights, and where $\widehat{\beta}_s$ is an OLS estimate based on observations from stratum s. In contrast, the WLS procedure applied to pooled data from all strata results in an estimator $\widehat{\beta}_{WLS}$,

$$\widehat{\beta}_{WLS} = \left(\sum_{s=1}^{S} W_s X_s' X_s \right)^{-1} \sum_{s=1}^{S} W_s X_s' y_s = \left(\sum_{s=1}^{S} W_s X_s' X_s \right)^{-1} \sum_{s=1}^{S} W_s X_s' X_s \widehat{\beta}_s,$$

which is in general not consistent for the weighted average of β_s.[25]

4.3. Random Effects Model

Assuming $\beta_{it} = \beta \ \forall i, t$ in Equation 4.1 one can impose a covariance structure on ϵs and apply the usual GLS approach. The random effects model (REM) specifies a particularly simple form of the residual covariance structure, namely $\epsilon_{it} = \alpha_i + u_{it}$ with $E[\alpha_i \alpha_j] = \sigma_\alpha^2$ if $i = j$ and is 0 otherwise. Other than that the only covariance is between u_{it} and u_{it} which is σ_u^2. We could also add a time random effect λ_t to ϵ_{it}.

Given this structure $V \equiv V(\underset{T\times 1}{\epsilon_i}) = \sigma_u^2 I_T + \sigma_\alpha^2 e_T e_T'$, where e_T is a $T \times 1$ column of numbers 1. We write down $E[\epsilon\epsilon']$ using V and invert V using the partitioned inverse formula to write down the GLS formula:

$$\widehat{\beta}_{GLS} = \left(\sum_{i=1}^{N} X_i' V^{-1} X_i \right)^{-1} \sum_{i=1}^{N} X_i' V^{-1} y_i \tag{4.5}$$

The GLS random effects estimator has an interpretation as a weighted average of a "within" and "across" estimator. We show this in class by first skipping to

[25] The WLS estimator is consistent for β if the parameter variation across strata is independent of the moment matrices and if the number of strata is large (see, e.g., Deaton, 1997, p. 70). Further, Pesaran et al. (2000) note that neglecting coefficient heterogeneity can result in significant estimates of incorrectly included regressors and bias other parameters even if the erroneously included variables are orthogonal to the true regressors.

the fixed effect model to describe the within estimator. Then we return to the above GLS formula, re-parametrize V^{-1} using the matrix $Q = I_T - \frac{1}{T}e_T e_T^{'}$, which takes things in deviation from time mean, and gain intuition by observing the two types of elements inside the GLS formula: (i) the "within" estimator based on deviations from mean $x_{it} - \overline{x_i}$ and (ii) the "across" estimator working off the time averages of the cross-sectional units, i.e. $\overline{x_i} - \overline{x}$. Treating α_i as random (and uncorrelated with x) provides us with an intermediate solution between treating α_i as being the same ($\sigma_\alpha^2 = 0$) and as being different ($\sigma_\alpha^2 \to \infty$). We combine both sources of variance: (i) over time within i units and (ii) over cross-sectional units.

As usual, the random effects GLS estimator is carried out as FGLS (need to get $\widehat{\sigma_u^2}$ and $\widehat{\sigma_\alpha^2}$ from OLS on within and across dimensions).

Remark 9. *With panel data one does not have to impose so much structure as in REM: (i) can estimate the person specific residual covariance structure, see the next Remark, and (ii) we can use minimum distance methods and leave the structure of error terms very flexible (see section 6.3.2).*

Remark 10. *Cross-sections with group-level variables can be thought of as panel data. If you are interested in the effect of the group-level variable, you need to admit that it does not vary independently across individual observations.*[26] *This is done by adjusting standard errors by clustering: Apply the White (1980) idea to estimate $\widehat{\Omega}$ while allowing for any unconditional heteroscedasticity as well as for correlation over time* within *a cross-sectional unit (or group):*[27]

$$\widehat{\Omega} = \sum_{i=1}^{N} X_i' \widehat{\varepsilon}_i \widehat{\varepsilon}_i{}' X_i.$$

Remark 11. *Along similar lines, one needs to adjust inference for multi-stage survey design (such as selecting randomly villages and then randomly households*

[26] A haiku by Keisuke Hirano:
T-stat looks too good
Try clustered standard errors—
Significance gone

[27] See 8.7 and http://www.stata.com/support/faqs/stat/cluster.html. Also see Wooldridge (2003, AER). Clustering may not work when the number of clusters is small (below 50). We come back to this topic in Section 6.2.

within villages—unobservables will be related within a village). There is an extensive set of commands for such adjustments available in Stata: see `svy` *commands.*

Remark 12. *Of course, group-level variables need not be just exogenous x controls. Group-level IVs are discussed in Remark 34. Sometimes we also ask why individuals belonging to the same group act in a similar way and whether they reflect each other's behavior—in that case, the group-level right-hand-side variable is the average of y for the group. This makes for some thorny identification issues; see Manski (1995), Durlauf (2002) and Brock and Durlauf (2001).*

5. What to Do When $E[\varepsilon \mid x] \neq 0$

5.1. The Fixed Effect Model

One of the (two) most important potential sources of bias in cross-sectional econometrics is the so called heterogeneity bias arising from unobserved heterogeneity related to both y and x.

Example 5.1. *Estimation of the effect of fertilizer on farm production in the presence of unobserved land quality; an earnings function and schooling when ability is not observed, or a production function when managerial capacity is not in the data, imply possibility of heterogeneity bias.*

If we have valid IVs (exclusion restriction), we can estimate our model by TSLS. If we have panel data, however, we can achieve consistency even when we do not have IVs available. If we assume that the unobservable element correlated with x does not change over time, we can get rid of this source of bias by running the fixed effect model (FEM). This model allows for an individual specific constant, which will capture all time-constant (unobserved) characteristics:

$$y_{it} = \alpha_i + x_{it}'\beta + \epsilon_{it} \tag{5.1}$$

When $T \geq 2$ the fixed effects α_i are estimable, but if N is large, they become nuisance parameters and we tend to get rid of them: by estimating the model on data taken in deviation from the time mean or by time differencing.

To summarize, the FEM is appropriate when the unobservable element α does not vary over time and when $COV[\alpha_i, X_i] \neq 0$. This nonzero covariance makes the $\widehat{\beta}_{OLS}$ and $\widehat{\beta}_{GLS}$ inconsistent. We'll come to the testing issue in section 6.

Suppose $x'_{it} = (w_{it}, z_i)$ and partition β appropriately into β^w and β^z. In this case note that we cannot separately identify β^z from α_i. This shows that when we run the fixed effect model, $\widehat{\beta}$ is identified from individual variation in X_i around the individual mean, i.e. $\widehat{\beta}$ is estimated off those who switch (change x over time). $\widehat{\alpha_i}$'s are unbiased, but inconsistent if T is fixed. Despite the increasing number of parameters as $N \longrightarrow \infty$, OLS applied to 5.1 yields consistent $\widehat{\beta^w}$ because it does not depend on $\widehat{\alpha_i}$. To see this solve the following exercise.

Exercise 5.1. *Let* $M_D = I_{NT} - D(D'D)^{-1}D'$, *where*

$$D = \begin{bmatrix} e_T & 0 & \dots & 0 \\ 0 & e_T & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & e_T \end{bmatrix} \text{ and } e_T = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Using the definition of M_D show $\widehat{\beta^w}$ is estimated by a regression of $y_{it} - \overline{y}_{i\cdot}$ on $w_{it} - \overline{w}_{i\cdot}$, where $\overline{w}_{i\cdot} = \frac{1}{T}\sum_{t=1}^{T} w_{it}$.

Remark 13. *For small* T *the average* $\overline{w}_{i\cdot}$ *is not a constant, but a r.v. Hence* $E[\epsilon_{it} \mid w_{it}] = 0$ *is no longer enough, we need* $E[\epsilon_{it} - \overline{\epsilon}_{i\cdot} \mid w_i] = 0$.

Remark 14. *Of course, we may also include time dummies, i.e. time fixed effects. We may also run out of degrees of freedom.*

Remark 15. *There is an alternative to using panel data with fixed effects that uses repeated observations on cohort averages instead of repeated data on individuals. See Deaton (1985) Journal of Econometrics.*

Remark 16. *While effects of time-constant variables are not identified in fixed effects models, one can estimate the change in the effect of these variables. Angrist (1995) AER.*

Remark 17. *There could be several fixed effects at different data dimensions (e.g., worker as well as firm time-constant unobservables). The estimation may get technically difficult, as shown by Abowd, Kramarz, and Margolis (Econometrica 67: 251–333) who discuss the estimation of three-way error-component models. Andrews, Schank and Upward (2006, Stata Journal 6 (4)) support the implementation of some of these methods into Stata (see* `st0112`*).*

5.2. Errors in Variables

([H] 3.9) One particular form of endogeneity of RHS variables was of concern in the previous section. We used the fixed effect model to capture time constant person-specific characteristics. The second most important potential source of bias is measurement error. Its effects are opposite to those of a typical unobserved fixed effect. Consider the model 5.1, where x is measured with error, i.e. we only observe $\widetilde{x}$ such that

$$\widetilde{x}_i = x_i + \nu_i \tag{5.2}$$

In the case of classical measurement error, when $E[\nu\varepsilon] = 0$, OLS is inconsistent and biased towards 0. For a univariate x_{it} we show in class that

$$\widehat{\beta}_{OLS} \xrightarrow{p} \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\nu^2}\beta \tag{5.3}$$

Note that what matters is the ratio of the 'signal' σ_x^2 to 'noise' σ_ν^2. Also note that adding additional regressors will typically exacerbate the measurement error bias because the additional regressors absorb some of the signal in $\widetilde{x}$.

Exercise 5.2. *Suppose there are two variables in x_{it}, only one of which is measured with error. Show whether the coefficient estimator for the other variable is affected as well.*

Remark 18. *In the case of misclassification of a binary variable $E[\nu\varepsilon] = 0$ cannot hold. This still biases the coefficient towards 0 (Aigner, 1973). However, the bias can go either way in other cases of non-classical measurement error.*

Within estimators (differencing) will typically make the measurement error bias worse. The signal-to-noise ratio will depend on σ_x^2 and on $\sigma_x^2 + \sigma_\nu^2[(1-\tau)/(1-\rho)]$ where τ is the first-order serial correlation in the measurement error and ρ is the first-order serial correlation in x. Again, the intuition is that differencing kills some of the signal in $\widetilde{x}$ because x is serially correlated, while the measurement error can occur in either period.

Exercise 5.3. *Derive the above-stated result.*

Exercise 5.4. *Explain how we could use a second measurement of x_{it} to consistently estimate β.*

Example 5.2. *In estimating the labor supply equation off PSID data the measure of wages is created as earnings over hours. If there is a measurement error in hours, the measurement error in wages will be negatively correlated with the error term in the hours equation.*

Remark 19. *When you don't have an IV, use reliability measures (separate research gives you these).*

Remark 20. *IV estimation method for errors in variables does not generalize to general nonlinear regression models. If the model is polynomial of finite order it does: see Hausman et al. (1991). See Schennach for use of Fourier transformation to derive a general repeated-measurement estimator for non-linear models with measurement error.*

Exercise 5.5. *Assume a simple non-linear regression model $y_i = \beta f(x_i) + \varepsilon_i$ with one regressor x_i measured with error as in Equation 5.2. Use Taylor series expansion around $\widetilde{x}$ to illustrate why normal IV fails here.*

Griliches and Hausman (1986): "Within" estimators are often unsatisfactory, which was blamed on measurement error. Their point: we may not need extraneous information. If $T > 2$ differencing of different lengths and the deviations-from-mean estimator will eliminate fixed effects and have a different effect on potential bias caused by measurement error. Therefore, differencing may suggest if measurement error is present, can be used to test if errors are correlated, and derive a consistent estimator in some cases. Note that here again (as with the fixed effect model) panel data allows us to deal with estimation problems that would not be possible to solve in simple cross-section data in absence of valid instruments.

Example 5.3. *Aydemir and Borjas (2006) estimate the impact of the share of immigrants on a local labor markets on wages. The regional share of immigrants π_r is estimated from random samples of the population, which introduces sampling error in the key independent variable. The variance of the sampling error is binomial $(\pi_r(1-\pi_r)/n_r)$ for very small sampling rates (τ). It can be shown that the inconsistency takes the form*

$$\widehat{\beta}_{OLS} \xrightarrow{p} \left(1 - \frac{(1-\tau)E_r[\pi_r(1-\pi_r)/n_r]}{(1-R^2)\sigma_p^2}\right)\beta, \tag{5.4}$$

where σ_p^2 is the variance of the observed π_r across the labor markets and where the R^2 comes from an auxiliary regression of π_r on all right-hand-side variables in *the model. The authors show that the use of regional fixed effects, typical in this literature, makes the auxiliary-regression R^2 close to 1 and the attenuation bias can be very large.*[28] *Equation (5.4) can be used to predict the correct coefficient or run IV using two measures. Note that the same problem is present when regressing outcomes on group-specific shares of unemployed, women, foreign firms, etc. as long as random samples are used to estimate the group-specific shares.*

6. Testing in Panel Data Analysis

Tests like *Breusch-Pagan* tell us whether to run OLS or random effects (GLS). What we really want to know is whether we should run fixed effects or random effects, i.e., is $COV[\alpha_i, X_i] \neq 0$?

Remark 21. *Mundlak's formulation connects random and fixed effects by parametrizing α_i (see [H] 3).*

6.1. Hausman test

- Basic idea is to compare two estimators: one consistent under both null hypothesis (no misspecification) and under the alternative (with misspecification), the other consistent only under the null. If the two estimates are significantly different, we reject the null.

	$\widehat{\beta}_{LSDV}$ fixed effects	$\widehat{\beta}_{GLS}$ random effects
$H_0 : COV[\alpha_i, X_i] = 0$	consistent, inefficient	consistent, efficient
$H_A : COV[\alpha_i, X_i] \neq 0$	consistent	inconsistent

- The mechanics of the test:

Theorem 6.1. *Under H_0 assume $\sqrt{n}(\widehat{\beta}_j - \beta) \xrightarrow{D} N(0, V(\widehat{\beta}_j)), j \in \{LSDV, GLS\}$ and $V(\widehat{\beta}_{LSDV}) \geq V(\widehat{\beta}_{GLS})$ and define $\sqrt{n}\ \widehat{q} = \sqrt{n}(\widehat{\beta}_{LSDV} - \widehat{\beta}_{GLS}) \xrightarrow{D} N(0, V(\widehat{q}))$ where*

$$V_q \equiv V(\widehat{q}) = V(\widehat{\beta}_{LSDV}) + V(\widehat{\beta}_{GLS}) - COV(\widehat{\beta}_{LSDV}, \widehat{\beta}'_{GLS}) - COV(\widehat{\beta}_{GLS}, \widehat{\beta}'_{LSDV}).$$

then

$$COV(\widehat{\beta}_{LSDV}, \widehat{\beta}'_{GLS}) = COV(\widehat{\beta}_{GLS}, \widehat{\beta}'_{LSDV}) = V(\widehat{\beta}_{GLS})$$

so that we can easily evaluate the test statistic $\widehat{q}' V_q^{-1} \widehat{q} \longrightarrow \chi^2(k)$.

[28] A back-of-the-envelope percent bias is $(1-\tau)\left(\overline{\pi}(1-\overline{\pi})/\overline{n}\right)\left[(1-R^2)\sigma_p^2\right]^{-1}$.

We prove the theorem in class using the fact that under H_0 the $\widehat{\beta}_{GLS}$ achieves the Rao-Cramer lower bound.

Remark 22. *The Hausman test asks if the impact of X on y within a person is the same as the impact identified from both within and cross-sectional variation.*

Remark 23. *Similar to the Hansen test (see Section 8), Hausman is an all-encompassing misspecification test, which does not point only to $COV[\alpha_i, X_i] \neq 0$, but may indicate misspecification. Of course, tests against specific alternatives will have more power.*

Remark 24. *The power of the Hausman test might be low if there is little variation for each cross-sectional unit. The fixed effect $\widehat{\beta}$ is then imprecise and the test will not reject even when the βs are different.*

Remark 25. *There is also a typical sequential testing issue. What if I suspect both individual and time fixed effects: which should I first run Hausman on. Since T is usually fixed, it seems safe to run Hausman on the individual effects, with time dummies included. But then we may run out of degrees of freedom.*

6.2. Inference in "Difference in Differences"

One of the most popular (panel-data) identification strategies, pioneered by Ashenfelter and Card (1985), is to estimate the difference in before/after time differences in the mean outcome variable across (two) groups, where some (one) of the groups experience a treatment (policy change). For example, think of using repeated cross-sections of workers across U.S. states, where some of the states change the statutory minimum wage at some year, and estimate the following fixed effect model of the effect of this treatment T on y:

$$y_{ist} = \alpha_s + \delta_t + \gamma x_{ist} + \beta T_{st} + \varepsilon_{ist}, \tag{6.1}$$

where i denotes workers, s denotes states (groups) and t denotes time (year).[29]

There is a serial correlation inference problem similar to the within-group correlation problem we encountered in Remark 10 (and Example 8.7). Moreover, traditional inference only considered uncertainty entering through sampling error

[29] Of course, one can go further and focus on the effect on a sub-group within ist, defining a difference-in-difference-in-differences estimate.

in the estimation of the group (before/after) means of y. Recent work, however, argues that one should allow for additional sources of uncertainty when forming inference, in particular, uncertainty in the quality of the control groups (Abadie, Diamond, and Hainmueller, 2007). So far, the practical upshot of the recent work is that when there are many treatment and control groups (many changes in laws and many control states), it is ok to just cluster ϵ at group (state) level (again, see Remark 10). If the number of policy changes is small, however, there is an inference problem.[30]

Bertrand et al. (2001) suggest that (6.1) gives wrong standard errors because (i) it relies on long time series, (ii) the dependent variables are typically highly positively serially correlated, and (iii) the treatment dummy T_{st} itself changes little over time. In their paper, placebo laws generate significant effects 45% of the time, as opposed to 5% (they repeatedly generate 'law' changes at random and estimate the model and they count the number of estimates suggesting a significant effect).[31] If the number of groups is small, they suggest the use of randomization inference tests: Use the distribution of estimated placebo laws to form the test statistic.

Conley and Taber (NBER TWP 312) propose a more general inference method for cases where the number of treatment groups is small, but the number of control groups is large.[32] They propose to use estimates obtained from the many controls to characterize the small-sample distribution of the treatment parameter. Think of comparing the (change in the) outcome in the treatment state to the distribution of the corresponding outcomes from all other states—is it an outlier? β is obviously not consistent with only few treatment groups, but we can get confidence intervals based on the empirical distribution of residuals from the (many) control groups.

Donald and Lang (2007, REStat) focus on the case where the number of groups (both treatment and control) is small. They show that clustering (as well as the Moulton random effect correction) is misleading. An alternative is the two step approach where one first estimates an individual-level regression with fixed effects corresponding to the group-level variation; in the second stage one runs these fixed

[30] In the famous Card and Krueger 1994 AER minimum wage study, there is only one treatment group (NJ) and only one control group (PA). No matter how many restaurants are covered in each state, the treatment effect may not be consistently estimated.

[31] As a solution they propose to aggregate up the time series dimension into pre- and post-treatment observations or allow for arbitrary covariance over time and within each state. These solutions work fine if the number of groups is sufficiently large—simply cluster.

[32] In terms of the Card-Krueger study, compare NJ to not only PA, but all other states.

effects on the group-level RHS variable.[33] If the number of individuals within each group is large, this two-step estimator is efficient and its t-statistics are distributed t if the underlying groups errors are normally distributed.[34] If, however, the number of members of each group is small, one needs special circumstances to argue that the two-step estimation is efficient with t-statistics having t distributions. They even recommend running the first stage separately for each group. This approach, together with the whole clustering discussion, is also covered in Wooldridge (2003, AER).[35] A workable bootstrap-t solution for standard errors (based on resampling of clusters) when the number of clusters is as low as six is provided in Cameron et al. (2008). Barrios et al. (NBER WP No. 15760) illustrate the behavior (bias) of clustered standard errors when there is some correlation between clusters (for example, when state policies exhibit substantial spatial correlations).

6.3. Using Minimum Distance Methods in Panel Data

Hausman test might reject $COV[\alpha_i, X_i] = 0$ and one may then use of the fixed effect model. But the fixed effect model model is fairly restrictive and eats up a lot of variation for α_is. When T is small we can test the validity of those restrictions using the MD methods. The same technique allows for estimation of β with a minimal structure imposed on α, allowing for correlation between the unobservable α and the regressors x. We will first understand the MD method and then apply it to panel data problems.

[33] One can think of the second stage as a Minimum Distance problem (see Section 6.3.1) where one ought to weight with the inverse of the variance of the estimated fixed effects. One may also need to think about the different implied weighting in the individual- and group-level regressions, and about omitted variable bias (see, e.g., Baker and Fortin, 2001).

[34] They think of the problem as follows: a small number of groups (NJ and PA, see note n. 30) are drawn from the population of potential groups. Given the groups, one gets the group averages based on random samples taken within groups. The NJ-PA dif-in-difs is based on 4 means, each of which is obtained from a large sample—so they are approximately normal, such that the resulting dif-in-difs parameter should also be approximately normal. But Donald and Lang reject this idea and suggest inference based on the t distribution, which can get degenerate.

[35] See also an update of the AER paper, Wooldridge (2006), and Hansen (2007), who offers some simulations. One of the upshots is, again, that one should not cluster with few groups and large group sizes. A useful way to test for the restrictions involved in these different approaches is to frame the issue as a minimum distance problem (see Loeb and Bound, 1996). See also Athey and Imbens (2006) for a non-parametric difference-in-differences.

6.3.1. The Minimum Distance Method

Suppose we have a model that implies restrictions on parameters that are hard to implement in the MLE framework. When estimation of an unconstrained version of our model is easy (OLS) and consistent, the MD method offers a way to impose the restrictions and regain efficiency and also to test the validity of the restrictions ([H] 3A).

Denote the unconstrained estimator as $\widehat{\pi}_N$, where N is the sample size in the unconstrained estimation problem, and denote the constrained parameter of interest as θ. Next, maintain the assumption that at the true value of θ the restrictions $\pi = f(\theta)$ are valid. The objective is to find $\widehat{\theta}$ such that the distance between $\widehat{\pi}$ and $f(\widehat{\theta})$ is minimized:[36]

$$\widehat{\theta}_N = \arg\min\{S_N\} \text{ where } S_N = N[\widehat{\pi}_N - f(\theta)]' A_N [\widehat{\pi}_N - f(\theta)], \tag{6.2}$$

and where $A_N \xrightarrow{p} A$ is a weighting matrix and $\sqrt{N}[\widehat{\pi}_N - f(\theta)] \xrightarrow{D} N(0, \Delta)$.[37]

The minimization problem 6.2 is of considerably smaller dimension than any constrained estimation with the N data points!

Theorem 6.2. *Under the above assumptions and if f is 2nd order differentiable and $\frac{\partial f}{\partial \theta'}$ has full column rank then a) $\sqrt{N}[\widehat{\theta}_N - \theta] \xrightarrow{D} N(0, V(A))$, b) the optimal $A = \Delta^{-1}$, and c) $\widehat{S_N} \xrightarrow{D} \chi^2(r)$ where $r = \dim(\pi) - \dim(\theta)$ is the number of overidentifying restrictions.*

We provide the proof in class. To show a) simply take a FOC and use Taylor series expansion to relate the distribution of $\widehat{\theta}_N$ to that of $\widehat{\pi}_N$.

Remark 26. *Note that the Minimum Distance Method is applicable in Simultaneous Equation Models to test for exclusion restrictions.*

$$\Gamma y_t + B x_t = u_t \ \Rightarrow y_t = \Pi x_t + v_t \text{ where } \Pi = -\Gamma^{-1} B$$

and we can test zero restrictions in Γ and B.

Remark 27. *MD is efficient only among the class of estimators which do not impose a priori restrictions on the error structure.*

[36]Find the minimum distance between the unconstrained estimator and the hyperplane of constraints. If restrictions are valid, asymptotically the projection will prove to be unnecessary.

[37]See Breusch-Godfrey 1981 test in Godfrey, L. (1988).

6.3.2. Arbitrary Error Structure

When we estimate random effects, $COV[\alpha, x]$ must be 0; further, the variance-covariance structure in the random effect model is quite restrictive. At the other extreme, when we estimate fixed effects, we lose a lot of variation and face multicollinearity between α_i and time constant x variables.

However, when T is fixed and $N \longrightarrow \infty$,[38] one can allow α to have a general expectations structure given x and estimate this structure together with our main parameter of interest: β (Chamberlain 1982, [H] 3.8). That is we will not eliminate α_i (and its correlation with x) by first differencing. Instead, we will control for (absorb) the correlation between α and x by explicitly parametrizing and estimating it. This parametrization can be rich: In particular, serial correlation and heteroscedasticity can be allowed for without imposing a particular structure on the variance-covariance matrix. In sum, we will estimate β with as little structure on the omitted latent random variable α as possible.[39] The technique of estimation will be the MD method.

Assume the usual fixed effect model with only $E[\varepsilon_{it} \mid x_{it}, \alpha_i^*] = 0$

$$y_i = e_T \alpha_i^* + \underset{T\times K}{X_i} \beta + \varepsilon_i \tag{6.3}$$

and let $x_i = \text{vec}(X_i^{'})$.[40] To allow for possible correlation between α_i and X_i, assume $E[\alpha_i^* \mid X_i] = \mu + \lambda^{'} x_i = \sum_{t=1}^{T} \lambda_t^{'} x_{it}$ (note μ and λ do not vary over i) and plug back into 6.3 to obtain

$$y_i = e_T \mu + (I_T \otimes \beta^{'} + e_T \lambda^{'}) x_i + [y_i - E[y_i \mid x_i]] = e_T \mu + \underset{T\times KT}{\Pi} x_i + \upsilon_i \tag{6.4}$$

We can obtain $\widehat{\Pi}$ by gigantic OLS and impose the restrictions on Π using MD.[41] We only need to assume x_{it} are *iid* for $t = 1, \dots, T$. Further, we do not need to assume $E[\alpha_i \mid X_i]$ is linear, but can treat $\mu + \lambda^{'} X_i$ as a projection, so that the error term υ_i is heteroscedastic.

Exercise 6.1. *Note how having two data dimensions is the key. In particular, try to implement this approach in cross-section data.*

[38] So that $(N - T^2 K)$ is large.

[39] The omitted variable has to be either time-invariant or individual-invariant.

[40] Here, *vec* is the vector operator stacking columns of matrices on top of each other into one long vector. We provide the definition and some basic algebra of the *vec* operator in class.

[41] How many underlying parameters are there in Π? Only $K + KT$.

Remark 28. *Hsiao's formulae (3.8.9.) and (3.8.10.) do not follow the treatment in (3.8.8.), but use time varying intercepts.*

6.3.3. Testing the Fixed Effects Model

Jakubson (1991): In estimating the effect of unions on wages we face the potential bias from unionized firms selecting workers with higher productivity. Jakubson uses the fixed effect model and tests its validity. We can use the MD framework to test for the restrictions implied by the typical fixed effect model. The MD test is an omnibus, all-encompassing test and Jakubson (1991) offers narrower tests of the fixed effect model as well:

- The MD test: Assume

$$y_{it} = \beta_t x_{it} + \epsilon_{it} \text{ with } \epsilon_{it} = \gamma_t \alpha_i + u_{it}$$

  where α_i is potentially correlated with $x_i \in \{0,1\}$[42]. Hence specify $\alpha_i = \lambda' \underset{T\times k}{x_i} + \xi_i$. Now, if we estimate

$$y_i = \underset{T\times T}{\Pi} x_i + \nu_i$$

  the above model implies the non-linear restrictions $\Pi = diag(\beta_1, \ldots, \beta_T) + \gamma\lambda'$ which we can test using MD. If H_0 is not rejected, we can further test for the fixed effect model, where $\beta_t = \beta \ \forall t$ and $\gamma_t = 1 \ \forall t$.

- Test against particular departures:[43]

  - Is differencing valid? Substitute for α_i to get

$$y_{it} = \beta_t x_{it} + (\frac{\gamma_t}{\gamma_{t-1}})y_{it-1} - (\beta_{t-1}\frac{\gamma_t}{\gamma_{t-1}})x_{it-1} + [u_{it} - (\frac{\gamma_t}{\gamma_{t-1}})u_{it-1}]$$

    Estimate overparametrized model by 3SLS with x as an IV for lagged y, test exclusion restrictions (see Remark 28), test $(\frac{\gamma_t}{\gamma_{t-1}}) = 1$ (does it make sense to use Δy_{it} on the left-hand side?), if valid test $\beta_t = \beta \ \forall t$.

[42] If α_i is correlated with x_{it} then it is also correlated with x_{is} $\forall s$.

[43] These tests are more powerful than the omnibus MD test. Further, when MD test rejects H_0 then the test against particular departure can be used to point to the *source* of misspecification.

– Is the effect "symmetric"?

$$\Delta y_{it} = \delta_{1t} ENTER_{it} + \delta_{2t} LEAVE + \delta_{3t} STAY + \Delta\mu_{it}$$

– Does the effect vary with other Xs?

Remark 29. *In the fixed effect model we rely on changing x_{it} over time. Note the implicit assumption that union status changes are random.*

7. Simultaneous Equations and IV Estimation

Simultaneous Equations are unique to social science. They occur when more than one equation links the same observed variables. Reverse causality. Identification.

Solution: IV/GMM: find variation in the X that suffers from simultaneity bias which is not related to the variation in the ϵs, i.e., use $\widehat{X}$ instead—the projection of X on Z—the part of variation in X that is generated by an instrument.[44] Theory or intuition is often used to find an "exclusion restriction" postulating that a certain variable (an instrument) does not belong to the equation in question. We can also try to use restrictions on the variance-covariance matrix of the structural system errors to identify parameters which are not identified by exclusion restrictions.

Example 7.1. *Consider the demand and supply system from Econometrics I (and of Haavelmo):*

$$\begin{aligned} q_D &= \alpha_0 + \alpha_1 p + \alpha_2 y + \varepsilon_D \\ q_S &= \beta_0 + \beta_1 p + \quad + \varepsilon_S \\ q_D &= q_S \end{aligned}$$

where S stands for supply, D stands for demand and p is price and y is income. We solve for the reduced form

$$\begin{aligned} p &= \pi_1 y + \upsilon_p \\ q &= \pi_2 y + \upsilon_q \end{aligned}$$

[44] Another intuitive approach is to model the unobservables directly (as the residuals in the first stage regression) and include them as an explanatory variable into the main equation. This is the so called *control function* approach (CF). In cases where the endogenous variable enters linearly, the CF approach and 2SLS are equivalent. However, CF is advantageous for models that are non-linear in endogenous variables (even if linear in parameters) and especially for models that are non-linear in parameters. See also Section 12.2.2.

and note that one can identify β_1 by instrumenting for p using y which is excluded from the demand equation. Here we note that in exactly identified models like this the IV estimate $\widehat{\beta_1} = \frac{\widehat{\pi}_2}{\widehat{\pi}_1}$ (show this as an exercise); this is called indirect least squares and demasks IV. To identify α_1 estimate Ω , the variance-covariance matrix of the reduced form, relate the structural and reduced form covariance matrices and assume $COV(\varepsilon_D, \varepsilon_S) = 0$ to express α_1 as a function of β_1.

Remark 30. *Here, we think of 2SLS as a simultaneous equation system, where both equations come from one economic model. A very different approach is when we have one structural equation with an endogeneity problem and we find a "natural" or controlled experiment" (see Sections 1 and 13) to use only the exogenous portion of the variation in the variable of interest (Deaton, 2009, NBER WP no. 14690).*

Along the same divide, one can think of identification as either corresponding to our ability to go from reduced-form to structural parameters (within an economic model) or to the (potential) presence of an experiment that could answer the question posed. For some questions, there are no counterfactuals, no available interpretation of the results of an experiment.[45]

Remark 31. *Of course, next to simultaneity (and reverse causality), the other two sources of bias that are dealt with using IV strategies are measurement error and omitted variable bias.*

A good instrument Z must be correlated with the endogenous part of X (in the first-stage regression controlling for all exogenous explanatory variables!)[46] and it must be valid, i.e., not correlated with ε. The next two sections discuss testing of each desired IV property.[47]

[45] For example, answering whether starting school at seven is better than doing so at six is impossible when using elementary school tests as the outcome variable because it is not possible to disentangle the effect of maturation from the effect of schooling. One can aswer the question with data on adults, where age no longer has an effect on ability (Angrist and Pischke, 2009, p.6).

[46] See [W] exercise 5.11. The key is that the prediction error from the first stage, which appears in the residual of the second stage, is not made orthogonal to the other exogenous variables by means of the first stage projection.

[47] See Murray (2006, J of Economic Perspectives) for an easy-to-read overview of the material covered in this section; he provides several interesting and useful examples from applied work that are not used here.

Note that the lack of IV-ϵ correlation corresponds to two underlying statements: (a) the IV is as good as randomly assigned (conditional independence assumption, traditionally referred to as *exogeneity*), and (b) the IV operates through a single known causal channel (the IV is correctly excluded from the main equation with y on the left-hand side, *exclusion restriction*). (Deaton, 2009, prefers to call IVs that satisfy the conditions (a) and (b) external and exogenous, respectively.)

Remark 32. *Often, IVs are derived from "natural experiments".*[48] *These "natural" IVs are clearly not affected by the y variables and so are exogenous (external in Deaton's terms, there is no simultaneity), which is, however, only a necessary condition for being correctly excluded (exogenous in Deaton's terms)! The question (assumption) is whether these IVs have any direct effect on y (other than through their effect on the endogenous x).*

Remark 33. *In 2SLS, consistency of the second-stage estimates does not depend on getting the first-stage functional form right (Kelejian, JASA 1971; Heckman, Econometrica 1978). In other words, a simple linear probability model for an endogenous dummy is sufficient in the first stage to obtain consistent estimates in the second-stage regression in 2SLS.*

Exercise 7.1. *If the endogenous variable enters in a quadratic form, one needs two instruments (such as the original instrument and its square, if it's not a dummy) and two first-stage regressions, one for x and the other one for x^2. Show why using only one first stage and plugging in the square of the predicted endogenous variable could be a problem. As usual, consider that the difference between the original and the predicted values is part of the residual of the main equation.*[49]

Remark 34. *IVs are often group specific. These work by averaging (eliminating) within-group heterogeneity (using the law of large numbers, similar to using group*

[48] Draft lottery predicts participation in the military, which drives earnings, Angrist, 1990; distance from the equator predicts per capita GDP, which drives religiosity, McCleary and Barro, 2006; rivers instrument for the number of school districts in explaining educational outcomes, Hoxby, 2000; month of birth as an instrument for years of schooling in an earnings regression, Angrist and Krueger, 1991; or rainfall as an instrument for economic growth in explaining civil war, Miguel, Satyanath, and Sergenti, 2004.

[49] Similarly, consider dealing with measurement error by using one noisy measure (proxy) of a true (latent) x as instrument for another proxy, see Exercise 5.4. In a model with interactions in two latent variables on the RHS, such as $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$, we need three IVs—for example z_1, z_2, and $z_1 z_2$. If only x_2 is measured with error, then we need only two IVs—z_2 and $z_2 x_1$.

averaging to reduce measurement error bias) and using only between group variation. Obviously, the cost is the loss of information within groups. This strategy stands in contrast to within-data estimation (fixed effect model) used in panel data settings, where we discard between group heterogeneity. In particular, with only one IV, $\widehat{\beta}_{IV} = COV_N(\widehat{X}, Y)/VAR(\widehat{X}) = COV_N(Z, Y)/COV(Z, X)$ *and when* Z *is binary, one can further write this as*

$$\widehat{\beta}_{IV} = \frac{E_N[Y|Z=1] - E_N[Y|Z=0]}{E_N[X|Z=1] - E_N[X|Z=0]},$$

where E_N *is a shorthand for sample mean.*[50] *This is also known as the Wald IV estimator.*

When there are many (exclusive) groups, the IV estimator can be expressed as consisting of all pairwise (Wald IV) group comparisons, i.e. consisting of all IV comparisons between each pair of groups (binary Wald IV estimators) and these pairwise Wald-IV estimators are efficiently (GLS) weighted in the grand 2SLS (see Angrist, 1991, Remark 8, Section 4.2, and Section 13.3).

7.1. Testing for exclusion restrictions and IV validity

When we have more instruments than endogenous variables, we can test for their validity using the Sargan's test.[51] This test asks if any of the IVs are invalid based on maintaining the validity of an exactly-identifying sub-set of the IVs. But which IVs belong to the valid subset? The test is problematic when all IVs share a common rationale (see Example 7.2 below). The test can never test whether *all* of the IVs are valid. In particular, not rejecting H0 in a test of over-identifying IVs (restrictions) is consistent with all instruments (restrictions) being invalid while rejecting H0 is consistent with a subset of IVs being correct!

Remark 35. *For a test of the validity of over-identifying exclusion restrictions, which we have already covered, see remark 26.*

Example 7.2. *For a simple test of an exclusion restriction, see Card (1993) who estimates returns to schooling using proximity to college as an instrument for education and tests for exclusion of college proximity from the wage equation. To do*

[50] This is based on $COV(Z, Y) = E[ZY] - E[Z]E[Y] = (E[ZY]/P(Z=1) - E[Y])P(Z=1) = \{E[Y|Z=1] - (E[Y|Z=1]P[Z=1] + E[Y|Z=0]P[Z=0])\}\, P(Z=1) = (E[Y|Z=1] - E[Y|Z=0])\, P[Z=1](1-P[Z=1]) = (E[Y|Z=1] - E[Y|Z=0])VAR[Z]$

[51] See stata `ivreg2`. See the next section on GMM for a general way of testing for overidentifying restrictions—the Hansen test.

this he assumes that college proximity times poverty status is a valid instrument and enters college proximity into the main wage equation. Notice that you have to maintain just identification to test overidentification. Also note that, unfortunately, the instrument, which is maintained as correctly excluded, in order to allow for this test, is based on the same economic argument as the first instrument, which is being tested.

Aside from econometric tests for IV validity (overidentification), one can also conduct intuitive tests when the exogenous variation (IV) comes from some quasi-experiment. Chiefly, this consists of asking whether there is a direct association between the instrument and the outcome in samples where there was never any treatment.

Example 7.3. *For example, Angrist in the 1990 Vietnam-era draft lottery paper asks if earning vary with draft-eligibility status for the 1953 cohort, which had a lottery, but was never drafted.*

Example 7.4. *Altonji, Elder and Taber (2005, JHR) provide several evaluations of IVs. They study the effect of studying in a Catholic (private) high school, CH, (as opposed to a public one) on several education outcomes y. In this literature, people use the following IVs (Z_i): being Catholic (C), distance to a Catholic school (proximity, D), and their interaction ($C * D$). So run $y_i = \alpha CH_i + X_i^{'}\gamma + \epsilon_i$ and IV for CH. Let's focus on evaluating the first IV: C:*

1. *Ask what the direct effect of C is on y for those who attend public eighth grades, because this group of students never attends Catholic high schools regardless of their religion or proximity.[52] They find some direct effect, suggesting the IV is not valid.*

2. *Use an approach that guesses about the degree of selection on unobservables from the measured degree of selection on observables. We usually believe that unobservables and observables are correlated. Indeed, in their data, there are large gaps in observables in favor of Catholic students. They find that all of the IVs lead to substantial upward biases in 2SLS. The 2SLS estimates are implausibly large.*

[52] Let proj refer to the lest squares projection: $\text{proj}(C_i|X_i) = X_i^{'}\pi$ and $\text{proj}(CH_i|X_i, C_i) = X_i^{'}\beta + \lambda C_i$. Define $\widetilde{C}_i \equiv C_i - X_i^{'}\pi$. Then $\widehat{\alpha}_{IV} \xrightarrow{p} \alpha + \frac{COV(\widetilde{C}_i, \epsilon_i)}{\lambda VAR(\widetilde{C}_i)}$. Now, run a regression of y on C and X for those who have $\Pr(CH_i = 1) = 0$. The coefficient on C_i will converge to $\frac{COV(\widetilde{C}_i, \epsilon_i)}{VAR(\widetilde{C}_i)}$. So, obtain an estimate of the bias ψ by running $y_i = X_i^{'}\delta + [\widetilde{C}_i\widehat{\lambda}]\psi + \omega_i$.

3. *Finally, they not only ask about whether exclusion restrictions are correct but also compare power of the IV to that of non-linearity in identifying the coefficient of interest in non-linear models. Specifically, they compare 2SLS to bivariate Probits and find that non-linearities, rather than the IVs (exclusion restrictions), are the main sources of identification, which is why the bivariate Probits give more plausible estimates.*[53] *The lesson is to avoid non-linearities, since otherwise you may get an estimate seemingly based on an IV while all you're getting is actually based on the arbitrary distributional assumption.*

Remark 36. *What if there is some correlation between the instrument and the error term ϵ, i.e., what if the IV is imperfect? Nevo and Rosen (2008, NBER WP No. 14434) assume that (a) the IV-ϵ correlation is of the same sign as the correlation between the endogenous x and ϵ, and (b) that the IV is less correlated with ϵ than the x is. Based on these assumptions, they derive analytic bounds for the parameters.*

A separate class of IV models, referred to as **regression discontinuity** design, arises when the endogenous variable, such as assignment to some treatment, is (fully or partly) determined by the value of a covariate lying on either side of an (administrative) threshold. Such assignment may be thought of as a natural experiment. Assume that the covariate has a *smooth* relationship with the outcome variable, which can be captured using parametric or semi-parametric models, and infer causal effects from discontinuity of the conditional expectation of the outcome variable related to assignment to treatment, which was determined by the 'forcing' variable being just below or just above the assignment threshold.[54] The position of the forcing covariate just above/below the threshold plays the role of the IV here.[55]

[53] First, compare the Probits with and without IVs. (Strictly speaking, the bivariate Probit assumptions (linearity and normality) are enough to identify α even in absence of exclusion restrictions.) Second, enter in the second stage not $\Phi(X_i^{'}\beta + Z_i^{'}\lambda)$ but two separate predicted probabilities for CH_i holding either X_i or Z_i constant at their sample means. This is pretty informal, but it helps to focus on where the identification works: off the IV or not.

[54] Clearly, there is some need for 'local' extrapolation (there is 0 common support in terms of the terminology of Section 13.2), so one asssumes that the conditional regression function is continuous.

[55] When it predicts perfectly, we talk of *sharp* regression discontinuity. When the first-stage R2 is below unity, we talk of *fuzzy* regression discontinuity. See Hahn, Todd and VanderKlaauw (2001).

Exercise 7.2. *For specific IV testing (specification testing) of the regression discontinuity models, see Imbens and Lemieux (2007)*[56]*, Van der Klaauw (2008) and NBER WP No. 14723 by Lee and Lemieux.*

Example 7.5. *Angrist and Lave (1998) study of the class-size effect using the Maimonides rule: not more than 40 pupils per class. Class size is endogenous because of potential quality sorting etc. Assuming cohorts are divided into equally sized classes, the predicted class size is*

$$z = \frac{e}{1 + int[(e-1)/40]},$$

where e denotes the school enrollment. Note that in order for z to be a valid instrument for actual class size, one must control for the smooth effect of enrollment because class size increases with enrollment as do test scores.

Example 7.6. *Matsudaira (2009, JEcm) studies the effect of a school program that is mandatory for students who score on a test less than some cutoff level.*

Example 7.7. *Or think of election outcomes that were just below or just above 50%.*

7.2. Dealing with Weak Instruments

IV is an asymptotic estimator, unlike OLS. 2SLS is biased in small samples and it needs large samples to invoke consistency. Other then testing for $COV(Z,\varepsilon) = 0$, one needs to consider the *weak instrument* problem, that is make sure that $COV(X,Z) \neq 0$. Even a small omitted variable bias ($COV(Z,\varepsilon) \neq 0$) can go a long way in biasing $\widehat{\beta}$ even in very large samples if $COV(X,Z)$ is small because $p\lim \widehat{\beta} = \beta_0 + COV(Z,\varepsilon)/COV(X,Z)$. See [W]5.2.6.[57]

When IVs are weak (and many), 2SLS is biased (even in really big samples).[58] The finite sample bias of 2SLS is larger when instruments are weaker and when

[56] It is an NBER WP no. 13039 and also the introduction to a special issue of the *Journal of Econometrics* on regression discontinuity. You need to wait with reading this one until after we cover Section 13.

[57] The ratio may not be well approximated by the Normal distribution even in large samples when $COV(X,Z) \simeq 0$, but by Cauchy (the ratio of two normals centered close to zero).

[58] The bias of over-identified 2SLS with weak IVs goes towards the probability limit of the corresponding OLS (inconsistent itself thanks to endogeneity) as the first stage gets closer to 0. Just-identified 2SLS is approximately unbiased, but that's of little help since with a weak IV 2SLS will be very noisy.

there are more instruments (high over-identification), leading to a trade-off between efficiency and bias (Hahn and Hausman, 2002).[59] Furthermore, weak IVs also bias inference in 2SLS, making standard errors too small.

Remark 37. *With grouped-data IV, the IV will be weak if the groups are small so there is not enough averaging out of the unobservables.*

Two responses have been developed: (a) test for the presence of weak IVs so you don't use them, (b) devise estimation methods that work with weak IVs.

(a) Bound et al. (1995) suggest the use of F tests in the first stage to detect weak IVs. Also see Staiger and Stock (1997) who suggest that an F statistic below 5 suggests weak instruments See also Stock and Yogo (2005) for another test.[60]

(b) Even when IVs are weak, there is hope for correct inference (around wrong point estimates). Moreira (2003) develops a conditional likelihood ratio test, which overcomes the distortion in standard errors: he changes critical values of the test using an estimate of the first-stage-regression coefficients on instruments. He then calculates confidence intervals (around the 2SLS point estimates) as the set of coefficient values that would not be rejected at a given level of statistical significance. These need not be symmetric. This test statistic is robust to weak instruments.[61] The argument is that even with low first-stage F, it may be possible to obtain informative confidence intervals.[62] Of course, this method is mainly useful in the just-identified case (when you pick your best IV and report a specification that's not over-identified) or with low degrees of over-identification, as in these cases,

[59] According to Bound et al. (1995), the weak IV bias is proportional to $(k-2)\sigma_{\epsilon\nu}/NR^2\sigma_x^2$, where ν is the first-stage residual, k is the number of IVs, and R^2 is the unadjusted first-stage R-squared. However, this bias formula is based on approximations which may not work well with some weak IVs (Hahn and Hausmann, 2001; Cruz and Moreira, 2005).

[60] Pretesting for weak IVs changes the distribution of 2SLS (see our discussion of seqential testing in Section 2.2) as Andrews and Stock (2005) suggest. Also, see Cruz and Moreira (2005) for an argument that the ability to use standard inference in the second stage depends not only on the F from the first stage, but also on the degree of endogeneity of X. This is their motivation for developing the method described in point (b).

[61] The conditional test statistic is available in Stata (get the latest version). See Moreira and Poi (Stata Journal, 2003) and Cruz and Moreira (JHR, 2005) and Moreira, Andrews and Stock in Journal of Econometrics. For point estimators usable with weak IVs, see Andrews and Stock (2005) and the Fuller estimators 2 and 4 in `ivreg2`.

[62] However, the Moreira method is not asymptotically valid if one first decides to check the F of the first stage and proceeds with the Moreira method only when the F is not totally tragic. And one always does this. See Chioda and Jansson (2006) for a solution.

2SLS will not be very misleading in most cases. On the other hand, Limited information maximum likelihood (LIML) is approximately unbiased for over-identified models.[63]

Example 7.8. *Returning to Angrist and Krueger (1991) IV estimation of returns to education, Cruz and Moreira (2005) show that in some specifications with relatively low first-stage F statistics, the confidence intervals around the wrong 2SLS estimate is not too far off from the original range given by Angrist and Krueger (A-K). In the most interesting specification, they estimate a confidence interval that has the same lower bound as that given by A-K, but a much higher upper bound. They say this makes the correct β "likely to be larger" compared to that given by A-K.*

Remark 38. *A simple pre-testing approach is to estimate the so-called* the *reduced form regression of y on the exogenous x and on the instruments (while excluding the endogenous x). Whenever the reduced form estimates are not significantly different from zero, one presumes that "the effect of interest is either absent or the instruments are too weak to detect it" (Angrist and Krueger 2001; Chernozhukov and Hansen, 2008). This is so because the IV (2SLS) estimator can be written as the ratio of the reduced-form and first-stage coefficients.*

Remark 39. *Of course, at the end, you test whether there is any endogeneity affecting β. This is typically done by Hausman, which simplifies to an auxiliary regression test where we include the first stage predicted residual in the second stage and test its significance ([W] p.118)*[64]

Remark 40. *It is still possible to estimate IV if only the instrument Z and y are in one data set and the instrument and the endogenous X are available in another data set (Angrist and Krueger, 1992). In this case, Inoue and Solon (2005) suggest the use of the two-sample 2SLS estimator $\widehat{\beta}_{TS2SLS} = \left(\widehat{X}_1'\widehat{X}_1\right)^{-1}\widehat{X}_1'y_1$, where $\widehat{X_1} = Z_1\left(Z_2'Z_2\right)^{-1}Z_2'X_2$ instead of the two-sample IV estimator $\widehat{\beta}_{TSIV} =$*

[63] For a survey available solutions in highly over-identified cases, see Section 13 of the lecture notes to the Imbens/Wooldridge NBER Summer 07 course and the Flores-Lagunes (2007) simulation study.

[64] See also http://www.stata.com/statalist/archive/2002-11/msg00109.html .

$\left(Z_2'X_2/n_2\right)^{-1} Z_1'y_1/n_1$.[65] *See also Arellano and Meghir (1991) for a Minimum-Distance version.*

Remark 41. *Actually, this logic applies to OLS just as well. We can replace any of the sample averages in the OLS formula by other averages that have the right plim. For example, we can ignore missing observations on y when calculating the $X'X$ plim, use other data, "pseudo panels", etc.*

Remark 42. *A simple and probably powerful test of joint IV validity and non-weakness is proposed by Hahn and Hausman (2002, Econometrica). It consists of comparing the forward two stage estimate of the endogenous-variable coefficient to the inverse of the estimate from the reverse regression (the right hand side endogenous variables becomes dependent variable and the dependent variable from the forward regression becomes the new right-hand side variable) using the same instruments. It uses a more robust 2nd order asymptotic justification and compares two estimates of a structural parameter such that one can assess the economic magnitude of the departure from H0.*

8. GMM and its Application in Panel Data

Read at least one of the two handouts on GMM which are available in the reference folder for this course in the library. The shorter is also easier to read.

Theory (model) gives us population orthogonality conditions, which link the data to parameters, i.e., $E[m(X,Y,\theta)] = 0$. The GMM idea: to find the population moments use their sample analogues (averages) $\sum_{i=1}^{N} m(X_i, Y_i, \theta) = q_N$ and find $\widehat{\theta}$ to get sample analogue close to 0.

If there are more orthogonality conditions than parameters (e.g. more IV's than endogenous variables) we cannot satisfy all conditions exactly so we have to weight the distance just like in the MD method, and the resulting minimized value of the objective function is again χ^2 with the degrees of freedom equal to the number of overidentifying conditions. This is the so called **Hansen test** or J test or GMM test of overidentifying restrictions:

$$\widehat{\theta}_N^{GMM} = \arg\min\{q_N(\theta)' W_N q_N(\theta)\} \tag{8.1}$$

[65] With exact identification, TS-2SLS is asymptotically equivalent to TS-IV, but with overidentification TS-2SLS is more efficient because it implicitly corrects for differences in the empirical distributions of the instrumental variables between the two samples.

To reach χ^2 distribution, one must use the optimal weighting matrix, $\widehat{V(m)}^{-1}$, so that those moment conditions that are better estimated are forced to hold more closely (see Section 6.3.1 for similar intuition). A feasible procedure is to first run GMM with the identity matrix, which provides consistent $\widehat{\theta}$ and use the resulting $\widehat{\varepsilon}$s to form the optimal weighting matrix.

Remark 43. *GMM nests most other estimators we use and is helpful in comparing them and/or pooling different estimation methods.*

Example 8.1. *OLS:* $y = X\beta + \varepsilon$*, where* $E[\varepsilon|X] = 0 \Longrightarrow E[X'\varepsilon] = 0$ *so solve* $X'(y - X\widehat{\beta}) = 0$.

Example 8.2. *IV:* $E[X'\varepsilon] \neq 0$ *but* $E[Z'\varepsilon] = 0$ *so set* $Z'(y - X\widehat{\beta}) = 0$ *if* $\dim(Z) = \dim(X)$*. If* $\dim(Z) > \dim(X)$ *solve 8.1 to verify that here* $= \widehat{\beta}_{TSLS}$.

Example 8.3. *Non-linear IV:* $y = f(X)\beta + \varepsilon$*, but still* $E[Z'\varepsilon] = 0$ *so set* $Z'(y - f(X)\widehat{\beta}) = 0$.[66]

Example 8.4. *Euler equations:* $E_t[u'(c_{t+1})] = \gamma u'(c_t) \Rightarrow E_t[u'(c_{t+1}) - \gamma u'(c_t)] = 0$*. Use rational expectations to find instruments:* Z_t *containing information dates* t *and before. So* $E_t[Z_t'(u'(c_{t+1}) - \gamma u'(c_t))] = 0$ *is the orthogonality condition. Note that here* ε *is the forecast error that will average out to* 0 *over time for each individual but not for each year over people so we need large* T.

Example 8.5. *Maximum likelihood: Binary choice: Logit: Score defines* $\widehat{\beta}$*:*

$$\log L(\beta) = \log \prod_{n=1}^{N} \frac{\left(\exp(\beta' x_i)\right)^{y_i} 1^{1-y_i}}{1 + \exp(\beta' x_i)} = \beta' \sum_{n=1}^{N} x_i y_i - \sum_{n=1}^{N} \log\left(1 + \exp(\beta' x_i)\right)$$

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{n=1}^{N} x_i \left[y_i - \frac{\exp(\beta' x_i)}{1 + \exp(\beta' x_i)} \right] = \sum_{n=1}^{N} x_i \epsilon_i = 0$$

Example 8.6. *One can use GMM to jointly estimate models that have a link and so neatly improve efficiency by imposing the cross-equation moment conditions. For example, Engberg (1992) jointly estimates an unemployment hazard model (MLE) and an accepted wage equation (LS), which are linked together by a selection correction, using the GMM estimator.*

[66] See Kelejian (1971, JASA) for 2SLS with equations linear in parameters but non-linear in endogenous variables (such as a quadratic in X).

Remark 44. *GMM does not require strong distributional assumptions on ε like MLE. Further, when εs are not independent, the MLE will not piece out nicely, but GMM will still provide consistent estimates.*

Remark 45. *GMM is consistent, but biased in general as* $E[\widehat{\beta}_{IV}|X,Z] - \beta_0 = (Z'X)^{-1}E[\varepsilon|X,Z] \neq 0$ *because* $E[\varepsilon|X,Z] \neq 0$ *for some* Z *because* $E[\varepsilon|X] \neq 0$. *GMM is a* large *sample estimator. In small samples it is often biased downwards (Altonji and Segal 1994).*

Remark 46. *GMM allows us to compute variance estimators in situations when we are not using the exact likelihood or the exact* $E[y \mid x]$ *but only their approximations.*[67]

Remark 47. *The GMM test of over-identifying restrictions can often fail to reject in IV situations because the IV estimates are often imprecise. When IV estimates are precise, on the other hand, rejection may correspond not to identification failure, but, rather, to treatment effect heterogeneity (see Section 13.3).*

Example 8.7. *The GMM analogue to 2SLS with general form of heteroscedasticity is*

$$\widehat{\beta}_{GMM} = (X'Z\widehat{\Omega}^{-1}Z'X)^{-1}X'Z\widehat{\Omega}^{-1}Z'Y \tag{8.2}$$

and with panel data we can apply the White (1980) idea to estimate $\widehat{\Omega}$ *while allowing for any unconditional heteroscedasticity* and *for correlation over time* within *a cross-sectional unit:*

$$\widehat{\Omega} = \sum_{i=1}^{N} Z_i'\widehat{\varepsilon}_i\widehat{\varepsilon}_i{}'Z_i$$

where the $\widehat{\varepsilon}_i$ *comes from a consistent estimator such as homoscedastic TSLS.*[68]

Exercise 8.1. *Show that even with heteroscedastic errors, the GMM estimator is equivalent to TSLS when the model is exactly identified.*

Exercise 8.2. *Compare the way we allow for flexible assumptions on the error terms in the estimator 8.2 to the strategy proposed in section 6.3.2.*

---

[67] See section 5 of the GMM handout by George Jakubson in the library.

[68] For differences between the traditional 3SLS and GMM, see [W]8.3.5.

Example 8.8. *Nonlinear system of simultaneous equations. Euler equations.*

McFadden (1989) and Pakes (1989) allow the moments to be simulated: SMM (see Remark 55). Imbens and Hellerstein (1993) propose a method to utilize exact knowledge of some *population* moments while estimating θ from the *sample* moments: reweight the data so that the transformed sample moments would equal the population moments.

9. Dynamic Panel Data

Up to now, we relied mainly on the assumption of time-constant unobservables (α_i) to produce difference-in-differences estimates. But in many applications, this notion is not attractive. For example, when evaluating labor-market interventions, there is a famous temporary drop in y right before the treatment takes place (the Ashenfelter's dip, Dehejia and Wahba, JASA 1999). Abbring and van den Berg (2003) point out that when assignment to treatment (policy change) is anticipated, individuals may react in an attempt to explore the policy rules and such anticipatory behavior would render 'before' individuals with similar characteristics incomparable, invalidating the D-in-Ds design.

So, what if we have $y_{t-1,i}$ on the RHS? If we want to do both fixed effects and lagged y, we're probably asking too much of the data.[69] The problem is that after we get rid of α_i by first differencing, $\triangle\epsilon_{it}$ is correlated with $\triangle y_{i,t-1}$ because they both depend on $\epsilon_{i,t-1.}$ Solving this problem requires instrumenting with longer lags (having a long panel), but if there is too much serial correlation in $\epsilon_{i,t}$ (a realistic situation) there may be no consistent estimator.

See the Arellano and Bond (1991) method (in `Stata`) or Blundell and Bond (1998). Arellano and Bover (1995) propose a system estimator that jointly estimates the regression in levels and in differences in order to re-incorporate levels variation and reduce the likelihood of weak-instruments bias. See also the Arellano and Honore chapter in the Handbook of Econometrics on new IV methods for dynamic panel models.

[69]See Guryan (AER, 2004) for an argument that comparing the FEM (with no $y_{i,t-1}$) to a regression with no α_i but with $y_{i,t-1}$ provides a useful interval of estimates.

Part III
Qualitative and Limited Dependent Variables

10. Qualitative response models

Our usual regression methods are designed for a continuous dependent variable. In practice, we very often analyze a qualitative response - a discrete dependent variable. For example: decide to buy a car, quit a job, retire, move, work; choose among many alternatives such as how to commute to work; choose sequentially the level of education; influence the number of injuries in a plant, etc. While it was entirely plausible to assume that ε in our usual regression model with a continuous y had a continuous pdf, this assumption is not valid here. The usual $E[y \mid x]$ no longer does the job in those situations.

Most models are estimated by MLE which allows us to write down even very complicated models.[70] As a consequence, IV is not easily possible and panel data analysis is difficult. Further, heteroscedasticity or omission of an explanatory variable orthogonal to the included regressors cause bias unlike in the linear regression analysis![71] Since MLE crucially hinges on distributional assumptions, recent literature focuses on estimation methods not requiring specification of any parametric distribution.

10.1. Binary Choice Models

10.1.1. Linear Probability Model

In the Linear Probability Model we assume our usual linear regression even though $y_i \in \{0, 1\}$. As a consequence the interpretation of $E[y_i \mid x_i] = \beta' x_i$ being the probability the event occurs breaks down when $\widehat{\beta}' x_i \notin [0, 1]$.

[70] Also testing using the LR principle is very convenient.

[71] For example, Arabmazar and Schmidt (1981), give some examples of the asymptotic biases for the Tobit model, see section 11.1.

Exercise 10.1. *Show that given $E[\varepsilon_i] = 0$, the residuals ε_i which can take on only two values are heteroscedastic.*

The main advantage of the LPM is its ability to handle IV estimation easily (2SLS, see Remark 33).[72] Applying the LPM is perfectly valid asymptotically when the empirical $\widehat{y_i}$s are not close to 0 or 1 and in most applications, it is the preferred econometric model.[73] One should also allow the xs to enter as finite order polynomials, allowing for non-linearity, which will help with predicting out of admissible range.

Example 10.1. *Cutler and Gruber (1995) estimate the crowding out effect of public insurance in a large sample of individuals. They specify a LPM:*

$$Coverage_i = \beta_1 Elig_i + X_i \beta_2 + \varepsilon_i$$

Eligibility is potentially endogenous and also subject to measurement error. To instrument for $Elig_i$ they select a national *random sample and assign that sample to each state in each year to impute an average state level eligibility. This measure is not affected by state level demographic composition and serves as an IV since it is not correlated with individual demand for insurance or measurement error, but is correlated with individual eligibility.*

10.1.2. Logit and Probit MLE

The MLE methods transform the discrete dependent variable into a continuous domain using cumulative distribution functions. This is a natural choice as any $F(\cdot) \in [0, 1]$.

Assume existence of a continuous latent variable $y_i^* = \beta' x_i + u_i$ where we only observe $y_i = 1$ if $y_i^* > 0$ and $y_i = 0$ otherwise. Then for a symmetric $F(\cdot)$ we have

$$P[y_i = 1 \mid x_i] = P[u_i > -\beta' x_i] = 1 - F(-\beta' x_i) = F(\beta' x_i). \qquad (10.1)$$

[72] There are 2SLS procedures available for non-linear models (Achen 1986), but these require strong distributional assumptions (Angrist and Krueger, 2001; [W]). A method sometimes used here is a Probit with a *control function* approach (see note n. 44) based on joint normality of residuals in the first and second stage. See Rivers and Vuong (1988, JEconometrics) for Probit, Smith and Blundell (1986, Econometrica) for Tobit, and Blundell and Powell (2004, REStud) for semiparametric models with a control function approach to endogeneity.

[73] See Angrist and Pischke (2009) for a vigorous defense of linear regressions in any situation, including qualitative choice, limited-depedent-variable models or IV with heterogenous treatment effects.

Two common choices for $F(\beta' x_i)$ are $\Lambda(\beta' x_i) = \frac{\exp(\beta' x_i)}{1+\exp(\beta' x_i)}$ (logit) and $\Phi(\beta' x_i)$ (probit). The sample likelihood is then built under random sampling.[74]

Remark 48. *MLE maximizes the log-likelihood* $\mathcal{L}(\theta) = \sum_{i=1}^{N} \log f(x_i, \theta)$, *where* $f(x_i, \theta)$ *is the individual likelihood contribution, for computational convenience. It is a natural thing to do since*

$$E\left\{\mathcal{L}(\theta) - \mathcal{L}(\theta_0)\right\} \underset{iid}{=} nE\left\{\log\left[\frac{f(x_i,\theta)}{f(x_i,\theta_0)}\right]\right\} \underset{Jensen}{\leq} n\log\left[E\left\{\frac{f(x_i,\theta)}{f(x_i,\theta_0)}\right\}\right] = 0.$$

Therefore we construct a sample analogue to $E \log f(x_i, \theta)$ *and maximize w.r.t.* θ. *Random sampling guarantees that* $\frac{1}{N}\sum_{i=1}^{N} \log l(x_i, \theta)$ *converges to* $E \log f(x_i, \theta)$. *Hence, lack of independence will not be a problem if the marginals do not shift around, even though* $\mathcal{L}(\theta)$ *is no longer the right likelihood. Similar convergence property underlies the GMM.*

Remark 49. *Both models are suitable for non-linear optimization using the Newton-Raphson methods as the Hessian is always n.d.*

Remark 50. $\widehat{\beta}$*s from logit and probit are not directly comparable* ($\widehat{\beta}_{Logit} \simeq 1.6\widehat{\beta}_{Probit}$ *, see [M p.23]). Further, while* $\widehat{\beta}_{OLS} = \frac{\partial E[y_i|x_i]}{\partial x_i}$ *we need to find the probability derivatives for logits and probits, e.g.* $\widehat{\beta}_{Logit} \neq \frac{\partial P[y_i=1|x_i]}{\partial x_i} = \Lambda(-\beta' x_i)[1 - \Lambda(-\beta' x_i)]\widehat{\beta}_{Logit}$. *These derivatives will only most rarely be any different from the LPM coefficients, even when the mean probability is different from 0.5. (And this is probably true for Tobit marginal effects in many applications.)*

Remark 51. *Parametric methods (e.g., probit and logit) assume strict monotonicity and homoscedasticity.*[75]

Remark 52. *There are bivariate extensions in the SURE spirit ([G] 21.6).*[76]

[74] When one reject equality of coefficients, one cannot just interact all right-hand side variables with group dummies in the case of limited dependent variable models (as one would in the case of OLS) as this would lead to incorrect inference in presence of group-level unobservables. Williams (2009), “Using heterogeneous choice models to compare logit and probit coefficients across groups”, *Sociological Methods and Research*, 37:4, 531-59.

[75] There is a heterogeneity test for probit ([G]p.680).

[76] Also see section 5.2. of the GMM handout by George Jakubson in the library for an example with correlated probits and their univariate approximation.

Exercise 10.2. *Show that in Probit, one can only estimate* β/σ.

Exercise 10.3. *Estimates from binary response models are essentially WLS estimates: Find the corresponding GMM/IV interpretation for logit model using the FOC's of the MLE. Compare it to the corresponding probit expression and find the WNLLS interpretation for probit. Will they give the same answer as the MLE in small samples? Think of the intuition behind the size of the weight in the variance-covariance matrix as a function of* $x'\beta$.

Remark 53. *See Davidson and MacKinnon (1993) textbook, chapter 15.4 for a useful auxiliary regression connected to qualitative response models.*

Remark 54. *When you estimate 2SLS with an endogenous binary variable, it may be tempting to use Logit or Probit in the first stage and to plug in the first-stage predicted values. But this will only generate consistent estimates if the first-stage assumptions are exactly valid, which is unlikely. So just use LPM in the first stage. Only OLS makes absolutely sure that first-stage residuals are orthogonal to both fitted value and covariates (as the first-stage prediction error will, of course, show up in the residual of the main equation). See remark 33.*

10.1.3. The WLS-MD for Multiple Observations

([M] 2.8, [G] 21.4.6) Suppose we have n_i observations corresponding to x_i and that for m_i of them the event occurred. Then assume $\widehat{p_i} = \frac{m_i}{n_i} = p_i + u_i = \beta' x_i + u_i$ and correct for heteroscedasticity. For non-linear models we invert the *cdf* and we need a Taylor series expansion to find the form of heteroscedasticity.

Example 10.2. *For the logit model* $p_i = \Lambda(\beta' x_i)$ *and we get* $\Lambda^{-1}(p_i) = \ln \frac{p_i}{1-p_i} = \beta' x_i + u_i$.□

Exercise 10.4. *Show the WLS is a genuine MD.*

10.1.4. Panel Data Applications of Binary Choice Models

The usual suspects: Random and Fixed Effects. See [H] 7.

Random Effect Probit Probit does not allow the fixed effect treatment at all. Random effects model is feasible but has been difficult because of multidimensional integration. To prevent contamination of β's, we need to integrate the random effects α out. For MLE we must assume a particular distribution for α, say $g(\alpha)$ depending on parameters δ. Then allowing for correlation of α over time for the same person we can maximize the following with respect to both β and δ :

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \int \prod_{t=1}^{T} \Phi(\beta' x_{it} + \alpha)^{y_{it}} \left[1 - \Phi(\beta' x_{it} + \alpha)\right]^{1-y_{it}} dG(\alpha|\delta) \qquad (10.2)$$

Notice the multidimensional integral (each Φ is an integral inside the T-dimensional integral over α). We can simplify matters by assuming that the correlation of α between any two time periods is the same. Then we can look at each y_{it} and $\int P[y_{it} = 1 \mid x_{it}, \alpha_i] \; g(\alpha)d\alpha = P[y_{it} = 1 \mid x_{it}]$. For each y_{it} we then have a double integral.

Remark 55. *When we allow for general structure of $g(\alpha)$ we need the simulated method of moments (SMM) to evaluate the integrals* fast *(McFadden, 1988): When computing the $P[y_i = 1 \mid X_i] = P_i$ presents a formidable computational problem one solution is to use their unbiased estimates. To illustrate this method return back to a cross-section and consider the GMM interpretation of probit where $P_i = \Phi(\beta' x_i)$ (see exercise 10.3):*

$$0 = \sum_{i=1}^{N} (y_i - P_i) \frac{X_i \phi(x_i'\beta)}{P_i(1-P_i)} = \sum_{i=1}^{N} (\varepsilon_i) w_i.$$

Suppose P_i is hard to calculate. Solution? Use $w_i = X_i$, which will deliver inefficient but consistent estimates. You still need to evaluate the P_i inside the ε_i. To do this, let $I(\cdot)$ be the indicator function and consider

$$\Phi(\beta' x_i) = \int_{-\infty}^{\beta' x_i} \phi(s) ds = \int_{-\infty}^{\infty} \phi(s) I(s < \beta' x_i) ds = E_s[I(s < \beta' x_i)].$$

To simulate the integral generate R values $s_r \sim N(0,1)$ and evaluate $\frac{1}{R}\sum_{r=1}^{R} I(s_r < \beta' x_i)$ to obtain an unbiased estimate of P_i. (It's not consistent as long as R is finite so can't use it in the w_i.). To conclude, drive the simulated moment condition to 0.

Remark 56. *To allow (flexible) correlation between x_i and α_i we may follow Chamberlain (1980), but we now need the true regression function (see section 6.3.2) and a distributional assumption on the α equation error term.*

Remark 57. *There is a specific counterpart to random effects usable with the logit model: NP-MLE (Non-Parametric Maximum Likelihood, see Heckman and Singer, 1984, for such duration models). Simply approximate the $g(\alpha)$ with a discrete distribution. Estimate the points of support and the respective probabilities as part of your likelihood maximization. See Section 14.2.1 for an application of this approach.*

Conditional Fixed Effect Logit The motivation for a fixed effect model is similar as in panel data linear regression. In MLE the α_is are again consistent only with $T \longrightarrow \infty$. Since T is usually fixed and since MLE relies on consistency, the α_is must be swept out. But how do you "difference out" an additive element from a non-linear function?

Logit does allow for such a trick. Consider the T observations on y_{it} as one $T-$ variate observation y_i. The suggestion of Chamberlain (1980) is to maximize the conditional likelihood (see section 10.2.1) of y_i given $\sum_{t=1}^{T} y_{it}$ which turns out to remove the heterogeneity. Conditional on α_is we have independence over both i and t.

Exercise 10.5. *To verify this, write down the conditional likelihood contribution of $y_i^{'} = (0,1)$ when $T = 2$.*

Remark 58. *Again, use Hausman test to compare the fixed effect model with the $\alpha_i = \alpha$ simple pooled-data logit.*

Remark 59. *The conditional fixed effect logit is computationally cumbersome for $T \geq 10$.*

Exercise 10.6. *Explain why y_is with no change over time are not used in the estimation and show that observations with time constant xs are not used either.*

10.1.5. Relaxing the distributional assumptions

Parametric models of choice (like logit or probit) are inconsistent if the distribution of the error term is misspecified, including the presence of heteroscedasticity.

One can go fully non-parametric. Matzkin (1992): Let $E[y_i \mid x_i] = m(x) = F(h(x))$ and study the identification of h from F. This is the most general and least operational we can go.

Index models Cosslett (1981): max $\mathcal{L}(\theta)$ w.r.t. both β and $F(g(\beta, x_i))$, where $g(\cdot)$ is assumed parametric. Only consistency derived, but no asymptotic distribution. Further research on index models includes Ichimura (1993) with a $\sqrt{n}$ estimator and Klein and Spady (1993). All of these require ε and x to be independent.

Maximum rank estimators Manski's Maximum Score Estimator (1975, 1985) maximizes the number of correct predictions, is $n^{-1/3}$ consistent, and is in LIMDEP. The idea is based on $E[y_i \mid x_i] = F(\beta_0' x_i)$. Assume $F(s) = .5$ iff $s = 0$.[77] Then $\beta_0' x_i \geq (\leq) 0$ iff $E[y_i \mid x_i] \geq (\leq) .5$ and we use $\text{sgn}(\beta' x_i) - \text{sgn}(E[y_i \mid x_i] - .5) = 0$ as a moment condition.[78] Then

$$\widehat{\beta}_{MRE} = \arg \max_{s.t.\beta'\beta=1} \frac{1}{n} \sum_{i=1}^{n} \left[(2y_i - 1)\text{sgn}(\beta' x_i) \right] \tag{10.3}$$

Functionally related regressors are excluded by identification assumptions and $\widehat{\beta}_{MRE}$ is identified up to a scaling factor. Asymptotic distribution is not normal and not easy to use since variance is not the right measure of variation so we need to bootstrap. The method allows for conditional heteroscedasticity and generalizes to multinomial setting.

Smoothed MSE by Horowitz (1992) can be made arbitrarily close to $\sqrt{n}$ convergence. The idea is to smooth the score function to make it continuous and differentiable by using *cdf* in place of *sgn*.

Another method of maximizing correct predictions is based on the Powell's idea of comparing pairs of people.[79] Assume the model $y_i = d_i = 1\{x_i\beta + \epsilon_i > 0\}$ and assume ϵ_i independent of x_i (no heteroscedasticity), then $E[d_i - d_j | x_i, x_j] = E[d_i|x_i] - E[d_j|x_j] = F_\epsilon(x_i\beta) - F_\epsilon(x_j\beta) > 0$ iff $(x_i - x_j)\beta > 0$ so estimate β by maximum rank estimator such as

$$\max_{\beta} \sum_{i<j} sign(d_i - d_j) sign((x_i - x_j)\beta) \tag{10.4}$$

This, of course gets rid of the intercept, so Heckman (1990) proposed that in presence of exclusion restriction, one can get the intercept off those who have $p(d_i = 1)$ almost equal to one.

[77] The choice of the median can be generalized to any quantile.

[78] Note that $\text{sgn}(\cdot)$ is not invertible.

[79] See the discussion on selection the Powell's way below in Section 12.2.4

Finally, Sognian Chen (1999) uses the additional assumption of symmetry of the distribution of ϵ to allow for $\sqrt{n}$ estimation of the constant term. (All other semiparametric methods make for a slower rate for the constant even if they deliver $\sqrt{n}$ for the slope.) (You still need to normalize the scale.) He also allows for heteroscedasticity of a particular form: $f(\epsilon|x) = f(\epsilon|\widetilde{x})$ where $\widetilde{x}$ is a subset of x. Assuming f is symmetric implies that $E[d_i + d_j|x_i, x_j] = F_\epsilon(x_i\beta) + F_\epsilon(x_j\beta) > 1$ iff $(x_i + x_j)\beta > 0$ (draw a picture of f symmetric around 0 to see why). Note that sum does not get rid of the intercept. So, estimate something like

$$\max_{\beta} \sum_{i<j} sign(d_i - d_j - 1) sign((x_i + x_j)\beta). \tag{10.5}$$

10.2. Multinomial Choice Models

See McFadden (1984) for a summary of pioneering research and Pudney (1989), *Modelling Individual Choice*, [P], chapter 3, for discussion of the material in view of the underlying economic theory. ([M]2,3, [A]9, [G]21)[80]

10.2.1. Unordered Response Models

So far we talked about 0/1 decisions. What if there are more choices?

Example 10.3. *Choice of commuting to work, choice among occupations, purchasing one of many product brands, etc.*

We want to analyze *simultaneous* choice among m alternatives. The idea is to look at pairwise comparisons to some reference outcome:

$$\frac{p_j}{p_j + p_m} = F(\beta_j^{'}x) \Rightarrow \frac{p_j}{p_m} = \frac{F(\beta_j^{'}x)}{1 - F(\beta_j^{'}x)} = G(\beta_j^{'}x) \Rightarrow p_j = \frac{G(\beta_j^{'}x)}{1 + \sum_{k=1}^{m-1} G(\beta_k^{'}x)} \tag{10.6}$$

Remark 60. *Note that in our binary choice example ($m = 2$), we also started by defining $p_j(p_j + p_m)^{-1} = p/(p + 1 - p) = p = F(\beta_j^{'}x)$.*

[80] The Limdep v.7 (1995) manual discusses several extensions and examples of application.

Multinomial Logit (MNL) If $F(\beta_j^{'}x) = \Lambda(\beta_j^{'}x)$ then $G(\beta_j^{'}x) = \exp(\beta_j^{'}x)$ and the estimation does not require any integration. Simply define $y_{ij} = 1$ if person i chooses the j -th choice and, $y_{ij} = 0$ otherwise, and

$$\max_{\beta_1,\beta_2,\ldots,\beta_{m-1}} \log L = \sum_{i=1}^{N} \sum_{j=1}^{m} y_{ij} \log p_{ij}, \text{ where} \tag{10.7}$$

$$p_{ij} = \frac{\exp(\beta_j^{'}x_i)}{1 + \sum_{l=1}^{m-1} \exp(\beta_l^{'}x_i)} \text{ for } j = 1, ..., m-1 \text{ and } p_{im} = \frac{1}{1 + \sum_{l=1}^{m-1} \exp(\beta_l^{'}x_i)}.$$

Remark 61. *The FOCs again have the familiar GMM interpretation*

$$\sum_{i=1}^{N} (y_{ij} - \widehat{p}_{ij})x_i = 0 \text{ for } j = 1, ..., m-1$$

and again imply that if x consists only of a constant, the model predicts the actual frequencies (see exercise 10.3). This can be used to define a measure of fit based on comparing our log-likelihood with the benchmark that one would obtain by merely regressing the outcome on constants α_j.

Exercise 10.7. *Verify that the benchmark likelihood equals $\prod_{j=1}^{m} \left(\frac{N_j}{N}\right)^{N_j}$ where $N_j = \sum_{i=1}^{N} y_{ij}$.*

Exercise 10.8. *What happens in the commuting choice example when all males choose to drive?*

Remark 62. *To interpret the estimates $\widehat{\beta}$ we need the derivatives w.r.t. x_k (k-th element of x) even more as $\widehat{\beta}_{jk}$ shows up in p_l $l = 1, \ldots, m$. $\frac{\partial p_j}{\partial x_k} = p_j[\beta_{jk} - \sum_{s=1}^{m-1} p_s \beta_{sk}]$.*

Remark 63. *There is a utility maximization model of individual choice which leads to the multinomial logit, assuming additivity of disturbances $y_{ij}^{*} = V_j(x_i) + \epsilon_{ij}$ with $y_{ik} = 1$ iff $\forall j \neq k$ $y_{ik}^{*} > y_{ij}^{*}$ and assuming ϵ_j's are iid type I extreme value distribution.*[81] *$P[y_k = 1|x_i]$ corresponds to the joint occurrence of $V_k(x_i) + \epsilon_k >$*

[81] Because the difference of two random variables following type I extreme value actually follows logistic distribution. Of course, this is much simple with Normal distribution, where a difference is again Normal. See Multinomial Probit below.

$V_j(x_i) + \epsilon_j \ \forall j \neq k$, that is

$$P[y_k = 1|x_i] = \int \prod_{j \neq k} F(\epsilon_k + V_k(x_i) - V_j(x_i)) f(\epsilon_k) d\epsilon_k.$$

In class we show that this equals $\exp(V_k(x_i))/\sum_{j=1}^{m} \exp(V_j(x_i))$. Set $V_k(x_i) = x_i^{'}\beta_k$ to conclude.

Exercise 10.9. *Verify the derivatives in [M] p.36 and show $\frac{p_j}{p_k} = \exp[(\beta_j - \beta_k)^{'} x]$.*

McFadden's Conditional Logit So far we focused on the question of how individual characteristics influence the choice. Next, answer the question of how often will individuals choose a new alternative, i.e., express the probability of choice as a function of the *characteristics of the choice* k (as perceived by individual i), say z_{ik}, not necessarily the characteristics of the individual x_i.

$$P[y_i = k] = \frac{\exp(\beta^{'} z_{ik})}{\sum_{s=1}^{m} \exp(\beta^{'} z_{is})} \tag{10.8}$$

Remark 64. *Individual characteristics which do not change with the choice drop out unless we combine the two models, i.e., allow for both choice and personal characteristics.*

Exercise 10.10. *Show $\frac{p_j}{p_k} = \exp[(z_j - z_k)^{'}\beta]$.*

Remark 65. *The elimination by aspect model ([M]3.4) represents another way how to account for similarities between alternative choices.*

In the multinomial logit (MNL) model, the joint distribution of extreme value εs does not involve any unknown parameters and is therefore not capable of approximating a wide range of stochastic structures. Furthermore, the MNL assumes that disturbances are *independent* (see Remark 63). When there is correlation, consistency suffers. Consider for example the choice between a blue bus, a red bus and a train. Hence, multinomial logit conforms to the *IIA hypothesis* (independence from irrelevant alternatives). See exercise 10.9 which shows that $\frac{p_j}{p_k}$ does not depend on characteristics or even the existence of choices other than j or k. Hence an introduction of a new alternative means that all of the existing probabilities are reduced by the same amount, irrespective of the new choice degree of similarity to any of the existing ones. The model restricts the choice probabilities to share a uniform set of cross-elasticities.[82]

Inclusion of some potentially correlated alternatives can be tested with a typical Hausman test (Hausman and McFadden, 1984). Under H_0 : IIA, one can estimate a subset of the β_j parameters consistently but inefficiently by dropping the individuals who choose the potentially correlated alternatives. These $\widehat{\beta}_j s$ can then be compared to those estimated off the whole data set with all options. Of course, if IIA is violated, the latter will be inconsistent.

Remark 66. *In absence of some natural grouping of the alternatives, the choice of the subset to leave out is arbitrary and, hence, so is the test.*

Remark 67. *In study of social interactions (social capital, peer effects, herding behavior), one may want to condition on the group- or region-specific propensity of the particular choice. For example, will I be more likely to go to college when many of my schoolmates do; or will I be more likely to become self-employed when many of my neighbors do so. One then conditions on the predicted probability of a given choice in the group/neighborhood $E(y_g)$ (endogenous effect) together with group-specific z_g variables (contextual effects) and individual x_{ig}. In linear regressions, there is a formidable identification problem with estimating these social equilibria (see Manski's, 1995, "reflection problem" and Durlauf, 2002);[83] however, in non-linear models identification is easier (Brock and Durlauf, 2001). These group-specific propensities are correlated across choices and may absorb the correlation of ϵ across choices.[84] See also Brock and Durlauf (in press, JEcm).*

Multinomial Probit and Nested Logit Given how unattractive the IIA assumption of MNL is in this setting, there are two responses: Multinomial Probit and Nested Logit.[85]

[82] $\frac{\partial \log P[y_i=j|z]}{\partial \log z_k} = -P[y_i = k \mid z]\frac{\partial \beta' z_k}{\partial \log z_k}$. See Pudney, p. 118.

[83] Note that the regression $E[y|z,x] = \alpha + \beta' x + \gamma' z + \delta' E[y|z] + \lambda' E[x|z]$ has a social equilibrium: $E[y|z] = \alpha + \beta' E[x|z] + \gamma' z + \delta' E[y|z] + \lambda' E[x|z]$, which you can solve for $E[y|z]$ and plug back to show that the structural parameters are not identified.

[84] There is still selection bias from the choice of group membership.

[85] It is also possible to avoid the IIA assumption by applying the Random Coefficient Model of Section 4.2. Simply allow the βs inside the $V_j(x_i)$ of Remark 63 to vary across people as individual tastes imply that coefficients are correlated across choices with different characteristics. See McFadden and Train (2000) and the use of the Gibbs sampler (Section 17) for estimation. Another possibility is to follow the Heckman and Singer (1984) approach that we first mentioned in Remark 57 and that we cover in detail in Section 14.2.1.

Multinomial Probit (MNP) Unlike MNL, the MNP model allows for a full correlation structure with $\epsilon \sim N(0, \Sigma)$ and requires $m-1$ dimensional numerical integration. One has to impose normalization on the $m(m-1)$ free elements σ of the $m \times m$ matrix Σ.

With $m = 3$, the choice of the first alternative $P[y_i = 1|x_i]$ corresponds to the joint occurrence of $\eta_{12} \equiv \epsilon_1 - \epsilon_2 > V_2(x_i) - V_1(x_i)$ and $\eta_{13} \equiv \epsilon_1 - \epsilon_3 > V_3(x_i) - V_1(x_i)$. One can then derive the variance-covariance of the joint normal *pdf* of η_{12} and η_{13}, the 2x2 matrix $\widetilde{\Sigma}$, from the original σ elements. Finally,

$$P[y_i = 1|x_i] = \int_{-\infty}^{V_2 - V_1} \int_{-\infty}^{V_3 - V_1} \frac{1}{2\pi\sqrt{\left|\widetilde{\Sigma}\right|}} \exp\left[-\frac{1}{2}(\eta_{12}, \eta_{13})' \widetilde{\Sigma}^{-1} (\eta_{12}, \eta_{13})\right] d\eta_{12} d\eta_{13}.$$

The likelihood requires $m-1$ dimensional numerical integration, numerical 1st and 2nd derivatives and is therefore potentially messy with $m > 3$. The Simulated Method of Moments of Remark 55 is used here; see, e.g. Geweke, Keane and Runkle (1994).

Nested Logit Alternatively, the independence assumption of MNL can be relaxed using the generalized extreme value (GEV) models ([M]3.7). The GEV distribution generalizes the independent univariate extreme value *cdf*s to allow for ε correlation across choices:

$$F(\varepsilon_1, \varepsilon_2, ..., \varepsilon_m) = \exp\left[-G(\exp(-\varepsilon_1), ..., \exp(-\varepsilon_m))\right], \tag{10.9}$$

where the function G is such that F follows properties of (multinomial) *cdf*. The GEV approach has been widely used in the context of the nested multinomial logit model (see section 10.2.1).

Example 10.4. *With $G(a_1, a_2, ..., a_m) = \sum a_m$ we obtain the simple MNL model. With $m = 2$ and*

$$G(a_1, a_2) = \left[a_1^{\frac{1}{1-\sigma}} + a_2^{\frac{1}{1-\sigma}}\right]^{1-\sigma}$$

we can interpret the σ parameter as correlation. In this case

$$P[y_i = j|x_i] = \frac{\exp(\frac{V_j(x_i)}{1-\sigma})}{\exp(\frac{V_1(x_i)}{1-\sigma}) + \exp(\frac{V_2(x_i)}{1-\sigma})}$$

where V_j is the valuation of choice j (see Remark 63).

Our goal is to (a) study the use of the multinomial logit model in tree structures, and (b) use GEV to allow for departure from IIA *within* groups of alternatives, whilst assuming separability between groups.

Example 10.5. *Choice of house: choose the neighborhood and select a specific house within a chosen neighborhood. Choose to travel by plane, then choose among the airlines. Allow for unobservables to be correlated within neighborhoods.*

(a) In presence of a nested structure of the decision problem we assume the utility from house j in neighborhood i looks as follows: $V_{ij} = \beta' x_{ij} + \alpha' z_i$, where z_i are characteristics of neighborhoods and x_{ij} are house-specific characteristics. To facilitate estimation when the number of choices is very large but the decision problem has a tree structure, we use $p_{ij} = p_i p_{j|i}$,[86] where as it turns out $p_{j|i}$ only involves β but not α:

$$p_{j|i} = \frac{\exp(\beta' x_{ij} + \alpha' z_i)}{\sum_{n=1}^{N_i} \exp(\beta' x_{in} + \alpha' z_i)} = \frac{\exp(\beta' x_{ij})}{\sum_{n=1}^{N_i} \exp(\beta' x_{in})}. \tag{10.10}$$

Similarly,

$$p_i = \sum_{n=1}^{N_i} p_{ij} = \frac{\exp(I_i + \alpha' z_i)}{\sum_{m=1}^{C} \exp(I_m + \alpha' z_m)}, \text{ where } I_i = \log \left[\sum_{n=1}^{N_i} \exp(\beta' x_{in}) \right] \tag{10.11}$$

is the so-called inclusive value (the total contribution of each house in a neighborhood). One can therefore first estimate β off the choice within neighborhoods (based on $p_{j|i}$) and then use the $\widehat{\beta}$ to impute $\widehat{I}_i$ and estimate α by maximizing a likelihood consisting of p_i. This sequential estimation provides consistent estimates, but MLE iteration based on these starting values can be used to improve efficiency. If MLE gives different results it suggests misspecification.[87]

Remark 68. *The assumed forms of utility functions can differ across branches and decisions.*

Remark 69. *The NMNL gives identical fits to data as the hierarchical elimination by aspect model.*

[86] Of course $p_{ijk} = p_i p_{j|i} p_{k|i,j}$.

[87] The likelihood is no longer globally concave in all parameters. For estimation methods see [MF]p.1426.

(b) Next, we use generalized extreme value distribution to allow for correlation of the disturbances. Start off by assuming stochastic utility maximization along the lines of Example 63 but assume GEV instead of type I extreme value. This will lead to a generalization of the NMNL model that actually nests independence (and generalizes to multivariate setting):

$$p_i = \frac{\exp[(1-\sigma)I_i + \alpha' z_i]}{\sum_{m=1}^{C} \exp[(1-\sigma)I_m + \alpha' z_m]} \tag{10.12}$$

Here one can test for within-neighborhood correlation by asking whether $\widehat{\sigma} = 0$.

Remark 70. *Zanella (2007, JEEA) extends the social interactions literature (see Remark 67) by applying the nested logit structure to a random utility framework in order to build a model with social interactions* and *endogenous choice of the group membership (neighborhood). The micro-founded theory then suggests econometric identification strategies.*

10.2.2. Ordered Response Models

Example 10.6. *Ratings, opinion surveys, attained education level.* $0 < 1 < 2$ *but* $1 - 0 \neq 2 - 1$.

Use threshold "constants" to split the range of ϵs. A common β affects the decision among many alternatives.

Example 10.7. *With* 3 *ordered choices assume that the latent* $y_i^* = -\beta' x_i + u_i$. *Then (i)* $y_i = 1$ *if* $y_i^* < 0 \Leftrightarrow u_i < \beta' x_i$, *(ii)* $y_i = 2$ *if* $y_i^* \in (0, c) \Leftrightarrow \beta' x_i < u_i < \beta' x_i + c$, *(iii)* $y_i = 3$ *if* $y_i^* > c \Leftrightarrow \beta' x_i + c < u_i$, *where* c *is another parameter to be estimated. Following the usual logic the likelihood is based on a product of individual* i *contributions, which depend on choice:* $P[y_i = 1|x_i] = F(\beta' x_i)$ *while* $P[y_i = 2|x_i] = F(\beta' x_i + c) - F(\beta' x_i)$ *and* $P[y_i = 3|x_i] = 1 - F(\beta' x_i + c)$.

The model generates to multinomial settings. Interpreting the coefficients based on their sign (!) is *not* obvious in the ordered response model (see [G] p.738).

10.2.3. Sequential Choice Models

These models have a much richer set of coefficients than the ordered response models. They arise naturally when decisions take place at different points in time (e.g. choice of education level).

In the simplest case assume independence of disturbances and estimate the model using a sequence of independent binary choice models. (In doing so, one places severe restrictions on the underlying preferences and opportunity sets.) On the other hand, they have been used to lower the computational burden of simultaneous choice among m alternatives with correlated disturbances.

Example 10.8. *First, choose to graduate from high school or not (this occurs with probability $1-F(\beta_H^{'}x_i)$); if you do then choose to go to college ($F(\beta_H^{'}x_i)F(\beta_C^{'}x_i)$) or not ($F(\beta_H^{'}x_i)[1-F(\beta_C^{'}x_i)]$). Note that the likelihood can be optimized separately with respect to β_H and β_C – we can run two separate logit/probit likelihoods, one over the choice of high school, the other over the choice of college (for those who did graduate from high school).*

In the most advanced case of modelling intertemporal choice under uncertainty, it is more satisfactory to use dynamic programming techniques.

10.3. Models for Count Data

Example 10.9. *Number of accidents in a given plant. Number of visits to a doctor.*

The essential limiting form of binomial processes is **Poisson** distribution: $P[y=r]=\exp(-\lambda)\lambda^r(r!)^{-1}$. Assume the number of accidents in each plant follows Poisson with plant-specific parameter λ_i and that these processes are independent across plants. To bring in $x'\beta$ assume $\ln\lambda_i=\beta^{'}x_i$ and maximize the likelihood:

$$\max_{\beta} L=\prod_i \exp(-\lambda_i)\lambda_i^{y_i}(y_i!)^{-1}. \tag{10.13}$$

However, Poisson is restrictive in many ways: First, the model assumes independence of number of occurrences in two successive periods. Second, the probability of occurrence will depend on time length of interval. Third, the model assumes the equality of mean and variance:

$$E[y_i \mid x_i]=V[y_i \mid x_i]=\lambda_i=\exp(\beta^{'}x_i) \Longrightarrow \frac{\partial E[y_i \mid x_i]}{\partial x_i}=\lambda_i\beta \tag{10.14}$$

The last assumption is relaxed by the **Negative Binomial** extension of Poisson, which allows for overdispersion: Let $\ln \lambda_i = \beta' x_i + \epsilon$ where $\epsilon \sim \Gamma(1, \alpha)$. Integrate the ϵ out of likelihood before maximization (as in Random Effect Probit) and maximize w.r.t. both β and α, the overdispersion parameter.[88]

See LIMDEP manual, section 26.2, for extensions of both models to panel data, censoring and truncation, the zero-inflated probability (see immediately below) and sample selection (see section 12).

10.4. Threshold Models

Combine a binary choice model with other likelihoods. In case of the count data estimation this approach has been coined as the zero inflated probability model:

Consider the example of accident counts in plants. The zero-inflated version allows for the possibility that there could not be any accidents in plant i: When we observe $y_i = 0$, it can either correspond to our usual Poisson data generating process where out of luck, there were no accidents in the given time period or it can correspond to a plant where the probability of having an accident is zero (in that event $Z_i = 1$): that is we can express $P[y_i = 0|x_i]$ as

$$\begin{aligned} P[0|x_i] &= P[Z_i = 1|x_i] + P[Z_i = 0|x_i] P[y_i^* = 0|x_i] = \\ &= F(\gamma' x_i) + (1 - F(\gamma' x_i)) \exp(-\lambda_i) \lambda_i^{y_i} (y_i!)^{-1}. \end{aligned} \tag{10.15}$$

Ideally, there is at least one variable affecting Z, but not y_i^* to aid identification.

[88] $V[y_i \mid x_i] = E[y_i \mid x_i](1 + \alpha E[y_i \mid x_i])$

11. Limited Dependent Variables

See [W] 16, [M]6, [G]22, [P]4. Let's combine qualitative choice with continuous variation.

Example 11.1. *Wage data censored from above at the maximum social security contribution level.*

Example 11.2. *Zero expenditure and corner solutions: labor force participation or R&D expenditure. A change in the xs affects both the usual intensive margin and the extensive margin of the corner solution.*

Note the fundamental difference between the two examples.

The difference between censoring and truncation, which both have to do with thresholds on observable ys, is in observing the xs for the censored values of ys.

Technical Preliminaries

(a) Means of truncated distributions:

$$E[\varepsilon \mid \varepsilon \geq c] = \int_c^\infty \frac{\varepsilon f(\varepsilon)}{1 - F(c)} d\varepsilon \tag{11.1}$$

Exercise 11.1. *Show that if $\epsilon \sim N(\mu, \sigma^2)$ then $E[\varepsilon \mid \varepsilon \geq c] = \mu + \sigma\lambda(\frac{c-\mu}{\sigma})$, where $\lambda(\cdot) = \frac{\varphi(\cdot)}{1-\Phi(\cdot)}$ is the so called inverse of the Mills' ratio.*[89] *Also find $V[\varepsilon \mid \varepsilon \geq c]$.*

(b) Means of censored distributions, where $\varepsilon^c = \max\{c, \varepsilon\}$:

$$E[\varepsilon^c] = F(c)c + [1 - F(c)]E[\varepsilon \mid \varepsilon \geq c] \tag{11.2}$$

11.1. Censored Models

Example 11.3. *When actual income is above \$ 100 000, the reported income is \$ 100 000.*

The structural model is using the concept of an underlying latent variable y_i^* :

$$\begin{aligned} y_i^* &= \beta' x_i + u_i \text{ with } u_i \sim N(\mu, \sigma^2) \\ y_i &= y_i^* \text{ iff } y_i^* > c \\ y_i &= c \text{ iff } y_i^* \leq c \end{aligned} \tag{11.3}$$

[89] Using the fact that $\int \varepsilon\phi(\varepsilon)d\varepsilon = -\int d\phi(\varepsilon) = -\phi(\varepsilon)$

Tobit Model When the data are censored, variation in the observed variable will understate the effect of the regressors on the "true" dependent variable.[90] As a result, OLS will typically result in coefficients biased towards zero.

WLOG[91] suppose the threshold occurs at $c = 0$. OLS is inconsistent no matter whether we include or exclude the zero observations because $E[\widehat{\beta_{OLS}} - \beta]$ depends on the truncated expectation of u_i in either case.

Exercise 11.2. *Characterize the bias of the OLS estimator when applied only to the nonzero y observations and show that OLS estimator when applied to all y observations is inconsistent (we are essentially omitting the explanatory variable $\lambda(x_i'\beta/\sigma)$).*

The Tobit likelihood is

$$L = \prod_{y_i^* > c} \frac{1}{\sigma}\varphi\left(\frac{y_i - x_i'\beta}{\sigma}\right) \prod_{y_i^* \leq c} \Phi\left(\frac{c - x_i'\beta}{\sigma}\right), \tag{11.4}$$

which has a single maximum, but two step procedures have been devised by Heckman ([M]8.2) and Amemiya ([M]6.5).

Remark 71. *The two step procedure of Heckman starts with a Probit on $y_i > 0$ or not. This delivers consistent $\widehat{\beta/\sigma}$. In the second step, bring in the continuous information and consider*

$$\begin{aligned} E[y_i|x_i] &= P[y_i^* > 0]E[y_i|y_i^* > 0] + P[y_i = 0]E[y_i|y_i = 0] = \\ &= \Phi\left(\frac{x_i'\beta}{\sigma}\right)x_i'\beta + \sigma\varphi\left(\frac{x_i'\beta}{\sigma}\right) + 0 = \Phi_i x_i'\beta + \sigma\varphi_i. \end{aligned}$$

Use the first-step $\widehat{\beta/\sigma}$ to predict $\widehat{\Phi}_i$ and $\widehat{\varphi}_i$ and estimate $y_i = \widehat{\Phi}_i x_i'\beta + \sigma\ \widehat{\varphi}_i$ for a new set of $\widehat{\beta}$ and $\widehat{\sigma}$. As usual, drastic differences between first- and second-step estimates signal misspecification.

Remark 72. *The likelihood is only piece-wise continuous.*

---

[90] Plot a graph with censored data to see that you get no change in y for changes in x for the censored values of y.

[91] Assuming there is a constant in x ([M]p.159).

Remark 73. *The estimator is biased in the presence of heteroscedasticity. See Arabmazar and Schmidt (1981) for the potential magnitude of the bias. See Koenker and Bassett (1982) for quantile regression tests for heteroscedasticity. Pagan and Vella (1989) propose a test for heteroscedasticity when the dependent variable is censored. Need zero-expected-value residuals to construct the test. These can be obtained by a trimmed LS estimator (Powell 1986). See section 11.3 for recent heteroscedasticity-robust alternatives to Tobit such as CLAD.*

Of course, the model can be easily extended to censoring from above *and* below. The model also has extensions allowing for multiple and variable thresholds c_i, endogenous thresholds, heteroscedasticity,[92] panel data random effects, sample selection (see section 12), SEM, nested structures, and non-normality (see LIMDEP manual, Ch. 27).[93]

How do we interpret the coefficients? Depends on the type of data generating mechanism / question we ask. There are 3 types of predictions we can consider, using the definitions of $E[y^* \mid x]$ (for data censoring problems) and $E[y \mid x]$ and $E[y \mid x, y^* > 0]$ (for corner solution problems). But then we would not want to use a Tobit for a corner solution problem. Or other latent variable models for that matter, since what is the meaning of negative $y*$ latent health care expenditure? See also [W] 16.1, 16.5.[94]

Exercise 11.3. *Find the expressions for these 3 conditional mean functions and their derivatives w.r.t. x. $E[y \mid x]$ has a particularly simple form. What is the intuition? Marginal variation in x does not lead to anyone crossing the extensive margin. In the case of evaluating the effect of a dummy variable, the formula makes clear that the effect is much smaller than the corresponding parameter, which is intuitive given that the latent outcome always changes when the dummy switches, while the observed outcome stays the same (at 0) for many individuals.*

Remark 74. *There is little theoretical justification for Tobit in rational choice models (see [P]p.141).*

---

[92] The available methods use a parametric assumption on the form of heteroscedasticity. Semi-parametric estimators are a focus of much current research, see section 11.3.

[93] For a survey of Tobit specification tests see [P]4.1.5. For further reading see special issues of the *Journal of Econometrics* (84-1,86-1,87-1). One strand of tests is based on conditional moment restrictions, see [G]22.3.4d.

[94] Bauer and Sinning (2005 IZA DP; 2008 AStA) provide a Oaxaca-Blinder decomposition for Tobit and, more generally, for non-linear models.

Remark 75. *Under the strong assumption of joint normality of error terms, one can instrument in a Tobit.*

Grouped Data

Example 11.4. *Wages reported only in ranges, i.e.* $w_i \in [\$10000, \$20000)$, *i.e.* $w_i \in [c_{j-1}, c_j)$ $j = 1, \dots J$

The difference between this model and the ordered choice models is that the threshold values are known here. For $c_j = H$ and $c_{j-1} = L$ the likelihood contribution of observations with y_is in those ranges is

$$\ln L_{HL} = \sum_{i=1}^{N} \left\{ \ln[\Phi(\eta H - x_i^{'}\gamma) - \Phi(\eta L - x_i^{'}\gamma)] \right\}, \tag{11.5}$$

where $\gamma = \frac{\beta}{\sigma}$ and $\eta = \frac{1}{\sigma}$ (similar reparametrization of the likelihood is used in the estimation of the Tobit mode, see Olsen 1978). Use

$$E[y_i^* \mid x_i] = x_i^{'}\beta + \sigma \frac{\varphi_{iL} - \varphi_{iH}}{\Phi_{iH} - \Phi_{iL}} \tag{11.6}$$

for prediction. Again, the model can be extended to allow for sample selection.

11.2. Truncated Models

Example 11.5. *Only have data on low income households when studying the impact of variable* x *on income* y.

$$L = \prod_{y_i^* > c} \frac{1}{\sigma} \varphi \left(\frac{y_i - x_i^{'}\beta}{\sigma} \right) \left[1 - \Phi \left(\frac{c - x_i^{'}\beta}{\sigma} \right) \right]^{-1} \tag{11.7}$$

Tobit-type model is not feasible here as we do not observe xs for the $y = c$ observations. To evaluate the impact of x on y, in a simple truncation, use

$$E[y_i \mid x_i, y_i < c] = x_i^{'}\beta + \sigma \frac{\varphi(c/\sigma)}{\Phi(c/\sigma)}.$$

In a double truncation region, use Equation 11.6 for $E[y_i \mid c_{iL} \leq y_i \leq c_{iH}]$.

Finally, it is an opportune time, to note that the Tobit model is restrictive in constraining the coefficients and the xs affecting the extensive and intensive margins to be the same ([G]p.770).

Example 11.6. *Consider studying the impact of the age of a building on the cost from a fire in that building. In some buildings there is no fire and cost is zero, in other buildings you observe a fire and the associated cost. It is likely that older buildings are more likely to experience fire, while the cost of fire, conditional on having one, is likely to be higher in a newer building.*

We can relax the Tobit likelihood and split it into two (independent) parts: (i) 0/1 probit for whether there is a fire or not, and (ii) a truncated normal regression of the cost of fire estimated on those buildings where there was a fire. Further, we can allow different explanatory variables to enter each of the separate two likelihoods.

Remark 76. *Assuming the xs affecting both margins (equations) are the same, note that under the equality of coefficients, the relaxed two-part model boils down to the restricted Tobit model. Hence, the equality of coefficients is testable using an LR test:*

$$LR = -2\{\ln L_{PROB} + \ln L_{TRUNC} - \ln L_{TOBIT}\} \sim \chi^2(k) \text{ where } k = \dim(\beta).$$

Remark 77. *But the disturbances from the two separate equations are likely dependent, which is why we need a sample selection model!*

Remark 78. *Note that (without any covariates) a linear regression of an "expenditure" (corner solution) Y on a binary treatment would give the (unconditional) average treatment effect for expenditures, but a truncated regression (conditional on positive) would not have a causal interpretation in a randomized trial because the experiment changes the composition of the group with positive expenditures.*

11.3. Semiparametric Truncated and Censored Estimators

If the residual in a censored model is subject to heteroscedasticity of an unknown form or if we do not know the distribution of the ε for sure, then standard MLE will be inconsistent. Also, maximum likelihood estimation of censored panel-data fixed-effect models will be generally inconsistent even when we have the correct parametric form of the conditional error distribution (Honoré, 1992).

Below, we will continue to specify the regression function parametrically, but will try to do without assuming parametric distributions for ε. The estimators will alternate between additional "recensoring," which will compensate for the original censoring in the data, and a "regression" step using only the "trimmed" data part. For simplicity, consider only censoring or truncation from below at 0.

Symmetrically Trimmed Least Squares How can we estimate truncated or censored models without relying on particular distributional assumptions? Consider truncation from below at 0 in a model $y_i^* = x_i'\beta + \epsilon_i$. The idea of the estimator is to trim (truncate) the dependent variable *additionally* from above to make it symmetrically distributed. The new dependent variable will be symmetrically distributed around the regression function so we can apply least squares. But where do you trim from above? Depends on β. *Assume* that $f_\varepsilon(s|x)$ is symmetric around zero and unimodal. Then for $x_i'\beta > 0$, the ε is truncated at $0 - x_i'\beta$ so a symmetric truncation of ε is at $x_i'\beta - 0$. This corresponds to truncating y at $2x_i'\beta - 0$ (plot a distribution graph to see this point).

Powell's (1986) Symmetrically Trimmed LS is consistent and asymptotically normal for a wide class of symmetric error distributions with heteroscedasticity of unknown form. With data truncated from below, the estimator minimizes

$$\sum I\{x_i'\beta > 0\} I\{y_i < 2x_i'\beta\} \left[y - x_i'\beta\right]^2. \tag{11.8}$$

Alternatively, with censoring from below, apply the same idea (Symmetrically Censored LS) to minimize

$$\sum I\{x_i'\beta > 0\} \left[\min(y_i, 2x_i'\beta) - x_i'\beta\right]^2. \tag{11.9}$$

Censored Least Absolute Deviation [W] 16.6.4. Powell's (1984) CLAD is again based on *additional* censoring of y. The main idea is to look at median as opposed to mean, because median is not affected by censoring. (The main assumption of the estimator is therefore zero median of $F_\epsilon(s|x)$.) Median is not affected as long as we are in the uncensored part of the data. If we are below the censoring point, then the median does not depend on $x'\beta$. So: $median(y_i^*|x_i) = \max\{x_i'\beta, 0\}$. (Again, we work only with variation in the $x_i'\beta > 0$ area.) We note that medians are estimated using LAD.[95] The CLAD estimator is then found by minimizing

$$\sum \left| y_i - \max\{0, x_i'\beta\} \right|. \tag{11.10}$$

[95] Estimate δ, the median of z_i, by $\min \sum_i |z_i - \delta|$. LAD is not a least-squares, but a median (quantile) regression (these are in general more robust to outliers, see Section 16.5). Quantile regression fits a linear model for conditional quantiles, just as OLS fits a linear model for conditional means.

It alternates between deleting observations with estimates of $x_i'\beta$ that are outside of the uncensored region and estimating the median regression based on the remaining observations.

It can be used in a data censoring (Tobit) or corner solution (Heckman's λ of Section 12.2.2) setup and in presence of heteroscedasticity and it is even more robust than STLS. CLAD is programmed into Stata.[96] CLAD has a small sample bias opposite to the OLS bias; for a two-step remedy, see Khan and Powell (2001, JEcm). Honore, Khan and Powell (2002) allow the censoring thresholds to be not always observed. Blundell and Powell (2007, JEcm) allow for IVs.[97] Also see Newey and Powell (1990). More accessible treatment of related topics can be found in a book on *Methods of Moments and Semiparametric Econometrics for Limited Dependent Variable Models* by Myoung-jae Lee.[98]

12. Sample Selection

A focus of an enormous volume of empirical and theoretical literature. It involves features of both truncated and censored models. The treatment of the data where sampling depends on outcomes is different in cases where the variables determining selection are observable and when they are not (for an introductory discussion see [P]2.5).

We first consider situations where the econometrician (data collection agency) chooses to base sampling on y.

Example 12.1. *Consider a sample of families and estimate the impact of x on family income. However, the sample is such that low-income families are oversampled.*

Second, we consider situations where the individual behavior results in sample selection: situations where people/firms/etc. select themselves into different states based on potentially unobserved characteristics.

Example 12.2. *You can only measure the impact of x on wages (y) for those women who work (selection on y). Whether or not a woman works, depends on the wage she could get when working.*

---

[96] By D. Jolliffee, a former CERGE-EI faculty member, and by two of your co-students.

[97] http://www.ucl.ac.uk/~uctp39a/BlundellPowellCRQDec05.pdf

[98] Stata ado files for these semiparametric models can be downloaded from http://emlab.berkeley.edu/users/kenchay .

Example 12.3. *Average wage over the business cycle: seems flatter due to selective drop out of work.*

12.1. Sampling on Outcome

We first think about sampling based on qualitative choice and then on a continuous outcome variable.

12.1.1. Choice-based sampling

Consider a binary choice problem. To analyze a rare event when population probabilities $p(y_i) = P[y_i = 1]$ are tiny (training treatment, violent crime), we can decide to save money and instead of collecting a large random sample of the population (in order to have a decent number of $y_i = 1$ group members), we sample randomly within each y group to obtain $f(x_i \mid y_i)$. Then we note $f(x_i, y_i) = p(y_i \mid x_i)f(x_i) = f(x_i \mid y_i)p(y_i)$ and write the likelihood function for the two samples

$$L(\cdot) = \prod_{i \in S_1} f(x_i \mid y_i = 1) \prod_{i \in S_2} f(x_i \mid y_i = 0) \tag{12.1}$$

in terms of $p(y_i \mid x_i) = F(\beta' x_i)$ (in the bivariate example).[99]

Remark 79. $P[y_i = 1], P[y_i = 0]$ *usually come from a different data set (are known), but can be estimated as part of the problem.*

Next, consider a multinomial choice problem. Manski and McFadden (1981) set up an intuitive conditional maximum likelihood estimator using the formula for the conditional probability of i given x in the sample. For $j = 1, ..., M$ choices:

$$L(\theta) = \sum_{i=1}^{N} \ln \frac{p(y = y_i|x_i, \theta)H_{y_i}/p(y = y_i)}{\sum_{j=1}^{M} p(y_i = j|x_i, \theta)H_j/p(y_i = j)}, \tag{12.2}$$

where H_j is the probability of sampling from a strata j (can be unknown to the researcher).[100]

[99] See also Pudney (1989), chapter 3.2.3.

[100] However, as noted in Cosslett (1981) paper, this estimator is not efficient. Partly because the sample distribution of x actually depends on $\theta : g(x) = \sum_{s=1}^{S} H_s/p(s) \sum_{j \in I(s)} p(j|x, \theta)f(x)$. So we should use this information to help us estimate θ better. But then we do not get rid of $f(x)$, which was the beauty of 12.2.

Remark 80. *Replacing H with $\widehat{H}$ actually improves efficiency. Counterintuitive. Only works with an inefficient estimator.*[101]

To repeat: selecting "balanced" data with the same number of 0 and 1 outcomes and running a simple Logit is wrong, even though it is regularly done in practice. See Scott and Wild (1997) for a bias correction. There is much practical research on economically optimal sampling design, i.e., saving money on data collection and delivering efficiency. Some of the sampling designs combine sampling on the key right-hand-side variable with sampling on the outcome.[102]

12.1.2. Endogenous Stratified Sampling

It occurs when the probability that an individual is observed in the sample depends on the (continuos) y.[103]

Remark 81. *Stratification (over/undersampling) based on x variables presents no problem for OLS, as long as there is no parameter heterogeneity across strata (see Remark 8).*

Assume the final sample is obtained by repeated random drawings, with each draw being made from stratum i with probability $p_j = \frac{n_j}{Nj}$ which is independent of the xs. Here n_j is the number of observations in data from strata j and N_j

[101] Cosslett (1981) devised a pseudo-likelihood estimator, replacing the $f(x)$ with a set of discrete densities f_n. Counterintuitively, even though the number of parameters climbs with n, this estimator is efficient. (Think of estimating a mean this way.) However, it is not practical. So, Imbens (1992) comes up with a reparametrization of Cosslett moment conditions which is implementable. It is based on the intuition that to devise a moment condition based on x with many points of support, I do not need to know the points of support themselves (see the example below). He uses change in variables in the FOC (moment conditions) of the Cosslett estimator between a subset of the points of support of x and the population marginal densities p to come up with nice moment conditions.

Example: Suppose you want to estimate $\delta = \Pr(z > 0)$. If z is discrete with $\{z^1, z^2, ..., z^L\}$ points of support and unknown probabilities $\{\pi_1, \pi_2, ..., \pi_L\}$ one could efficiently estimate δ on the basis of $i = 1, ..., N$ independent observations of z_i by ML as $\widehat{\delta} = \sum_{l|z^l>0} \widehat{\pi}_l = \frac{1}{N}\sum_{n=1}^{N} I[z_n > 0]$ where the last representation of the estimator does not depend on the points of support. It can also be used when δ does not have a discrete distribution.

[102] This is a detailed Monte Carlo study of alternative sampling designs for Logit: http://www.occ.treas.gov/ftp/workpaper/wp2001-3.pdf

[103] Simple estimation using data split into subsamples based on the level of the dependent variable is a no-no thing in econometrics.

is the population size of strata j. Let y_{ij} denote the value of y for person i from strata j. The typical solution in practice is WLS:

$$\min \sum_{i,j} \frac{1}{p_j}(y_{ij} - x_{ij}^{'}\beta)^2,$$

which, however, only works asymptotically. In small samples it will be biased.

A potentially better solution is MLE. Consider an example of endogenous stratification (think oversampling or undersampling) with known threshold L and with 2 strata ($j = 1, 2$) of the level of y ([M]6.10.). Assume Normality and maximize a likelihood based on[104,105]

$$\begin{aligned} L(y_i|x_i) &= L_i^{-1} p_1 \phi((y_i - x_{ij}^{'}\beta)/\sigma) \text{ if } y_i < L \text{ and} \\ L(y_i|x_i) &= L_i^{-1} p_2 \phi((y_i - x_{ij}^{'}\beta)/\sigma)) \text{ if } y_i > L \text{ where} \\ L_i &= p_1 \Phi[(L - x_{ij}^{'}\beta)/\sigma] + p_2(1 - \Phi[(L - x_{ij}^{'}\beta)/\sigma]). \end{aligned} \tag{12.3}$$

In the next subsection, we will consider cases when truncation or censoring occurs with stochastic or unobservable thresholds and where individuals make the sample-inclusion choice, not the data collection agency.

12.2. Models with Self-selectivity

Now it is the individuals that we study who makes the sampling decision.

Example 12.4. *Fishing and hunting: the Roy's model ([M]9.1); workers choose their union status based on the wage "in" and on the wage "out"; labor force participation; returns to education; migration and income; effect of training programs, evaluation of social policy.*

There are two main types of models: first, when we do not observe the y under one choice and observe it under the other (labor force participation, Heckman's λ), second, when we observe y under all *chosen* alternatives (union wages, switching regression).

[104] The formulas in [M]6.10 are conditional on y_{ij} actually being drawn from a given strata j.
[105] See [Mp.173] for the asymptotic justification of WLS based on this MLE.

12.2.1. Roy's model

First consider a classical theory (paradigm) on the topic. A worker i chooses to either hunt or fish, depending on which of corresponding outputs y_{iH} and y_{iF} is larger. Note that we never observe both y_{iH} and y_{iF} for each worker, but only one of the two outcomes.[106]

Assuming that

$$\begin{pmatrix} y_{iH} \\ y_{iF} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_H \\ \mu_F \end{pmatrix}, \begin{pmatrix} \sigma_H^2 & \sigma_{HF} \\ \sigma_{HF} & \sigma_F^2 \end{pmatrix} \right),$$

one can show that

$$E[y_{iH}|y_{iH} > y_{iF}] = \mu_H + \frac{COV(y_{iH}, y_{iH} - y_{iF})}{\sqrt{V(y_{iH} - y_{iF})}} \frac{\phi(z)}{\Phi(z)}, \text{ where } z = \frac{\mu_H - \mu_F}{\sqrt{V(y_{iH} - y_{iF})}}.$$

In short, $E[y_{iH}|y_{iH} > y_{iF}] = \mu_H + \frac{\sigma_H^2 - \sigma_{HF}}{\sigma} \frac{\phi(z)}{\Phi(z)}$ and similarly for $E[y_{iF}|y_{iH} > y_{iF}]$. There are 3 possible cases of a Roy's economy:

1. If $\sigma_H^2 - \sigma_{HF} > 0$ and $\sigma_F^2 - \sigma_{HF} > 0$, those who hunt are better off than an average hunter (similarly for the fishermen). This is the case of absolute advantage.

2. When $\sigma_H^2 - \sigma_{HF} > 0$ and $\sigma_F^2 - \sigma_{HF} < 0$, those who hunt are better than average in both occupations, but they are better in hunting (comparative advantage).

3. Reverse of 2.

Remark 82. *Notice that individuals with better skills choose the occupation with higher variance of earnings. Also notice the importance of* $\sigma_{HF} \neq 0$ *(see Remark 77).*

Remark 83. *Note that* $\sigma_H^2 - \sigma_{HF} < 0$ *and* $\sigma_F^2 - \sigma_{HF} < 0$ *cannot happen due to Cauchy-Schwartz inequality.*

Remark 84. *Think of real-life applications when you really want to know the population-wide* μ_H.

---

[106] For purposes of policy evaluation we will need to deal with estimating the counterfactual. See subsection 13.

12.2.2. Heckman's λ

See Heckman (1980), [M]6.11, 8.4. Consider a two-equation behavioral model:

$$\begin{aligned} y_{i1} &= x_{i1}^{'}\beta_1 + u_{i1} \\ y_{i2} &= x_{i2}^{'}\beta_2 + u_{i2}, \end{aligned} \tag{12.4}$$

where y_{i1} is observed only when $y_{i2} > 0$.

Example 12.5. *Observe wages (y_{i1}) only for women who work ($y_{i2} > 0$).*

Note that the expectation of data on y_{i1} you observe depends on the selection rule which determines that y_{i1} is observable:

$$\begin{aligned} E[y_{i1}|x_i, y_{i2} &> 0] = x_{i1}^{'}\beta_1 + E[u_{i1}|selection\ rule] = \\ &= x_{i1}^{'}\beta_1 + E[u_{i1}|y_{i2} > 0] = x_{i1}^{'}\beta_1 + E[u_{i1}|u_{i2} > -x_{i2}^{'}\beta_2]. \end{aligned} \tag{12.5}$$

We have an omitted variable problem: x_{i2} enters the y_{i1} equation. Of course $E[u_{i1}|u_{i2} > -x_{i2}^{'}\beta_2] = 0$ if u_{i1} and u_{i2} are independent (again, think of Remark 77 and σ_{HF} in the Roy's model).

If we assume that u_{i1} and u_{i2} are jointly normal with covariance σ_{12} and variances σ_1^2 and σ_2^2 respectively, we know what $E[u_{i1}|u_{i2} > -x_{i2}^{'}\beta_2]$ looks like: It is the usual inverse of the Mills' ratio, which we will call here Heckman's lambda:

$$E[y_{i1}|x_i, y_{i2} > 0] = x_{i1}^{'}\beta_1 + \frac{\sigma_{12}}{\sigma_2}\frac{\phi(x_{i2}^{'}\beta_2/\sigma_2)}{\Phi(x_{i2}^{'}\beta_2/\sigma_2)} = x_{i1}^{'}\beta_1 + \sigma_\lambda\lambda(x_{i2}^{'}\beta_2). \tag{12.6}$$

While we can numerically identify σ_λ from β_1 even when $x_{i2} = x_{i1}$ because λ is a non-linear function, there is need for exclusion restrictions (variables in x_{i2} not included in x_{i1}) in order to avoid identification by functional form (i.e. by distributional assumption implying nonlinearity in xs).

The model can be estimated by FIML or in two stages. The two-stage estimation starts with a probit on $y_{i2} > 0$ which delivers $\widehat{\beta_2}$ which can be used to calculate $\widehat{\lambda_i} = \lambda(x_{i2}^{'}\widehat{\beta_2})$. In the second stage y_{i1} is run on x_{i1} and $\widehat{\lambda_i}$ to estimate $\widehat{\beta_1}$ and $\widehat{\sigma_\lambda}$. Of course, if $\widehat{\sigma_\lambda} = 0$, selection is not important.

The joint normality implies a particular form of heteroscedasticity at the second step regression (GLS matrix Γ). Further, we have to make another GLS correction for the fact that we're not using $\lambda_i(z)$ but only $\widehat{\lambda_i}(z)$ so that the error term contains the following: $\sigma_\lambda(\lambda_i - \widehat{\lambda_i}) \cong \frac{\partial\lambda_i}{\partial z}(\beta_2 - \widehat{\beta_2})x_{i2}$ evaluated at $z = x_{i2}^{'}\beta_2$

(this approach of using the first order Taylor series approximation is often called the Delta method–the point is general: whenever you use predicted regressors in a non-linear function, you need to correct your standard errors!). Hence, the variance-covariance of the error term in the second-step regression is composed of Γ plus $\left(\frac{\partial\lambda}{\partial z}\right)' Var(\widehat{\beta_2})\left(\frac{\partial\lambda}{\partial z}\right)$.

Recent semiparametric literature is relaxing the assumption of joint normality of disturbances (see section 12.2.4 below).

Example 12.6. *First run probit on labor force participation and obtain $\widehat{\lambda}$, then run the wage regression to get the effect of education on wages $\widehat{\beta}$ (and $\widehat{\sigma}$).*

Example 12.7. *Consider the hours labor-supply regression with wages on the RHS. First, you need to correct the hours equation for sample selection into labor force (only observe h for those who work). This correction comes from a comparison of behavior equations governing reservation wages w_i^R and market wages w_i which leads to a 0/1 participation estimation depending on $Z_i'\gamma$, where Z is the collection of RHS variables from both w_i^R and w_i equations. Second, you need to instrument for w_i which is likely endogenous. The first stage regression where you predict $\widehat{w}_i$ also needs to have a selection correction in it. Finally, you can estimate*

$$h_i = \delta\widehat{w}_i + x_i'\beta + \sigma\lambda(Z_i'\widehat{\gamma}) + \varepsilon_i.$$

There is serious need for exclusion restrictions: you need an exclusion restriction for running IV for w_i (that is a variable predicting wages but not hours) and you need another exclusion restriction to identify the selection correction in the first-stage wage equation (that is you need a variable affecting participation, but not wages).

Remark 85. *Asymptotic distribution: two stage methods are efficient in one iteration.*

Remark 86. *If the unobservable selection threshold is time constant we can use a fixed effect panel data model to deal with it.*

Example 12.8. *The λ method is applicable in unbalanced panel data, see Lafontaine and Shaw (1995) for an example. Franchisees that go out of business have shorter T_i. Their fixed effect model appears to eliminate most of the selection bias suggesting that within-firm variation in selection has little effect. In other words,* *the FEM approach will work when the reason why firms exit is the time constant unobservable. To apply the λ method, one needs (time changing) predictors of survival excluded from the y equation. Finally, in cases when it is reasonable that those who disappear from the data have y that is always below the median of the (surviving) sample, one can insert such a small value of y for the firms that went belly up and use them in an estimation of a median (LAD) regression, where their particular y value is of no importance, as long as it is below the median.*

Remark 87. *Estimating parametric Limdep models with a λ on the RHS is a big problem, especially with heteroscedasticity, which kills consistency. The λ problem is that selection affects the whole distribution and λ only fixes the expectation (centering).*

Remark 88. *Buchinsky (1998, JAppliedEcm) provides a (rarely used) method of estimating quantile (median) regressions with a sample selection correction. See Section 16.5 for an introduction to quantile regressions.*

12.2.3. Switching Regression

In this case we observe y (and x) under all *chosen* alternatives.

Example 12.9. *The union-nonunion or migrants-stayers wage model. The owner-rental housing demand. The privatized-state profit function.*

A first approximation in case of two choices is a restrictive **constant effect model** which pools data into one regression:

$$y_i = x_i'\beta + \alpha D_i + \varepsilon_i, \tag{12.7}$$

which is estimated by IV under the assumption that $y_{1i} - y_{0i} = \alpha$, where 0 and 1 denote the two different states (union/nonunion, treatment/control). The first stage is based on $P[D_i = 1 \mid z_i] = P[z_i'\gamma + \nu_i \geq 0 \mid z_i]$ so that $\widehat{D_i} = F_\nu(z_i'\widehat{\gamma})$ for a symmetric $F_\nu(\cdot)$.

Remark 89. *The consistency of the main equation is contingent on correct specification of the error distribution $F_\nu(\cdot)$. See Remarks 33 and 54.*

Remark 90. *Again, the standard errors of β need to be corrected for the estimation of γ, see Lee (1981), and the discussion of Delta method above for β_2. With multiple choices, estimate a MNL model in the first stage.*[107]

A more general and widely used approach called **switching regression** assumes there are two (or more) regression functions and a discrete choice model determining which one applies. The typical estimation is similar to Heckman's λ.

Example 12.10. *Munich, Svejnar and Terrell study wages of Czech workers during 1991-1996. The workers either stay in post-communist firms or enter newly started private enterprises (de novo jobs, DNJ). Munich et al. obtain a positive coefficient on λ for the movers and a negative coefficient for the stayers.*

Example 12.11. *Engberg and Kim (1995) study the intra-metropolitan earnings variation: is it caused by person or place effects? Person i chooses his/her location j (inner city/poor suburb/rich suburb) based on his/her (latent) wage in each location w^*(think Roy model) and the location's amenities:*

$$U_{ij}^* = w_{ij}^* \gamma_j + x_i^{'} \alpha_j + \nu_{ij}, \text{ where } w_{ij}^* = x_i^{'} \beta_j + \varepsilon_{ij}.$$

U^* is the latent utility of each location. Assuming that $\varepsilon_{ij}\gamma_j + \nu_{ij}$ is iid logit, they look up the appropriate λ sample-selection formula and proceed to run switching wage regressions. The place effect is measured as $\overline{x}'(\beta_{Suburb} - \beta_{City})$. Actually, they present 6 different results based on what kind of control method they choose, starting with unconditional means. Assuming MNL iid error terms for the choice equation is not appropriate given their maintained assumption of no location effects for the most highly educated workers. They have no credible exclusion restrictions. Identifying the model off functional form blows it up. So they use non-linearities for identification: these come from non-parametric estimation of the selection equation. Do you find this credible? Finally, they run a semi-parametric selection (of location) model:

12.2.4. Semiparametric Sample Selection

See Kyriazidou (1997) and Powell (1989). Assume $d_i = 1\{x_i\gamma + v_{1i} > 0\}$, $y_i = y_{i2} * d_i$, $y_{i2} = x_i\beta + v_{2i}$ and assume that $f(v_1, v_2)$ is independent of x_i. Then if I do

[107] See a recent paper by Bourguignon, Fournier, and Gurgand, which corrects the formulas from Lee (1983). This is implemented in the `selmlog` command in Stata.

not want to assume a particular form for the selection term (i.e., I am not willing to assume a particular distribution f), follow Powell and choose person i and j such that $x_i\gamma = x_j\gamma$ so that $y_{i2} - y_{j2} = (x_i - x_j)\beta + \lambda(x_i\gamma) - \lambda(x_j\gamma) = (x_i - x_j)\beta$. In practice do a Kernel on those pairs which are close, i.e., use an estimator such as

$$\left[\sum K\left(\frac{(x_i - x_j)\widehat{\gamma}}{n}\right)(x_i - x_j)(x_i - x_j)'\right]^{-1}\left[\sum K\left(\frac{(x_i - x_j)\widehat{\gamma}}{n}\right)(x_i - x_j)(y_i - y_j)\right] \tag{12.8}$$

Example 12.12. *Return to Engberg and Kim and note the use of their maintained assumption as both a measuring stick for their different methods and as an identifying assumption for the semi-parametric sample selection estimation, where the constant is differenced out: Using* $\overline{x}'(\beta_{Suburb} - \beta_{City}) = 0$ *for highly educated white males identifies the constant difference for other group.*

Notice that what we are doing is essentially matching pairs of observations with the same probability of selection/participation, i.e. we are *matching on propensity score*. Ahn and Powell further suggest that there is no need for any γ here and all can be done non-parametrically. See Rosenbaum and Rubin for early theoretical work on matching using the propensity score. We return to this issue in Section 13.2.

Remark 91. *Of course, all of these models assume away heteroscedasticity, which is most likely to exist in large micro-data. Songian Chen uses symmetry assumption on* $f(v_1, v_2)$ *to deal with heterogeneity of a particular form at both stages:* $f(|x) = f(v_1, v_2|\widetilde{x})$ *where* $\widetilde{x}$ *is a subset of* x.

The lesson from the example is that one should not attempt a problem of this type without an instrument (exclusion restriction). If there is an instrument, then a situation where we observe the outcome under both hypotheses allows for either selection-model-style switching-regression estimation or for simpler IV solutions. In the next section we will discuss the advantages and disadvantages of these two approaches. But note that the IV strategy is not available when we observe the outcome only for a subset of the individuals (one of the chosen alternatives).

13. Program Evaluation

Consider estimating the impact of a treatment (binary variable): a medical procedure, a training program, etc. Evaluation of social programs is what much of true micro-econometrics is all about. We ask how to estimate the effect of a social program (participation in a training program, change in college tuition) in absence of controlled experiments ([M]9.2.). How can we create counterfactual outcomes (such as what would have happened in absence of treatment)?[108]

Methodological advances in program evaluation are important for all of cross-sectional econometrics.[109] First, the Heckman's bivariate normal selection models dominated the field, but they were somewhat replaced by difference-in-differences models, which assumed that selection was based on time-constant unobservables. Recently, as data become richer, the matching method, introduced below, became popular; it focuses on controlling for observables, in contrast to previous approaches, which worried about selection on unobservables.

One of the main lessons of the recent literature, which we introduce in this section, is that the impacts of programs differ across individuals. Up to the discussion of the Roy model in Section 12.2.1, we estimated models where a given program (e.g., privatization) affected all participants in the same way. In fact, treatment effects often vary across individuals and these differences can be known to program participants or administrators such that they can act upon them, which makes those with the largest gains from the program treatment most likely to participate in the program.

At a fundamental level, we need to differentiate two types of problems: (1) *treatment effect problem*: What is the effect of a program in place on participants and nonparticipants compared to no program at all; (2) *structural problem*: What is the likely effect of a new program or an old program applied to a new setting. The latter problem is perhaps too ambitious and definitely requires more heroic assumptions. So focus on (1).

Crucially, we ask about *partial equilibrium* effects here; no answers given on across-board policy evaluation (such as making every student go to college) – no general equilibrium effects are taken into account!

[108] Return to the first class of the course for a broad introduction to *causal inference.*

[109] See Imbens and Wooldridge (in press) "Recent Developments in the Econometrics of Program Evaluation".

13.1. Setup of the Problem

Consider the effect of a (training) program within the Rubin Causal Model (Holland, 1986), which formulates causal questions as comparisons of potential outcomes: $y_{1i} = \mu_1 + u_{1i}$ are earnings with training, and $y_{0i} = \mu_0 + u_{0i}$ are earnings without training (go back to Roy's model in Section 12.2.1).

Consider the population of eligible workers. They first choose to apply for the training program or not. We observe y_{1i} only when $D_i = 1$ (the person applied for and took training) and observe y_{0i} only when $D_i = 0$ (these are the so called eligible non-participants, ENPs).

13.1.1. Parameters of Policy Interest

We may want to know $E[y_{1i} - y_{0i}] = \mu_1 - \mu_0$, i.e. the treatment effect under random assignment, the *average treatment effect* (ATE), a.k.a. "the true causal effect."[110] We may also want to know $E[y_{1i} - y_{0i}|D_i = 1]$, the *average effect of treatment on the treated* (ATT), which allows us to ask whether the current program, given the voluntary participation and all, is worth its costs or not. Finally, we may want to know the average effect on the untreated (ATU): $E[y_{1i} - y_{0i}|D_i = 0]$, which allows us to see whether we should expand the current program.[111]

Note that ATT = ATE + $E[u_{1i} - u_{0i}|D_i = 1]$, where the second term is the gain that the treatments obtain from treatment on top of the gain obtained by a random person. Of course, if the effect of treatment is the same for all individuals then ATE=ATT=ATU.

The fundamental problem: Consider estimation of ATT: The data only provides $E[y_{1i}|D_i = 1]$ and $E[y_{0i}|D_i = 0]$ but $E[y_{0i}|D_i = 1]$ is the "what-if" counterfactual.

$$\begin{aligned} E[y_{1i}|D_i = 1] - E[y_{0i}|D_i = 0] &= \text{ATT} + \{E[y_{0i}|D_i = 1] - E[y_{0i}|D_i = 0]\} \\ &= \mu_1 - \mu_0 + E[u_{1i} - u_{0i}|D_i = 1] + \\ &\quad \{E[u_{0i}|D_i = 1] - E[u_{0i}|D_i = 0]\}. \end{aligned}$$

[110] This is what biostatistics is after. Economists care about ATE when evaluating a mandatory program that affects for example all unemployed.

[111] Note that we may want to know the answer of the effect of expanding the program by 10% or by 30%. These two effects may differ is those who benefit more from the program are more likely to participate, all else equal.

The sample selection bias (the term in curly brackets) comes from the fact that the treatments and controls may have a different outcome if neither got treated.

13.1.2. Experimental Solution

An almost ideal (see end of Section) tool for measuring causal effects (think of medical trials using placebo). Randomization ensures that on average the sample selection bias is zero: $E[y_{0i}|D_i = 1] = E[y_{0i}|D_i = 0]$.

Basic structure: Take the $D = 1$ group and randomize into treatment ($R = 1$) and control ($R = 0$) group. Then construct the experimental outcome: $E[y_{1i}|D_i = 1, R_i = 1] - E[y_{0i}|D_i = 1, R_i = 0]$. This can be used as a benchmark for the accuracy of sample selection techniques that we need when we have no experiment.[112]

Results obtained from experiments are less controversial for general public than those based on non-experimental methodology and it may be harder to cheat when running an experiment compared to conducting a non-experimental study.

However, experiments are (sometimes) costly and often socially unacceptable (in Europe). Further, randomization may disrupt the operation of the program in question and people may behave differently knowing they are in an experiment (think of temporarily expanding medical coverage). Some important treatment questions cannot be subject to experimental research design (family income while young, etc.)

Even with experimental data, there are often problems of selectivity. Double blinding is not feasible in social sciences. First, worry about (obviously very selective) non-compliers: those with $R_i = 1$ who do not show up for training and those with $R_i = 0$ who get a similar treatment outside of the program. (Those who volunteer for the program but are randomized out of treatment may get treated elsewhere or may refuse to provide data.)[113] Second, dynamic sample selection may arise even in presence of initial randomization: see Ham and LaLonde (1996) for a duration study.

Remark 92. *The ATT is an average. It can be positive when most of the treatment group gets hurt by treatment, but a few participants receive very large ben-*

[112] LaLonde's (1986) study, which showed that non-experimental methods are unable to replicate experimental results, motivated both further development of non-experimental methods and more use of social experiments.

[113] In those cases, one can exploit the random assignment and estimate the "intention-to-treat" (ITT) parameter. However, if we use the original treatment assignment as an IV for actual treatment, we're back in the LATE IV word (see below), where one needs to model the behavior (non-compliance), i.e., the heterogeneity in the response to assignment.

efits. More generally, randomized controlled trials only inform about the mean of the treatment effect and not about the distribution.[114]

Remark 93. *Inference in experiments is not trivial. If we run $y_i = \beta_0 + \beta_1 T_i + \epsilon_i$, using the randomized treatment dummy T, we can use the standard error only if we make a heteroscedasticity correction because treatment affects the variance of y. The difference in means divided by the standard error (approximately) has the Student's t distribution only when the number of treatments and controls is the same. See Duflo, Glennerster and Kremer (2008) for a discussion.*

Remark 94. *There are other inference issues. Asking ex post, after conducting an experiment, if one group benefits more than another smells of data mining (there is no way to ex post adjust inference). Looking at multiple outcomes and finding the one that "works" is another example (although there is a way to adjust inference—see the Bonferroni's correction for multiple comparisons). Successful replication of "what works" experiments is more likely if we understand the mechanism of the treatment effect. Controlling for other Xs may introduce finite sample bias (because of the correlation with treatment that will be there in small samples even if the population correlation is zero) but may improve precision if these covariates predict y for both treatments and controls.*

If we do not have an experiment, how can we deal with the fundamental problem? Two answers are proposed, one when we have the key control variables (and do not have an instrument for treatment, 13.2), the other when we do have an exclusion restriction (which is needed because we do not have the key control variables, 13.3).

13.2. Matching

This is a non-experimental solution based on controlling for enough observables such that selection on unobservables is (assumed) irrelevant.[115] Consider estima-

[114] Note that both randomization and linearity are required for the 'magic' where we learn something about hypothetical outcomes without imposing any structure. In particular, randomization ensures that $E[y_{0i}|R_i = 1] - E[y_{0i}|R_i = 0] = 0$ such that $E[y_{1i}|R_i = 1] - E[y_{0i}|R_i = 0] = E[y_{1i}|R_i = 1] - E[y_{0i}|R_i = 1] + 0$ and thanks to expectation being a *linear* operator, this equals $E[y_{1i} - y_{0i}|R_i = 1]$,the ATT. We can't find out about the median impact as the median of the difference is not the difference in medians.

[115] The assumption is that a sufficiently detailed (non-parametric) conditioning on observables makes the sample selection problem go away.

tion of the ATT. When can we use $E[y_0|D=0]$ as a surrogate for the counterfactual $E[y_0|D=1]$? If $E[y_0|D=1,X]=E[y_0|D=0,X]$ then there is hope. This condition (assumption) says that selection occurs based solely on observables, i.e. that "conditioning" on X (within each X cell) assignment to treatment is random. This is similar to standard exogeneity assumptions in regressions.

Example 13.1. *An example of selection on observables is the following process of admission to university: Each applicant is ranked based on the number of "points" derived from observable characteristics. Admitted students are a random sample from those who pass the admission threshold (Angrist and Krueger, 1999).*

The assumption of unconfoundedness given X is that $(y_{0i}, y_{1i}) \perp D_i|X_i$.[116] To estimate only ATT (as opposed to ATE), one needs only a weaker version of the unconfoundedness assumption, namely that $y_{i0} \perp D_i|X_i$.[117] This assumption allows us to get at the counterfactual because

$$E[y_0|D=1] = E\{E[y_0|D=1,X]|D=1\} = E\{E[y_0|D=0,X]|D=1\}$$

so we estimate the last term by

$$\frac{1}{N_1}\sum_{i\in\{D=1\}} \widehat{E}[y_0|D=0, X=x_i].$$

We can estimate a regression using $D=0$ data but predict outcome using $D=1$. Alternatively, proceed non-parametrically and use matching to estimate the ATT: (i) for each value of X (combinations of x_k values) compute the difference in y between treatments and controls, (ii) average these differences across treatments' X. This is similar to a non-parametric (kernel) regression (see Section 16.1). The idea is to resemble an experiment—to make the distribution of observed determinants balanced across the two groups that we compare. Matching is a process of 're-building' an experimental data set.

But what if the X support of $E[y_1|D=1,X]$ and $E[y_0|D=0,X]$ does not coincide? We cannot predict out of sample in terms of X for $D=0$ non-parametrically. Lack of *common support* will occur frequently if the dimension of X is high and is natural when comparing treatments and controls who are not randomly assigned.

[116] We assume that the treatment status is conditionally mean independent from the potential outcomes. This is also called the Conditional Independence Assumption. See Cochrane and Rubin (1973) or Rosenbaum and Rubin (1983).

[117] Note that the identifying assumptions motivating the panel-data fixed effect model of Section 5.1 are $E[y_{0it}|\alpha_i, D_{it}, X_{it}] = E[y_{0it}|\alpha_i, X_{it}]$ together with a linear model for $E[y_{0it}|\alpha_i, X_{it}]$.

Remark 95. *For the matching estimation of ATT, we need the presence of an 'analogue' control for each treated; it does not matter if there are controls with no 'matched' treated individuals. We drop those treated who have no 'matched' controls. So if impacts vary across people, this means that estimates of ATT by methods that do not drop any observations (i.e., regressions) will have different population analogues.*

Matching on Propensity Score (Index Sufficiency) However, 'exact' matching on the combination of all X values is not practical with high-dimensional X because there will be little data in each 'data cell' corresponding to a particular X combination.[118] We deal with the curse of dimensionality (of matching based on a high-dimensional X) by conditioning only on a scalar: the propensity score $p(X) = E[D|X] = \Pr[D = 1|X]$. We match on $p(X)$ over the common support – compare the outcome for those individuals (ENPs compared to treatments) with similar probability of participation in the program.[119]

Rosenbaum and Rubin (1983) show that this works, i.e., that if $y_1, y_0 \perp D|X$ then $y_1, y_0 \perp D|p(X)$ using the fact that $D \perp X|p(X)$ (that is $\Pr[D = 1|X, p(X)] = \Pr[D = 1|p(X)] = p(X)$).

In practice, matching on $P(X)$ is not exact so we stratify the propensity score or do a Kernel (see Section 16.1) or Nearest Neighbor (see Section 16.2) non-parametric regression.[120] The choice of the matching method can make a difference in small samples (see Heckman, Ichimura and Todd, 1997). Standard errors are typically bootstrapped, although this has been shown to be invalid.[121] One of the

---

[118] Also see Abadie and Imbens (2002) on the bias of simple matching estimators when X dimension is high. Angrist and Hahn (2004) show that even though there is no asymptotic gain in efficiency from using the propensity score as opposed to matching on covariates, there will likely be a gain in efficiency in finite samples. Using the propensity score essentially corresponds to applying prior knowledge to reduce dimensionality and this will improve precision in small samples.

[119] So always start by plotting the pscore for the treatment and control group in the same graph. See Caliendo and Kopeinig (2008) for a survey of practical issues with implementing matching strategies.

[120] Stata programs to estimate treatment effects are available from Becker and Ichino (`att*`, 2002), Leuven and Sianesi (`psmatch2`, 2003) and Abadie et al. (`nnmatch`, 2004). See also notes by Andrea Ichino at `http://www.iue.it/Personal/Ichino/` Guido Imbens provides examples of matching programs for his re-evaluation of the LaLonde AER data at `http://emlab.berkeley.edu/users/imbens/estimators.shtml`

[121] Abadie and Imbens (2009, NBER WP No. 15301) provide the large sample distribution of propensity score matching estimators and a variance adjustment that corrects for the fact that

most widely used methods today is dif-in-difs for matched data.

Matching on $P(X)$ shifts attention from estimating $E[Y|X, D]$ to estimating $P(X) = E[D|X]$, which may be attractive when we have a better motivation or a model for the selection into treatment than for how the treatment operates. We often make parametric assumptions when estimating the pscore (otherwise, it is not clear that the curse of dimensionality would be reduced).

How much matching reduces the bias is an empirical question.[122] It depends on whether matching on observed Xs balances unobserved determinants or not. See Heckman, Ichimura, Smith, and Todd (1998), Angrist (1995) or Dehejia and Wahba (1998). One of the lessons from this literature on the effects of training programs on the labor market is that one should match on sufficiently lagged pre-treatment performance (Ashenfelter's dip). To the extent that past performance and current Xs do not capture individual's motivation or skills, the bias remains.

Remark 96. *One cannot directly test (reject) the unconfoundedness assumption, but in cases that there are two distinct control groups (as in Heckman, Ichimura and Todd, 1997), ideally two groups with different potential sources of bias, one can test whether the matched 'causal' effect estimate of being in one group as opposed to the other is zero as it should be. Another related test is to estimate the 'causal effect' of treatment on lagged outcome (for example wages before training). If this 'effect' is estimated to be close to zero, the unconfoundedness assumption is more plausible. Finally, one can provide (Rosenbaum, 2002) bounds on the degree of the departure from the unconfoundedness assumption that would make the estimated ATT insignificant.*[123] *For another strand of analysis of sensitivity of matching to the Conditional Independence Assumption, see Ichino et. al. (2006) or Nannicini (2006).*

Remark 97. *The balancing property holds given the true propensity score (Rosenbaum and Rubin, 1983), but no such result exists for the estimated pscore. Hence, to check that matching does its job, we need to use t tests to make sure that there*

---

the p-score itself is first estimated.

[122] Matching is the *estimator du jour* in the program evaluation literature (Smith and Todd, 2004). The meta-analysis of active labor market policy program evaluations by Card, Kluve and Weber (2009, IZA DP no. 4002) suggests that "experimental and non-experimental studies have similar fractions of significant negative and significant positive impact estimates, suggesting that the research designs used in recent non-experimental evaluations are unbiased."

[123] In Stata, `rbounds` gives the bounds for continuous variables and `mhbounds` works for binary outcome variables.

are no significant differences in covariates means for both groups after matching, i.e., to test for balancing given pscore. See Rosenbaum and Rubin (1985).

Remark 98. *One can implement the p-score matching by weighting observations using $P(X)$ to create a balance between treatment and control groups.*

Remark 99. *When the p-score cannot be consistently estimated because of choice-based sampling designs with unknown sampling weights, Heckman and Todd (2009, IZA DP No. 4304) show that the selection and matching procedures can be implemented using propensity scores fit on choice-based samples with misspecified weights, because the odds ratio of the propensity score fit on the choice-based sample is monotonically related to the odds ratio of the true propensity scores.*

Remark 100. *See IZA DP. No. 4841 for p-score-based estimation of distributional effects of interventions (i.e., not just of average effects).*

Matching with Multiple Treatments What if the treatment is not binary? Multiple-treatment evaluation has been developed by Imbens (2000) and Lechner (2001). Let $Y_i^0, Y_i^1, ..., Y_i^K$ correspond to outcomes with K different intensities of treatment or no treatment ($k = 0$). Let $T_i = k$ denote the actual occurrence. Now, define ATT of treatment k using all the $K + 1$ pairwise comparisons

$$E[Y^k - Y^{k'}|T = k] = E[Y^k|T = k] - E[Y^{k'}|T = k] \text{ for } k \in \{0, 1, ...K\}, k \neq k'$$

and estimate the second term (counterfactual) by $E_X[E(Y^{k'}|T = k', X)|T = k]$. Note that there is a lot of "common support" conditioning here (find a match for all k') so one really needs propensity score matching, which is estimated separately for all combinations of k and k' (ignoring the other groups).[124] The conditional independence assumption must hold for each of the k - k' comparison.

Matching and DiDs A popular approach suggested by Heckman, Ichimura, and Todd (1997) that makes the CIA assumption more likely to hold is to combine matching with diff-in-diffs, i.e., to compare the *change* in the outcome variable for the $k-$treated groups with the *change* in the outcome variable for all non-k groups, thus purging time invariant unobservables from the specification based on the familiar common trend assumption. Of course, this is done for binary treatments too. Conditioning is still based on pre-treatment Xs as post-treatment data would violate the CIA.

[124] Sianesi (2001) provides the Stata 8 code.

Regression or Matching? What if, instead of matching, we run a regression of y on D controlling for X? What is the difference? First, the regression approach imposes functional form (linearity) over the common support area while matching is non-parametric. Regressions also use their functional form to work off the common support, i.e., use areas of X with non-overlapping support of treated and controls, which can be highly misleading.

But what if we include a full set of interactions among xs to approximate a non-parametric solution (i.e., use a fully saturated model)? This regression still differs from the matching ATT estimator in the implicit weighting scheme: Matching gives more weight to X cells with high probability of treatment $P(X)$ (cells with high share of treated) and 0 weight to cells where $p(X) = 0$. Regression gives more weight to X cells where proportion of treated and untreated is similar, i.e., where the conditional variance of treatment is largest. (See Remark 8 on implicit weighting in regressions and Angrist, 1998). If those who are most likely to be selected (into training, the military, etc.) benefit less from the treatment (because they have the best earnings potential), then matching will give smaller effects than regression.

In the end, what matters most is not whether we run a regression or a matching exercise, but that we inspect the data on being balanced. Pscore estimation is useful for this. Recently, Crump, Hotz, Imbens, and Mitnik (2009) suggest that one select data on common support using propensity score and then run regressions on such balanced data. Further, inference in matching may be less standardized than in regressions.

13.3. Local IV

Now start worrying about unobservables again. Suppose, that we have instruments (affecting choice, but excluded from the y equation). There is an important contrast between IV methods and sample selection methods:

What are the *policy parameters of interest*? What do we want to know? ATE, ATT, ATU, or a treatment effect related to a specific new policy – affecting a specific subset of the population? If the effect of a given policy differs across parts of population (parameter heterogeneity) we need to be able to estimate the effect on each part. There are fundamental reasons why the treatment effect may differ across individuals.

13.3.1. Local Average Treatment Effect

Imbens and Angrist (1994) prove the LATE theorem, which provides interpretation for IV estimates when treatment effects are heterogenous in the population.[125] If the effect of x on y varies in the population, then it can be shown that IV estimates are weighted averages of these group-specific effects where higher weight is given to those groups whose x is better explained (predicted) by the instrument (see Remark 8 and 34). So the IV estimate is the treatment effect on specific groups–it is a "local" effect. Specifically, a local average effect of the treatment on those who change state (treatment status) in response to a change in the instrument (the *compliers*). It is not informative about the effect for the never-takers or always-takers, i.e., those for whom the IV value does not help to predict treatment status, or for the defiers—those who were assigned to treatment (should get treated given their IV value and the way the first stage generally works), but did not get treated.

Example 13.2. *Angrist and Krueger (1991) use quarter of birth and compulsory schooling laws requiring children to enrol at age 6 and remain in school until their 16th birthday to estimate returns to education. This approach uses only a small part of the overall variation in schooling; in particular, the variation comes from those who are unlikely to have higher education.*

Example 13.3. *Similarly, one may think of the Angrist (1990) estimate of the effect of military service as corresponding to the effect of the service on those drafted using the Vietnam-era lottery, but not those (majority) soldiers who volunteered.*

Remark 101. *Given that LATE identifies the effect on compliers, it is important to know as much as possible about the compliers, who, however, cannot be individually identified. See p.171 of Angrist and Pischke (2009) for a way to describe the distribution of compliers' characteristics.*

Remark 102. *Note, that this is a general problem of all estimation. The only difference is that IV selects a specific part of variation (we know who identifies the effect) whereas OLS can be thought of as weighted average of many sources of variation, some potentially endogenous.*

[125] The assumptions underlying the theorem are IV exogeneity and exclusion restriction, as usual, and also "monotonicity", which requires that anyone who would get treated if not induced so by the IV would also get treated if the IV is "pushing" for treatment.

For an introduction to the topic, see http://www.irs.princeton.edu/pubs/pdfs/415.pdf

There is a strong analogy between IV, which we assume to be exogenous, but which does not give an R^2 of 1 in the first stage, and a randomized experiment with non-perfect compliance with assigned status is important. In situations with no defiers, ATT is a weighted average of the effect of treatment on the compliers and on the always-takers, but LATE IV only identifies the effect on compliers. LATE will give ATT only when there are (almost) only compliers. IV will only give ATE under special conditions; for example, when something that was a matter of choice becomes legally binding such that everybody must comply (by law) and there are no never- or always-takers.[126]

Remark 103. *Estimating the so-called the reduced form (see Remark 38) of the outcome variable on the randomly assigned offer of treatment (the IV) gives the so-called* intention to treat *(ITT) parameter. LATE then equals ITT divided by the* compliance rate *(the first stage associated with the IV) based on the indirect least squares argument (see Example 7.1).*

Remark 104. *Note that because different instruments estimate different parameters, overidentification tests discussed in Section 7.1 are out the window.*

LATE is a treatment effect at the *margin of participation* (in treatment) *relative to the instrument*. Suppose that your instrument must have an economically small effect, such as the presence of a small fee for training materials affecting the choice to participate in a training program. Then the LATE corresponds to people who are just about indifferent between participation or not.

In *regression discontinuity* designs (see Section 7.1), one estimates a treatment parameter, but also only a local one—for those who are at the regression discontinuity and only for compliers.

Example 13.4. *Costa Dias, Ichimura and van den Berg (2008, IZA DP no. 3280) propose a simple matching-and-IV method based on sharp discontinuity in treatment probability. They match on X and then apply a correction based on an IV Z that shifts $P(X)$ to zero for values of Z above a certain threshold. In their case, the discontinuity arises thanks to eligibility rules for treatment: offer no training to unemployed workers above some age level. Just re-do matching (with kernel weights) but apply additional weights based on the distance from age eligibility cutoff.*

---

[126] See Manski (1990, 1996, 2003) for an approach that leads to bounds for ATE given assumptions one is willing to make.

The correction is based on the following argument: Assume that $y_0 \perp Z|X$ *and* $P[D = 0|X, Z = z^*] = 1$*. Then* $E[y_0|X] = E[y_0|X, Z]$ *and*

$$E[y_0|X, Z] = E[y_0|X, Z, D = 0]P[D = 0|X, Z] + E[y_0|X, Z, D = 1]P[D = 1|X, Z]$$

and at $Z = z^*$ *this implies that* $E[y_0|X] = E[y_0|X, Z = z^*, D = 0]$*. Next,*

$$\begin{aligned} E[y_0|X, D = 1] &= \frac{E[y_0|X] - E[y_0|X, D = 0]P[D = 0|X]}{P[D = 1|X]} \\ &= \frac{E[y_0|X, Z = z^*, D = 0] - E[y_0|X, D = 0]P[D = 0|X]}{P[D = 1|X]} \\ &= E[y_0|X, D = 0] + \frac{E[y_0|X, Z = z^*, D = 0] - E[y_0|X, D = 0]}{P[D = 1|X]}, \end{aligned}$$

where the second term will equal zero if the conditional independence assumption that underpins standard matching holds (this can be tested).

Example 13.5. *For an application of a wide array of matching and non-parametric IV estimators, see Ham, Li and Reagan (2009) migration study, which applies, among others, the Frölich (2007) non-parametric LATE estimator with covariates.*

13.3.2. Marginal Treatment Effect

Next, consider the recent work by Heckman and coauthors, who use the concept of the marginal treatment effect (MTE) introduced by Bjorklund and Moffitt (1987) to provide a unified perspective on the IV and sample selection literature.[127] To start from scratch: there are two possible states of the world:

(a) There is one "true" effect of D on y, namely β. There may be correlation between D and unobservables so we run IV for D in estimating:[128]

$$Y_i = \alpha + X_i\theta + \beta D_i + U_i.$$

(b) Return to Roy and Section 13.1. Individuals know (estimate) their *varying* returns from choosing D and *act upon the size of the return*:

$$\begin{aligned} Y_{i0} &= \alpha_0 + X_i\theta_0 + U_{i0} \\ Y_{i1} &= \alpha_1 + X_i\theta_1 + U_{i1}, \end{aligned}$$

[127] See http://jenni.uchicago.edu/underiv/userguide_march_22_2006.pdf

[128] Alternatively, write the common effect model as $Y_0 = \alpha + U$, $Y_1 = \alpha + \beta + U$.

so that the causal effect

$$\beta_i = Y_{i1} - Y_{i0} = \alpha_1 - \alpha_0 + X_i(\theta_1 - \theta_0) + U_{i1} - U_{i0}. \tag{13.1}$$

There is, therefore, a *distribution* of returns (correlated random coefficients, ex post causal effects) that cannot be summarized by one number β.[129] Individuals are likely to act on their knowledge of the size of their own gain (effect).[130]

To complete the model (similar to Heckman's λ switching regressions), introduce an "instrument" Z and postulate the selection equation:

$$\begin{aligned} D_i^* &= \theta_D Z_i + U_{iD} \\ D_i &= 1 \text{ if } D_i^* \geqq 0. \end{aligned}$$

Next, we define a new parameter: the marginal treatment effect $\text{MTE}(x, u_D) = E[\beta|X = x, U_D = u_D]$ as the effect of D on a person with values (x, u_D) that is just indifferent between choosing $D = 1$ and 0.[131] It is a willingness to pay measure for those at the margin of indifference given X and U_D.

The typical range of policy parameters (ATT, ATE, etc.) can be expressed as differentially weighted averages (integrals over population) of the MTE. IV and OLS can also be expressed as weighted averages of MTE, but the weights are not those of ATT, ATE, etc. IV weights are related to the type of instrument (LATE interpretation). Heckman et al. conclude that while IV estimation may be more statistically robust compared to sample selection methods, IV may often not answer any economically interesting questions.

Example 13.6. *Consider an application of this approach to estimating the wage* ***returns to education*** *(Carneiro, Heckman and Vytlacil, 2002) where $D \equiv S \in \{0, 1\}$ is college or high school. Either (a) human capital is homogenous (Griliches, 1977) and there is one β or (b) human capital is heterogeneous and people know their returns from S when making schooling decisions (Willis and Rosen, 1979).*

[129] In the standard IV literature (a), we worry about (i) unobservable heterogeneity bias ($COV(D, U) \neq 0$), (ii) downward measurement error bias ($COV(D, U) \neq 0$), and (iii) the weak instrument bias, where $COV(U, D)/COV(IV, D)$ is large because $COV(IV, D)$ is small. Here, under (b), there are more econometric problems: $COV(D, U_0) \neq 0$ as before, but also $COV(\beta, U_0) \neq 0$ and crucially $COV(\beta, D) \neq 0$.

[130] For a test of whether subjects act on the knowledge of gains, see NBER Working Paper No. 15463.

[131] Start with writing down the LATE for the case of a binary IV (values z and z') and let z' go to z. This is MTE. The people who are affected by such a small change in IV are indifferent between the two choices.

To deal with ability and measurement error biases, Card uses college proximity as an IV (see example 7.2). Typically $\beta_{IV} > \beta_{OLS}$. Now think of the LATE IV interpretation: β_{IV} is the effect of college on wages for those people whose college participation is affected by whether or not they grow up near college – these are students from low income families. Card therefore interprets $\beta_{IV} > \beta_{OLS}$ as saying that students from low income families have high β, but don't attend college because of credit constraints. The instruments affect those with high MTE values disproportionately. However, these cost constraints must be really strong, because if we think of OLS as the average effect and IV as the marginal effect, one would expect the marginal student (affected by instrument, typically cost-related IV) to have lower returns than the average (typical) student who always goes to college (because benefits always exceed costs). Heckman et al. say Card's interpretation is wrong, because OLS is not the average effect. They say that there is a large positive sorting gain (comparative advantage in Roy model), Willis and Rosen prevail (marginal is below average), and true β > IV > OLS.

So what is MTE? First, run a probit to get $\widehat{D} = E[D|Z] = P(Z)$. Second, using Equation 13.1, the observed outcome is

$$Y_i = \alpha_0 + X_i\theta_0 + D_i[\alpha_1 - \alpha_0 + X_i(\theta_1 - \theta_0)] + \{U_{i0} + D_i(U_{i1} - U_{i0})\}.$$

Now,

$$E[Y|Z,X] = \alpha_0 + X\theta_0 + E(D|Z)[\alpha_1 - \alpha_0 + X(\theta_1 - \theta_0) + E(U_1 - U_0|D = 1, Z)], \tag{13.2}$$

where the term in square brackets is the ATT. Next, invoke index sufficiency and condition on $P(Z)$ instead on Z (this is like instrumenting for D) and find the marginal treatment effect MTE$(x, P(z))$ as $\frac{\partial E[Y|X,P(Z)]}{\partial P(z)}$.

Note that the non-linearity of Equation 13.2 implies heterogeneity in the treatment effects that is correlated with treatment status. This highlights the limitation of linear IV methods: the relationship between $E[Y|X, P(Z)]$ and $P(Z)$ is non-linear, so linear IV is misspecified and IV estimates depend on the instrument used (LATE). *Outcomes are non-linear function of participation probabilities.* The MTE is then simply the slope of this function at a given $P(z)$.

The ATE equals the slope of the line that connects the $E[Y|P(Z)]$ at $P(Z) = 0$ and $P(Z) = 1$. Similarly, the TT(z) connects $E[Y|P(Z)]$ at $P(Z) = 1$ and $P(z)$. However, in practice, we do not have (treatments') data at $P(Z) = 0$. The LATE IV in the case of a binary instrument (Z taking only two values) just connects $E[Y|P(Z)]$ at the two values z and z'.

Heckman (1990) shows that the sample-selection model is not identified without distributional assumptions (non-parametrically) unless we observe $P(Z)$ at 0 and at 1 for some Z (we need the IV to be unbounded for this "identification at infinity", which does not happen in practice). So how does one estimate the MTE? There are several ways. Heckman et al. use a **Local IV** method.[132] Think of a switching regression (with non-parametric λs) applied at specific ranges of U_D (their rich instrument affects behavior in many parts of the ability distribution). See also Angrist (2004).

A simpler approach is proposed by Moffitt (2007, NBER WP no. 13534) who uses simple series estimation (splines in or polynomials of p-score, in his case $\Phi(z'\theta_D)$) to non-parametrically trace out the shape of the relationship between participation probability and the outcome. Specifically, he runs a linear regression with regressors non-linear in $\Phi(Z)$ to approximate $Y_i = \alpha_0 + X\theta_0 + P(Z)g(P(Z)) + \epsilon_i$. The g function here is the TT effect (see equations 13.1 and 13.2 to see this) such that $MTE = g(P(Z)) + P(Z)g^{'}(P(Z))$.

Example 13.7. *Moffitt (2007) can extrapolate out of the range of observed $P(Z)$, but one would have little trust in such results. He uses 3 IVs in order to operate in different portions of the $P(Z)$ distribution. He first shows a histogram of $\widehat{P(Z)}$, which is thin towards $P(Z) = 1$,but he also explains that what matters is not only where the data is, but also where the instruments have incremental effects (where in the range of $P(Z)$ they are not weak). Do see Figure 2 of the paper on this point. In his case, both conditions are satisfied in the region from 0.3 to 0.6.*

Remark 105. *Return to the issue of why $\beta_{IV} > \beta_{OLS}$ in the returns to education estimation with constant returns discussed in example 13.6. For MTE > (local) OLS (as in the Card explanation of $\beta_{IV} > \beta_{OLS}$), it is necessary that MTE > TT (outside of the neighborhood of $P(Z) = 0$ or 1). Moffitt (2007) goes through a proof of this argument and rejects this explanation of why $\beta_{IV} > \beta_{OLS}$ using his UK data.*

Remark 106. *An easy-to-read summary (including a discussion of identification of Equation 13.2) can be found in Manning (2003).*[133]

[132] They use non-parametric estimation by breaking the regression down into steps. See Section 16.3 for the definition of Local Linear Regression. $\frac{\partial PE(U_1 - U_0|P)}{\partial P}$ is approximated using discrete differences. The local IV method has now been extended to unordered multiple choice models: http://ftp.iza.org/dp3565.pdf.

[133] Available at http://econ.lse.ac.uk/staff/amanning/work/econometrics.html

Remark 107. *One can obtain MTE in the standard Heckman's λ approach using formulas in Heckman, Tobias, and Vytlacil (2000) NBER WP No. 7950. Assuming a parametric choice equation ($\theta_D Z$),*[134] *the parameters of interest are:*[135]

$$
\begin{aligned}
ATE(x) &= x(\theta_1-\theta_0) \\
MTE(x,u_D) &= x(\theta_1-\theta_0)+E(U_1-U_0|U_D=u_D) \\
TT(x,z,D(z)=1) &= x(\theta_1-\theta_0)+E(U_1-U_0|U_D \geqq -z\theta_D) \\
LATE(x,D(z)=0,D(z')=1) &= x(\theta_1-\theta_0)+E(U_1-U_0|-z'\theta_D \leqq U_D \leqq -z\theta_D)
\end{aligned}
$$

Assuming joint normality of (U_D, U_1, U_0) and normalizing variance in choice Probit, we have

$$
\begin{aligned}
TT(x,z,D(z)=1) &= x(\theta_1-\theta_0)+(\sigma_1-\sigma_0)\frac{\varphi(z\theta_D)}{\Phi(z\theta_D)} \\
LATE(x,D(z)=0,D(z')=1) &= x(\theta_1-\theta_0)+(\sigma_1-\sigma_0)\frac{\varphi(z'\theta_D)-\varphi(z\theta_D)}{\Phi(z'\theta_D)-\Phi(z\theta_D)} \\
MTE(x,u_D) &= x(\theta_1-\theta_0)+(\sigma_1-\sigma_0)u_{D.}
\end{aligned}
$$

Now take integrals over sub-populations; e.g., for ATT average of the $D=1$ population. We assume the functional form (shape) of the non-linear relationship between $P(Z)$ and $E[Y|Z]$ so even a binary instrument (just observing two points) is enough to know the whole curve and the whole range of policy relevant parameters.

Remark 108. *Heckman (IZA DP no. 3980) extend the MTE approach to unordered choice.*

Remark 109. *So far, we covered a binary treatment variable and so we discussed "local average treatment effects". With a continuous X, the 2SLS estimates "local average partial effects", it has a "average derivative" interpretation.*

In general, the lesson is that when responses are heterogeneous, there is no guarantee that 2SLS (with a valid instrument) is going to identify parameters of economic (policy) interest (any more than OLS). Especially when responses to choices vary among individuals and this variation affects the choices taken.

[134] We now hide $\alpha_1-\alpha_0$ inside the $x(\theta_1-\theta_0)$ to save space.

[135] Go back to example 12.11.: we estimated the ATE.

When an IV is correlated with heterogeneous responses, 2SLS will not reveal the average partial effect. Heterogeneity is not a technical problem, but more often than not, it will correspond to the mechanics of the treatment effect. On the other hand, if the variation in the IV closely corresponds to the policy you have in mind (changing college tuition by the amount that corresponds to cost difference between those that live near a college or not), then LATE 2SLS is of interest and has high internal validity.

Even more generally, LATE 2SLS is what we get in the "natural experiment" literature, where instead of estimating a simultaneous equation *model* (with exclusion restrictions) corresponding to economic theory, we estimate one structural equation whilst using an ad hoc external instrument (see Remarks 30 and 32 and Section 2 of Deaton 2009, NBER WP no. 14690). In this literature, the "questions [we answer]... are *defined* as probability limits of estimators and not by well-formulated economic problems" (Heckman, 2009, IZA DP no. 3980). The choice of the instrument determines the question we answer. We estimate unspecified "effects" instead of questions we care about.[136] Deaton (2009) adds that by using the LATE IV approach, we do not learn why and how the treatment works, only whether it works, which is a major issue for generalizing and applying the findings.

Example 13.8. *Consider estimating the effect of railroads construction on poverty in a linear regression with an IV for railroad construction, such as whether the Government of China designates the given area as belonging to an "infrastructure development area" (Deaton 2009). The heterogeneity in the treatment parameter corresponds to different ways (context) of how railroads may alleviate poverty. The variation in the parameter is about the* mechanisms *that ought to be at the main objects of the enquiry. The deviation in the parameter from its average will be in the residual and will not be orthogonal to the IV if,* within *the group of cities designated as infrastructure zones, those that build railway stations are those that benefit the most in terms of poverty reduction, which one hopes to be the case. The Heckman's local IV approach asks about how cities respond to their designation—asks about the mechanisms.*

Example 13.9. *Similarly, in the Maimonides rule regression-discontinuity example 7.5, the children that are shifted to classrooms of different sizes are not a*

[136] Heckman and Vytlacil (2005) then provide functions of IVs that answer well-posed questions (see above).

random sample of all children in these classes. We need to know how children end up in different classes. See Urquiola and Verhoogen (2008).

Therefore, researchers recently *combine experimental variation with structural models*, as a way of testing the structure or to identify it. Structural models are then used to deliver parameters of policy interest (see Example 15.3 below). *Experiments ought to be theory driven* to deliver information on the mechanisms of treatment effects, to test predictions of theories that are generalizable, rather than to just test "what works". For further reading, see "Giving Credit Where it is Due" by Banerjee and Duflo and the 'theory-based impact evaluation' calls by *3ie*.[137]

[137] At this point, read Blundell and Costa Dias (2008, IZA DP No. 3800). They present six related evaluation approaches in empirical microeconomics that we have covered in this course: (i) randomized (social) experiment, (ii) natural (quasi) experiment (i.e., difference in differences and when the method will identify the ATT), (iii) discontinuity design methods, (iv) matching, and (v) IV and control function methods.

14. Duration Analysis

[W] 20. Here we return to simpler reduced-form distribution-based maximum likelihood modelling, which is designed to fit the processes that result in variation in duration (length).[138]

Example 14.1. *Length of a job, duration of a marriage, how long a business lasts, when a worker retires, duration of a strike, length of an unemployment spell, length of a stay in a hospital depending on the type of insurance, spacing of births, time spent off drugs while fighting addiction, etc.*

The advantage of duration models is in their ability to handle time changing xs (both with respect to calendar and duration time), duration dependence, and right censoring. The models can also handle multiple exits and multiple states. Read Kiefer (1988), [G]22.5, [L].

14.1. Hazard Function

Duration models build upon the concept of a hazard function $\lambda(t)$, which is defined as the probability of leaving a given state at duration t *conditional* upon staying there up to that point. Using this definition one can build a likelihood function for the observed durations and estimate it using standard methods (MLE, GMM). For example, if the hazard does not depend on either xs or duration t, then we can express the unconditional probability of observing a spell of duration t, denoted $f(t)$ as $f(t) = \lambda(1-\lambda)^{t-1}$. The probability that a spell lasts at least T periods is called survival $S(T) = \Pr[t \geq T] = 1-F(t) = (1-\lambda)^{T-1}$. This type of spell, where we do not observe the end of the spell, is called *right censored*. A *left censored* spell occurs when we do not observe the first part of the spell, but do observe when it ended. What makes a tremendous difference is whether we know when a left censored spell started or not. Of course $\lambda(t) = \frac{f(t)}{S(T)}$.

Exercise 14.1. *Suppose the hazard depends on t and write down the likelihood contribution for a completed spell and for a right censored spell. Next assume that there is no duration dependence and write down the likelihood contribution of a left censored spell. Finally, how would your last answer differ in presence of duration dependence, depending on whether you know when a left censored spell started.*

[138] Think of how we built a model from Poisson distribution as the natural model for count processes.

Remark 110. *A first approximation to the hazard, ignoring both observed and unobserved differences is the so called Kaplan-Meier statistic (also called empirical hazard):*

$$\lambda(t) = \frac{\#[exit(t)]}{\#[risk(t)]} \text{ with } \sigma_\lambda(t) = \sqrt{\frac{\lambda(t)(1-\lambda(t))}{\#[risk(t)]}}. \tag{14.1}$$

Exercise 14.2. *Verify the formula for σ_λ. Also, think of how you would estimate the empirical hazard in a case of competing risks, i.e., when there are two or more ways how to leave a given state.*

One can use either **discrete time or continuous time** hazard models. In a discrete time model, the transition can occur at most once in a given time period, i.e., these models depend on the unit of the time interval. In a continuous time model

$$\lambda(t) = \lim_{h \to 0} \frac{1}{h} \Pr(t \le t^* < t + h \mid t^* \ge t) \tag{14.2}$$

A widely used continuous time model is the *proportional hazard* model, $\lambda_i(t) = \exp(h(t))\exp(x_i'\beta) = \lambda_0(t)\exp(x_i'\beta)$, where $\lambda_0(t)$ is the so called baseline hazard.

Remark 111. *Note that in continuous time, the hazard equals*

$$-\frac{d\ln S(t)}{dt} = -\frac{d\ln[1-F(t)]}{dt} = \frac{f(t)}{1-F(t)} = \lambda(t),$$

which implies that

$$S(t) = \exp - \int_0^t \lambda(\tau)d\tau, \text{ and } f(t) = \lambda(t)\exp - \int_0^t \lambda(\tau)d\tau.$$

Example 14.2. *One possible choice of a discrete time hazard is the logit specification:*

$$\lambda_j(t, x_t|\theta_k^j) = \frac{1}{1 + e^{-h_j(t,x_t|\theta_k^j)}}$$

where $h_j(t, x_t|\theta_k^j) = \beta_j' x_t + g_j(t, \gamma_j) + \theta_k^j$. Here, $g_j(t, \gamma_j)$ is a function capturing the duration dependence.[139]

Exercise 14.3. *Can the logit model be interpreted as an approximation to a proportional hazard model?*

Remark 112. *One can trick* LIMDEP *or other software to estimate the logit duration model.*

---

[139] For proportional hazard models, Elbers and Ridder show that the heterogeneity distribution and the duration dependence can be separately identified.

14.2. Estimation Issues

First, there is a possibility of the so called *length-biased (stock) sampling:* correct sampling is from inflow during a certain time window (sampling frame). Sampling from stock oversamples long spells (somebody starting a quarter ago with a short spell will not be in today's stock).

Second, *left censored spells* with an unknown date of start create a difficult estimation problem (see Exercise 14.1 and below).[140]

Third, it is well known that in the presence of *unobserved person-specific characteristics* affecting the probability of exit, all of the estimated coefficients will be biased.[141]

One of the widely used methods of controlling for unobserved factors is the flexible semi-parametric heterogeneity MLE estimator proposed by Heckman and Singer (1984) (also called NP-MLE as in non-parametric MLE). They show that if there is a parametric continuous distribution of unobservables, the estimated distribution has to be that of a discrete mixing distribution with a step function nature. (Think of random effect probit.) Using simulations, a small number of points of support has been shown to remove the bias in βs. There is no known way of correctly constructing the asymptotic standard errors, since the dimension of the parameter space depends the sample size. So assume the number of points of support is fixed to invoke standard asymptotics and determine that actual number of points of support of the distribution of unobservables from the sample likelihood.[142]

[140]We can fix things if we know start of spell unless there are unobservables, which would lead to dynamic distorsion of the distribution of unobservables by selection on who of the left-censored makes it into the sample.

[141]Here, we are concerned with the effects of unobserved heterogeneity in duration models. For an example of similar methods in other settings, see Berry, Carnall and Spiller (1995), where they explicitly allow for two types of airline customers (businessmen vs. tourists), unobserved by the econometrician.

[142]Two simulation studies, one already published in the *Journal of Econometrics*, are important here. Baker and Melino (2000) suggest using Schwarz or Akaike criterion for picking the number of points of support. The most recent evidence on this issue comes from Xianghong Li and Barry Smith, who also provide guidance on the performance of different types of optimization algorithms with respect to finding the global optimum (depending on starting values). They advocate the use of a step function for duration dependence and prefer the simulated annealing (SA) algorithm in place of derivative-based optimization techniques. They also suggest a bootstrap procedure for choosing the number of support points and argue that likelihood ratio tests may still be appropriate for this purpose.

Remark 113. *The heterogeneity bias in duration dependence coefficients has been shown to be negative. To see why, think of two flat hazards* $\lambda_{M/S}(t)$ *of married and single women and construct the empirical hazard in absence of the marital status info.*

Remark 114. *Note that if there is no variation in the* x*s independent of duration, identification will be difficult.*

Remark 115. *Competing risk models are generally unidentified in the sense that for every dependent distribution of time to exit by cause, one can find an independent one that is observationally equivalent. Typically, people assume independence of causes conditional on X (random censoring) or they assume a parametric model.*

Example 14.3. *Honore and Lleras-Muney (Econometrica, 2006) look at the competing risk of dying of cardiovascular diseases (CVD) and cancer. Age-adjusted mortality rate from cancer has been flat for 30 years and some view this as evidence on little progress on cancer. However, if the causes of CDV and cancer are dependent, this interpretation is wrong. Honore and Lleras-Muney make no assumptions of the joint distribution of the underlying durations and estimate bounds on the marginal distributions of each of the competing duration variables.*

Example 14.4. *A recently popular* timing-of-events *approach (Abbring and van den Berg, 2003, Econometrica) based on the mixed proportional hazard (MPH) model is used when estimating the effect of a treatment (assigned without randomization) on exit from a given state, when we know the exact time of assignment to treatment, which occurs during an on-going spell in a given state. Identification is established of the causal relationship between the two durations (outcome duration, duration to treatment) in absence of IVs or CIA (based on observables only), based on assuming (i) MPH structure, (ii) no anticipation, and (iii) conditional independence when conditioning on both observables and unobservables, which are estimated using the NP-MLE approach. As usual, multiple-spell data make identification easier.*[143]

[143] In a Monte Carlo study Gaure, Røed and Zhang (2007) argue that separating causality (of some treatment and/or duration dependence) from selection (on unobservables) within non-experimental duration data by means of estimating the mixing discrete distributions of unobservables (i.e., by means of the non-parametric maximum likelihood estimator, NPMLE) is a piece of cake (and restrictions on the number of support points proposed by Baker and Melino (2000) in a single-spell framework may cause more harm than good).

14.2.1. Flexible Heterogeneity Approach

Let us concentrate on a discrete time logit hazard model. We need to allow the likelihood to pick up the presence of unobservable person-specific heterogeneity. To use the "random effects" approach, estimate a discrete mixing distribution $p(\theta)$ of an unobserved heterogeneity term θ as a part of the optimization problem. In doing so, one can approximate any underlying distribution function of unobservables.

More specifically, let $\lambda_j(t, x_t|\theta_k^j)$ be the conditional probability (hazard) of leaving a given state at time (duration) t for someone with person specific characteristics x_t, conditional upon this person having the unobserved factor θ_k^j, $k = 1, 2, ..., N_\theta^j$. The j subscript stands for the different ways of leaving a given state and serves, therefore, as a state subscript as well. For example one can leave unemployment for a $n$ew job or for a $r$ecall, in which case $j \in \{r, n\}$, or one can leave employment through a $q$uit or through a $l$ayoff, in which case $j \in \{q, l\}$. This is often referred to as a *competing risk model*. Below, we will use the example of quit, layoff, recall and new job. See also the discussion in [P]6.5.1.

To give an example of how the sample likelihood is evaluated using the concept of a hazard function, assume away any complications arising from the competing risks for now. Let λ denote the overall hazard out of a given state. In the absence of any unobserved heterogeneity, the likelihood function contribution of a single employment spell which ended at duration t would be

$$L_e(t) = \lambda(t, x_t) \prod_{v=1}^{t-1} [1 - \lambda(v, x_v)]. \tag{14.3}$$

In a competing risks specification with layoff and quit hazards (not allowing for unobserved factors), the unconditional probability of someone leaving employment through a quit at duration t would become

$$L_e^q(t) = \lambda_q(t, x_t) \prod_{v=1}^{t-1} [1 - \lambda_q(v, x_v)][1 - \lambda_l(v, x_v)], \tag{14.4}$$

where λ_q and λ_l denote the quit and layoff hazards respectively. Similarly, for someone who gets laid off in week t of an employment spell, the likelihood contribution becomes

$$L_e^l(t) = \lambda_l(t, x_t) \prod_{v=1}^{t-1} [1 - \lambda_q(v, x_v)][1 - \lambda_l(v, x_v)]. \tag{14.5}$$

Hazard models are natural candidates for dealing with the problem of right-censoring. For an employment spell which is still in progress at the end of our sampling frame (i.e., no transition out of employment has been observed), one enters the survival probability

$$S_e(T) = \prod_{v=1}^{T} [1 - \lambda_q(v, x_v)][1 - \lambda_l(v, x_v)]. \tag{14.6}$$

Here, T denotes the highest duration at which we observe the spell in progress and $S_e(T)$ gives the probability of a given spell lasting at least T periods. The sample likelihood then equals the product of individual likelihood contributions. Now, if we introduce the unobserved heterogeneity, the likelihood function contribution for someone leaving $u$nemployment at duration t for a $n$ew job would be

$$L_u^n(t) = \sum_{k=1}^{N_\theta^n} \sum_{m=1}^{N_\theta^r} p(\theta_k^n, \theta_m^r) L_u^n(t|\theta_k^n, \theta_m^r), \tag{14.7}$$

where $p(\theta_k^n, \theta_m^r)$ is the probability of having the unobserved components θ_k^n and θ_m^r in the new job and recall hazards respectively, and where

$$L_u^n(t|\theta_k^n, \theta_m^r) = \lambda_n(t, x_t|\theta_k^n) \prod_{v=1}^{t-1} [1 - \lambda_n(v, x_v|\theta_k^n)] [1 - \lambda_r(v, x_v|\theta_m^r)]. \tag{14.8}$$

The likelihood of leaving an employment spell in week s, denoted $L_e(s)$, is specified in a similar fashion (with quit and layoff being the different reasons for exit here).

The previous discussion focuses on examples with a single spell of each type. Equation 14.9 gives the likelihood contribution of a person with two completed spells of employment. The first spell starts in week $t+1$ and ends with a layoff in week s (at duration $s-t$); the second spell starts in week $r+1$ and ends with a quit in week w (at duration $w-r-s-t$).

$$L(s, w) = \sum_{k=1}^{N_\theta^q} \sum_{m=1}^{N_\theta^l} p(\theta_k^q, \theta_m^l) L_e^l(s|\theta_k^q, \theta_m^l) L_e^q(w|\theta_k^q, \theta_m^l) \tag{14.9}$$

Here θ^q and θ^l denote the unobserved terms entering quit and layoff hazards respectively and

$$L_e^l(s|\theta_k^q, \theta_m^l) = \lambda_l(s, x_s|\theta_m^l) \prod_{v=t+1}^{s-1} [1 - \lambda_q(v, x_v|\theta_k^q)] [1 - \lambda_l(v, x_v|\theta_m^l)] \ , \tag{14.10}$$

$$L_e^q(w|\theta_k^q,\theta_m^l) = \lambda_q(w,x_w|\theta_m^l)\prod_{v=r+1}^{w-1}\,[1-\ \lambda_q(v,x_v|\theta_k^q)]\ [1-\lambda_l(v,x_v|\theta_m^l)]\ .$$

Using multiple spell data provides greater variation and improves identification of the unobserved heterogeneity distribution (need to separate duration dependence from unobserved heterogeneity). However, use of this type of data raises the possibility of *selection bias*; i.e., the individuals with more than one spell of either type may be a non-random sample. To control for this problem, one can estimate the whole duration history of all states jointly while allowing the unobserved heterogeneity to be correlated across these spells. To continue in the example we used up to now, the unemployment and employment hazard have to be estimated *jointly* in order to control for selection bias into multiple spells. One has to take into account the joint density of the unobservables across the two hazards, denoted by $p(\theta^u,\theta^e)$. Suppose we want to estimate a competing risks specification for quits and layoffs jointly with an overall hazard for unemployment. The likelihood contribution of someone leaving the first unemployment spell after t weeks, then getting laid off after $s-t$ weeks on a job and staying in the second unemployment spell till the date of the interview, say at $T-s-t$ weeks into the last spell, then becomes

$$L^{u,l,u}(t,s,T) = \sum_{k=1}^{N_\theta^u}\sum_{m=1}^{N_\theta^q}\sum_{n=1}^{N_\theta^l} p(\theta_k^u,\theta_m^q,\theta_n^l)L_u(t|\theta_k^u)L_e^l(s|\theta_m^q,\theta_n^l)S_u(T|\theta_k^u), \quad (14.11)$$

where

$$L_u(t|\theta_k^u) = \lambda_u(t,x_t|\theta_k^u)\prod_{v=1}^{t-1}\,[1-\ \lambda_u(v,x_v|\theta_k^u)]\ .$$

The employment contribution, L_e^l is defined in equation 14.10 . Finally

$$S_u(T|\theta_k^u) = \prod_{v=s+1}^{T}[1-\lambda_u(v,x_v|\theta_k^u)]$$

is the survivor function expressing the probability of a given spell lasting at least T periods.

One can compute individual contributions to the sample likelihood for other labor market histories in a similar way. The number of points of support of the distribution of unobservables (N_θ^u, N_θ^q and N_θ^l) is determined from the sample likelihood.[144] Note the assumption of θ^u, θ^q and θ^l staying the same across multiple unemployment and employment spells respectively. There are many possible choices for the distribution of unobservables:

Heterogeneity Distributions

1. Independent Heterogeneity: $p(\theta^u, \theta^e) = p_u(\theta^u)p_e(\theta^e)$

2. Bivariate Heterogeneity Distribution:

	θ_1^l	θ_2^l	...	θ_N^l
θ_1^q	$p(\theta_1^q, \theta_1^l)$	$p(\theta_1^q, \theta_2^l)$	...	$p(\theta_1^q, \theta_N^l)$
θ_2^q	$p(\theta_2^q, \theta_1^l)$	$p(\theta_2^q, \theta_2^l)$	...	$p(\theta_2^q, \theta_N^l)$
...	...	...	...	...
θ_M^q	$p(\theta_M^q, \theta_1^l)$	$p(\theta_M^q, \theta_2^l)$	...	$p(\theta_M^q, \theta_N^l)$

3. One factor loading: pairs of $\{\theta^l, \theta^q\}$ such that

$p(\Theta_1)$	$\Theta_1 = \{\theta_1^l, c\theta_1^l\}$
$p(\Theta_2)$	$\Theta_2 = \{\theta_2^l, c\theta_2^l\}$
...	...
$p(\Theta_N)$	$\Theta_N = \{\theta_N^l, c\theta_N^l\}$

4. Heterogeneity distribution with 3-tuples (corresponding to one way of leaving unemployment and 2 ways of leaving employment.)

$p(\Theta_1)$	$\Theta_1 = \{\theta_1^u, \theta_1^l, \theta_1^q\}$
$p(\Theta_2)$	$\Theta_2 = \{\theta_2^u, \theta_2^l, \theta_2^q\}$
...	...
$p(\Theta_N)$	$\Theta_N = \{\theta_N^u, \theta_N^l, \theta_N^q\}$

5. 'Stayer' heterogeneity: Suppose that we want to allow for the possibility of never leaving employment through a quit (or for the possibility of never returning to a prison.) Assume, for now, that the only way to transit out of employment is to quit. Furthermore, assume that there is no unobserved

[144] Simulation provide important guidance. See Baker and Melino (2000) and more importantly Xianghong and Smith (2009).

heterogeneity. A typical stayer model would then parametrize an individual's contribution to the likelihood as follows:

$$L(t) = p_s + (1 - p_s)\{\lambda_q(t, x_t) \prod_{v=1}^{t-1} [1 - \lambda_q(v, x_v)]\ \},$$

where p_s is the probability of never leaving employment and λ_q is a quit hazard. See Jurajda (2002) for details on estimation.

6. Continuous parametric distributions of heterogeneity, for example Weibull.

14.2.2. Left Censored Spells

We need to know when they started. In presence of unobserved heterogeneity, dropping left censored spells will cause bias. See Ham and Lalonde (1997) for an example where the bias matters. Heckman and Singer (1984) suggest to model the interrupted spells with a separate hazard, i.e., a new hazard with a different β from the fresh spells. See also exercise 14.1.

14.2.3. Expected Duration Simulations

How to evaluate the magnitude of coefficients? Use the unconditional probability of leaving a given state to compute the expected durations under different values of xs. Interpret the difference between expected durations as the magnitude of the particular β. The expected duration is computed as

$$E(t|X) = \sum_{i=1}^{I} \frac{\sum_{t=1}^{\infty} t f_i(t)}{I}, \qquad (14.12)$$

where I is the number of spells in the sample, x_{it} is the vector of all explanatory variables for a spell i at duration t, and X represents the collection of all x_{it} vectors.[145] Finally, using the example of a $r$ecall and $n$ew job hazard out of unemployment, the unconditional probability of leaving unemployment at duration t denoted $f_i(t)$ is computed as follows:

[145] A simpler (biased) approach is to evaluate the expected duration at a mean individual $\overline{x}$.

$$
\begin{aligned}
f_i(t) &= \sum_{k=1}^{N} p(\theta_k^r, \theta_k^n) f_i(t|\theta_k^r, \theta_k^n), \text{ where} \\
f_i(t|\theta_k^r, \theta_k^n) &= \{\lambda_r(t, x_{it}|\theta_k^r) + \lambda_n(t, x_{it}|\theta_k^n) - \lambda_r(t, x_{it}|\theta_k^r)\lambda_n(t, x_{it}|\theta_k^n)\} \times \\
& \prod_{v=1}^{t-1} [1 - \ \lambda_r(v, x_v|\theta_k^r)] \ [1 - \ \lambda_n(v, x_v|\theta_k^n)] \ .
\end{aligned}
$$

Remark 116. *In multiple-state multiple-spell data, single-spell duration simulations do not provide a full picture. See, e.g., Jurajda (2002).*

14.2.4. Partial Likelihood

Cox (1972, 1975): estimate β in the proportional hazard model $\lambda_i(t) = \lambda_0(t)\exp(x_i^{'}\beta)$, without specifying the form of the baseline hazard $\lambda_0(t)$. Order the completed durations t_i into $t_{(i)}$. The conditional probability that individual 1 concludes a spell at time $t_{(1)}$, given that all other individuals could have completed their spells at that duration is

$$
\frac{\lambda(t_{(1)}, x_{(1)})}{\sum_{i=1}^{n} \lambda(t_{(1)}, x_{(i)})} = \frac{\exp(x_{(1)}^{'}\beta)}{\sum_{i=1}^{n} \exp(x_{(i)}^{'}\beta)}. \quad (14.13)
$$

In the absence of information about the form of duration dependence, only the information about the *order* of the spell durations is used.

Remark 117. *This method alone does not allow the expected duration simulations. It is possible, though, to construct a nonparametric estimate of the baseline hazard using the estimated* $\exp(\widehat{x_i^{'}\beta})$. *See [P].*

Part IV
Selected Additional Topics

15. Structural Estimation

Is possible without closed form solutions. It has been applied to (RE) dynamic models of discrete choice (for example ICAPM) by Miller (1984), Wolpin (1984), Pakes (1986), and Rust (1987). For not so recent surveys see Eckstein and Wolpin (1989) and Rust (1992, 1994).

Example 15.1. *In a stopping problem Hotz and Miller (1993) provide a new method of estimating dynamic discrete choice models, not requiring evaluation of the value functions. They can estimate the parameters without the need to solve the problem numerically using an inversion results between* conditional choice probabilities *(which one can estimate from the cell data) and a difference of the value functions.*

Example 15.2. *Other applications include equilibrium models of unemployment (e.g. van den Berg and Ridder 1993) or local jurisdictions*[146] *(Epple and Sieg 1996).*

Recently, researchers combine (more reasonable) structural model estimation, which allows for the generation of policy predictions, with exogenous (natural-experiment) identification. You either test your structure using the experiment or identify it using the experiment.

Example 15.3. *See, e.g., papers by Attanasio, Meghir & Santiago (in progress) or Todd and Wolpin (in press in AER) on the Progressa experiment in Mexico or the discussion of the use of structural estimation for evaluating labor-supply effects of earned-income-tax-credit policies by Blundell (2005, Labour Economics)*[147] *Todd and Wolpin: The Progressa experiment offered one level of subsidies to parents who send their children to school in Mexican villages. Todd and Wolpin*

[146] Look at an equilibrium distribution of households by income across communities.

[147] http://www.ucl.ac.uk/~uctp39a/Blundell%20-%20Adam%20Smith%20Lecture.pdf

estimate a model where parents make sequential decisions about sending their children to school or to work, as well as about the timing and spacing of births. The model is estimated off the control group and it predicts well the response of the treatment group to the subsidy; it can therefore be used to infer the optimal size of the subsidy. Because in absence of treatment, there is no direct cost of schooling, they use observed child wages (the opportunity cost of attending school) to identify the model. The model is solved numerically, integrals are simulated, likelihood contributions are calculated by a smoothed frequency simulator.

16. Nonparametrics

The very opposite of the structural models. We already mentioned the use of semi-parametric methods in the estimation of discrete choice models (section 10.1.5). We apply non-parametric approaches when matching on unobservables (as in the selection bias model of Section 12.2.4) as well as observables (see section 13.2). Here, we will discuss the basic non-parametric methods.

16.1. Kernel estimation

A typical OLS regression will use information from the whole range of $x \in [\underline{x}, \overline{x}]$ to estimate $E[y_i \mid x = x_i] = \beta' x_i$. Here, we will estimate a conditional expectation function $E[y \mid x] = m(x)$ using 'local' information from an area $A(x)$ 'close' to x:

$$\widehat{E[y \mid x]} = \widehat{m(x)} = \frac{\sum_{i=1}^{n} I\{i \in A(x)\} y_i}{\sum_{i=1}^{n} I\{i \in A(x)\}} = \sum_{i=1}^{n} w_i(x) y_i.$$

Two questions: (i) how to define $A(x)$, (ii) are the weights $w_i(x)$ from above optimal. Instead of the indicator function $I\{\cdot\}$ let us use a bounded, symmetric *Kernel* function $K(\cdot)$ such that $\int K(u)du = 1$. For asymptotic theory on choosing the optimal Kernel and bandwidth[148], see [N] and Silverman (1986).

16.2. K-th Nearest Neighbor

Define $J(x) = \{i : x_i \text{ is one of the } K \text{ nearest neighbors}\}$ and use $w_i(x) = \frac{1}{K}$ if $i \in J(x)$. Kernel estimation lets precision vary and keeps bias constant. KNN does the opposite.

[148] The bandwidth can be also data-dependent.

16.3. Local Linear Regression

See Fan and Gijbels (1996). Kernel estimation has problems at the boundary of the space of x which LLR is able to remedy.

$$\widehat{m(x_0)} = \widehat{\alpha}, \text{ where } \widehat{\alpha} = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \{y_i - \alpha - \beta(x_i - x_0)\}^2 K\left(\frac{x_i - x_0}{a_n}\right)$$

The kernel K and a_n are chosen to optimize the asymptotic MSE.[149]

Kernels used in practice are:

- Epanechnikov: $K(u) = \frac{3}{4}(1 - u^2)I\{\mid u \mid \leq 1\}$ (optimal K in both LLR and Kernel estimation, optimal a_n differ)
- Quartic: $K(u) = \frac{15}{16}(1 - u^2)^2 I\{\mid u \mid \leq 1\}$
- Triangle: $K(u) = (1 - \mid u \mid)I\{\mid u \mid \leq 1\}$

The choice of a_n can be made using

- a point-wise plug-in method which relies on an initial estimate,
- a cross-validation method which chooses *global* a_n to minimize the MSE $\sum_{i=1}^{n}(y_i - \widehat{m_i}(x_i))^2$.
- a fishing expedition: increase a_n as long as linearity is not rejected.

Remark 118. STATA *has a kernel smoother, kernel density estimation (*`kdensity`*), and does local linear regression. Advanced programs are available on the www for* S-PLUS *(http:\\lib.stat.cmu.edu\).*

Remark 119. *Kernel estimation is basically a LLR on just the constant term.*

Remark 120. *LLR are used in the regression discontinuity design (see Section 7.1).*

[149] The bias in Kernel estimation depends on the distribution of regressors and on the slope of the regression function. The LLR bias only depends on the second derivative of the regression function. The asymptotic variance of the two methods is close unless data is sparce or m is changing rapidly around the x_0 data point.

Remark 121. *There are also extensions of the localization idea to the MLE framework, see Tibshirani and Hastie (1987).*

Remark 122. *There are other local regressions. For example, see the* LOWESS *procedure in* S-PLUS. *See Fan and Gijbels (1996) book (p.201) for local versions of quantile regressions.*

16.4. Multidimensional Extensions and Semiparametric Applications

The curse of dimensionality is severe. To have a reasonable speed of convergence we need very large samples. There are a few ways how to proceed:

- Regression trees: recursively split x to estimate step functions; derive a stopping rule to minimize mean square error.
- Impose additive separability or Projection pursuit regression:

$$m(x) = g_1(x\beta_1) + g_2(x\beta_2) + \ldots.$$

- Partial Linear Model: For a model $y = z\beta + f(x) + \varepsilon$, where both z and x are scalars, estimators of β can be constructed which are asymptotically normal at the $\sqrt{n}$ speed of convergence. See Yatchew, A. (1998).
- Average derivative estimation: If I am interested in $\theta = E\left\{\frac{\partial m(x_i)}{\partial x_i}\right\}$ then θ can be estimated with $\sqrt{n}$ speed of convergence. Example: binary choice or Tobit models.
- Index sufficiency. See the semi-parametric Heckman's λ application by Powell (1989) in Section 12.2.4 or matching on propensity score in Section 13.2.
- See Athey and Imbens (Econometrica, 2006)[150] for a generalization of the difference-in-differences method (see Section 6.2). They relax linearity and allow treatment effects to vary across individuals and average treatment effects to vary across groups (such as states that do and don't adopt some policy).

[150] http://kuznets.fas.harvard.edu/%7Eathey/CIC.pdf

16.5. Quantile Regression

A regression method robust to outliers and censoring. Let's start with the 50th quantile. Sometimes we want to fit the regression line through medians of $y|x$, not the means, as OLS does. The LAD estimator

$$\min_{\beta} \sum_{i=1}^{n} \left| y_i - x_i^{'}\beta \right|$$

corresponds to such median regression. One goal of this regression method is to reduce the influence of outliers (the LAD regression is not affected by taking a value of one y that is above the median and multiplying it by 1000). Quantile regressions only use ordinal information; see, e.g. the use of CLAD in censored data in Section 11.3.

However, note that OLS and LAD will only give the same answer when the distribution of $\epsilon|x$ is symmetric around zero. A comparison of LAD and OLS when y is a measure of wealth (a skewed y) therefore does not necessarily speak about the importance of outliers.

Now consider estimating regression parameters $\beta(q)$ corresponding to any quantile q you are interested in:

$$\min_{\beta(q)} \sum_{i=1}^{n} \left(y_i - x_i^{'}\beta \right) \left(q - I \left[y < x_i^{'}\beta \right] \right).$$

The objective function is piece-wise linear and continuous. For intuition on this objective, think first of $q = 0.5$. In this case, the second bracket is either $+0.5$ or -0.5.[151] When you are above (below) the regression line, you get 0.5 (-0.5) multiplying a positive (negative) ϵ. So, the objective function is just trying to minimize the sum of absolute values of ϵs (LAD), i.e., fit the line through (conditional) medians. Now think of $q = 0.1$. When you are above (below) the quantile regression line, you multiple $|\epsilon|$ by 0.1 (0.9). Why? You are minimizing and you want to punish for predicting below the line, because you want 90% of the points to be above the line.

You need to bootstrap for standard errors. There are now IV methods for quantile regressions as well as panel data quantile regressions.[152]

[151] You can think of this the *sgn* fuction. Remember the MRE (Section 10.1.5)?

[152] For the quantile IV regressions see Abadie, Angrist and Imbens (2002, Econometrica) or Powel (1983). See Section 14 of the Imbens/Wooldridge NBER Summer '07 course for panel

Remark 123. *Quantile regressions tell us about y distributions (for example, wage inequality) within x groups. Quantile regression coefficients speak about effects on distributions, not on individuals. For example, if T_i has a positive effect on a lower decile of the income distribution, this does not mean that someone who would have been poor will be less poor with treatment T_i, but that those who are poor with treatment are less poor than they would have been without treatment.*

17. Miscellaneous Other Topics

Bootstrap. A simulation-based set of techniques which provide estimates of variability, confidence intervals and critical levels for test procedures. They are used when asymptotic results are not available or hard to compute. Also, they may turn out to be more accurate than asymptotic theory because they are constructed based on the right sample size (i.e., the bootstrap sampling distribution of test statistics may be closer to the finite-sample distribution of interest than the corresponding asymptotic approximation; see Hall's book from 1992 or Horowitz, 2001). The idea is to create k replications of the original data set of size N by randomly drawing N data points from it with replacement. The model is re-estimated on each simulated sample and the variation in $\widehat{\beta}$ over k is used to answer questions about its distribution etc. In the residual bootstrap the resampling population is not the data set, but $\widehat{\epsilon}$.

Gibbs Sampler. A Bayesian approach to estimation introduced by Geman and Geman (1984). Related methods: data augmentation, Metropolis algorithm. Unlike Newton-Raphson, these methods allow us to obtain the *marginal* of the likelihood function or posterior density. Alternatively, they can be used to obtain *a sample of parameter values.* The idea is to draw from the joint distribution by drawing successively from various conditional distributions to avoid direct evaluation of the likelihood. These methods require a random input stream and iteration. See Tanner (1993). For an example of multinomial probit estimation see McCulloch and Rossi (1994).

For a 2007 mini-course overview of what's new in econometrics, which covers

---

data applications. In particular, the correlated random-effects quantile regression of Abrevaya and Dahl (2007) and the penalized fixed effects estimator of Koenker (2004). See Remark 88 for a reference to a sample selection correction for quantile regressions. Machado and Mata (2005, JAppliedEcm) provide a Oaxaca-Blinder style gap decomposition for quantile regressions.

Of course, when the IVs vary at group level and we care about y distributions within groups, we can just run 2SLS at the group level with the group-specific quantiles on the LHS.

the topics we discussed in this course but also many others, see http://www.nber.org/WNE/WNEnotes.pdf.

18. A Check-list for Empirical Work

Here are some tips for how to do (and present) empirical research, based largely on John Cochrane.[153]

Based on this course, you should understand that the three most important things for empirical work are Identification, Identification, Identification. So, in your written work, describe your identification strategy clearly. In particular:

1. Explain what *economic mechanism* caused the dispersion in your right hand variables.

2. Explain what *economic mechanism* constitutes the error term. What things other than your right hand variable cause variation in the left hand variable? (For example, is it prediction error?) This will help you explain in *economic terms* why you think the error term is uncorrelated with the right hand variables.

3. Explain the *economics* of why your instruments are correlated with the right-hand-side variable (sufficiently so that they are not weak) and not with the error term.[154] If your instrument has no economics to it, only quasi-randomized assignment, is it not only *external* (see Remark 32) but also *exogenous*? Will the variation in treatment within the group assigned to treatment be random or related to the size of the effect?

4. Are you sure causality doesn't run from y to x, or from z to y and x simultaneously? (Do interest rates cause changes in housing demand or vice versa

[153] http://faculty.chicagogsb.edu/john.cochrane/research/Papers/phd_paper_writing.pdf
He also provides many useful writing and presentation tips, such as: get to the main result of the paper as fast as possible (both in a presentation and in the text of the paper), move everything that's not essential to the main story line of the paper to an appendix, use active and present tense, simple words and sentences, etc.

[154] Also, do you understand the difference between an instrument and a control? In regressing y on x, when should z be used as an additional variable on the right hand side and when should it be an instrument for x? How would you use an ability measure when running a wage regression with education on the right-hand side (i.e., when you are worried about unobservable ability bias)?

(or does the overall state of the economy cause both to change)? Does the size of police force affect crime rates or vise versa?)

5. Describe the source of variation in the data that drives your estimates, for every single number you present. (Think of fixed effects vs. random effects.)

6. Are you sure you are looking at a demand curve, not a supply curve? As one way to clarify this question, ask "whose behavior are you modeling?" Example: Suppose you are interested in how interest rates affect housing demand, so you run the number of new loans on interest rates. But maybe when housing demand is large for other reasons, demand for mortgages drives interest rates up. Are you modeling the behavior of house purchasers or the behavior of savers (how savings responds to interest rates)?

7. Consider carefully what controls should and should not be in the regression. You may not want to include all the "determinants" of y on the right hand side. High R2 is usually bad — it means you ran y on a version of y and some other x. Do you want to condition on group fixed effects in situations when you do not need to remove group unobservables? If you do include them, you are asking about the effect of x within these groups, not across. (Think of wage regressions with industry dummies.) Similarly in matching exercises: economic theory should tell us what x need to be controlled for.

8. For every parametric or distributional assumption (a column in your table of results based on traditional techniques) provide a more flexible (semi-parametric) version to assess the sensitivity to arbitrary assumptions.

9. Explain the *economic* significance of your results.[155] Explain the economic

[155] But watch out: Economists often estimate log-linear (Mincerian, Euler, production-function, gravity) equations and present elasticities that are easy to interpret in terms of their magnitude. However, see Santos Silva and Tenreyro (2006) for how misleading such elasticities estimated by OLS from log-linearized models can be in the presence of heteroscedasticity.

To see why, start with a constant-elasticity model $y_i = \exp(\alpha)x_{1i}^{\beta_1}x_{2i}^{\beta_2}\epsilon_i$ or simply $y_i = \exp(\alpha)\exp(x_i^{'}\beta)\epsilon_i$ where $E[\epsilon_i|x_i] = 1$ and, assuming that $y_i > 0$, arrive at $\ln(y_i) = \alpha + x_i\beta + \ln(\epsilon_i)$. Now, the Jensen's inequality ($E[\ln y] \neq \ln E[y]$) implies that $E[\ln(\epsilon_i)|x_i] \neq 0$.This residual expectation will be a constant (i.e., will affect only the constant α of the model) only under very strong assumptions on the distribution of ϵ (because of the non-linear transformation of the dependent variable, the conditional expectation of ϵ depends on the the *shape* of the conditional distribution of ϵ). Specifically, with non-negative y_i the conditional variance of y (and of ϵ) vanishes when y is near zero. With heteroscedasticity (with the variance of ϵ depending on x), magnitude of the central estimates, not just their statistical significance.[156]

You should be able to link these suggestions to the discussion of the search for exogenous variation and the sensitivity to parametric and distributional assumptions provided in Section 1.

the expectation $E[\ln(\epsilon_i)|x_i]$ will be a function of x, thus rendering OLS of the log-linearized model inconsistent. Solution? WNLLS on levels, i.e., minimize squares of $(y_i - \exp(x_i^{'}\beta))$, but one needs an assumption on the functional form of $V[y_i|x_i]$ to be able to weight efficiently (see exercise 10.3). They end up proposing a Poisson pseudo-maximum likelihood estimator.

[156] Ziliak and McCloskey (2008, U of Michigan Press) offer a million examples of how statistical significance is misused in empirical work and argue that one should primarily focus on economic size of effects. Of course, a large imprecisely estimated effect can be more important for the real world than a tiny precisely estimated one. Especially if several studies obtain a similar estimate: large, but statistically imprecise within an individual sample. It's also clear that sampling-based confidence intervals (having to do with noise in small samples) capture only one of many sources of error in estimation. Yet, statistical significance can help differentiate among competing models.

On top of the point 9 given in the main text, they offer several useful pieces of practical advice: (i) Adjust α to sample size and deal with the power of the test. (ii) Report coefficients in elasticity form or in a way that allows one to easily see the economic magnitude of the estimates. (iii) Do not drop from regression specifications statistically insignificant controls with economically large estimated effects. (iv) Present an independent simulation that would allow some perspective on whether the coefficient estimates are of reasonable magnitude.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2007) "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *National Bureau of Economic Research Working Paper*: 12831, January 2007, 51 p.

Abadie, Alberto, Joshua Angrist, and Guido Imbens (2002) "Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings," *Econometrica*: 70(1), Jan. 2002, pp. 91–117.

Abadie, Alberto, and Guido Imbens (2002) "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," *National Bureau of Economic Research Technical Paper*: 283, October 2002, 57 p.

Abowd, John M., Francis Kramarz, and David N. Margolis (1999) "High wage workers and high wage firms," *Econometrica* 67(2), March 1999: pp. 251–333.

Abrevaya, J. and Dahl, C.M. (2008) "The effects of birth inputs on birthweight," *Journal of Business and Economic Statistics*: 26(4), pp. 379–397.

Achen, Christopher H. (1986) *The statistical analysis of quasi-experiments*. University of California Press, 1986, 186 p.

Ahn, H., L.J., Powell (1993) "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*; 58(1-2), July 1993, pp. 3-29.

Aigner, Dennis J. (1973) "Regression with a binary independent variable subject to errors of observation," *Journal of Econometrics*: 1(1), March 1973, pp. 49-59.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber (2005) "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*: 113(1), February 2005, pp. 151–184.

Altonji, Joseph G. and Lewis M. Segal (1994) "Small Sample Bias in GMM Estimation of Covariance Structures", *National Bureau of Economic Research Technical Paper*: 156, June 1994.

Amemiya T. (1984) "Tobit Models: A Survey," *Journal of Econometrics* 24(1-2), January-February 1984, pg: 3-61.

Amemiya T. (1985): *Advanced Econometrics*, Harvard U. Press, 1985.

Andrews, Schank and Upward (2006) "Practical fixed-effects estimation methods for the three-way error-components model," *Stata Journal* 6 (4), December 2006, pp. 461-481.

Andrews, D. and J. Stock (2005) "Inference with Weak Instruments," *Advances in Economics and Econometrics*, Vol III, Blundel„ Newey and Persson (eds.), 122-173.

Angrist, Joshua D. (1990) "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*; 80(3), June 1990, pages 313-36.

Angrist, Joshua D. (1991) "Grouped-data estimation and testing in simple labor-supply models," *Journal of Econometrics*: 47(2-3), February 1991, pp. 243-266.

Angrist, Joshua D. (1995) "Conditioning on the Probability of Selection to Control Selection Bias", *National Bureau of Economic Research Technical Paper*: 181, June 1995.

Angrist, Joshua D. (1998) "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica* 66(2), March 1998, pages 249-88

Angrist, Joshua D. (2004) "Treatment effect heterogeneity in theory and practice," *Economic Journal*: 114(494), March 2004, pp. C52-C83.

Angrist, Joshua D. and Hahn, J. (2004) "When to control for covariates? Panel asymptotics for estimates of treatment effects," *The Review of Economics and Statistics*: 86(1), pp. 58-72.

Angrist, Joshua D. and Alan B. Krueger (1992), "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association*, 87, 328–336.

Angrist, Joshua D. and Alan B. Krueger (1999) "Empirical Strategies in Labor Economics," Ashenfelter, Orley; Card, David, eds. *Handbook of Labor Economics*. Volume 3A. Handbooks in Economics, vol. 5. Amsterdam; New York and Oxford: Elsevier Science, North-Holland, 1999, pages 1277-1366.

Angrist, Joshua D. and Alan B. Krueger. (2001) "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives*; 15(4), Fall 2001, pages 69-85.

Angrist, Joshua D. and V. Lavy (1999) "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement", *Quarterly Journal of Economics* 114, pp. 533-575.

Angrist, J.D. and Pischke, J. (2009): *Mostly Harmless Econometrics: An Empiricist's Companion* , Princeton University Press, 2009.

Anselin (1988): *Spacial Econometrics: Methods and Models*, Kluwer Academic Press, 1988.

Arabmazar A., P. Schmidt (1981) "Further Evidence on the Robustness of the Tobit Estimator to Heteroskedasticity", *Journal of Econometrics*, 17(2), Nov. 1981, pg: 253-58.

Arellano, Manuel and Stephen Bond (1991) "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *The Review of Economic Studies*: 58(2), April 1991, pp. 277–297.

Arellano, Manuel and O. Bover (1995) " Another look at the instrumental variable estimation of error-components models," *Journal of Econometrics*: 68(1), July 1995, pp. 29–51.

Arellano Manuel, Bo Honoré: *Panel Data Models: Some Recent Developments*, <ftp://ftp.cemfi.es /wp/00/0016.pdf>

Arellano Manuel, C. Meghir (1992) "Female Labor Supply & On-the-Job Search: An Empirical Model Estimated Using Complementarity Data Sets," *Review of Economic Studies*, 59, 1992, 537-559.

Ashenfelter, Orley and David Card (1985) "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *The Review of Economics and Statistics*: 67(4), Nov. 1985, pp. 648-660.

Ashenfelter, O., Harmon, C. and Oosterbeek, H. (1999) "A review of estimates of the schooling/earnings relationship, with tests for publication bias, " *Labour Economics*, 6(4),pp. 453-470.

Ashenfelter, O. and A. Kruger (1994) "Estimates of the Economic Return to Schoolong from a New Sample of Twins," *American Economic Review* 84: 1157-1173.

Athey, Susan and Guido W. Imbens (2006) "Identification and Inference in Nonlinear Difference-in-Differences Models," *Econometrica*: 74(2), March 2006, pp. 431-497.

Attanasio, Meghir & Santiago (in progress) “Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to evaluate Progresa”

Aydemir, Abdurrahman and George J. Borjas (2006) “A Comparative Analysis of the Labor Market Impact of International Migration: Canada, Mexico, and the United States,” *National Bureau of Economic Research* (NBER) WP No. 12327, June 2006, 66 p.

Baker, M., A. Melino (2000) “Duration Dependence and Nonparametric Heterogeneity: A Monte Carlo Study,” *Journal of Econmetrics*, 96(2), June, pp.357-393

Baker, Michael and Nicole M. Fortin (2001). “Occupational Gender Composition and Wages in Canada, 1987-1988.” *Canadian Journal of Economics*, 34(2), 345-376.

Bauer, T. K. and Sinning, M. (2005) “Blinder-Oaxaca Decomposition for Tobit models,” *IZA* Discussion Papers: 1795.

Bauer, T. K. and Sinning, M. (2008) “An extension of the Blinder-Oaxaca Decomposition to nonlinear models, ” *AStA Advances in Statistical Analysis*: 92(2), May 2008, pp. 197-206.

Becker, S.O. and Ichino, A. (2002) “Estimation of average treatment effects based on propensity scores, ” *Stata Journal*, StataCorp LP, 2(4), November 2002, pp.358-377.

Berg, Van den, J. Gerard, G. Ridder (1993) “On the Estimation of Equilibrium Search Models from Panel Data,” van Ours, Jan C., Pfann, Gerard A., Ridder, Geert, eds. Labor demand and equilibrium wage formation. *Contributions to Economic Analysis*, vol. 213. Amsterdam; London and Tokyo: North-Holland, pages 227-45.

Berry, S., M. Carnall, T.P. Spiller, (1996) “Airline Hubs: Costs, Markups and the Implications of Customer Heterogeneity,” *National Bureau of Economic Research* WP 5561, May 1996, pp. 20.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2002) “How Much Should We Trust Differences-in-Differences Estimates?” *NBER* Working Paper No. 8841, March 2002, 39 p.

Björklund, Anders and Robert Moffitt (1987) “The Estimation of Wage Gains and Welfare Gains in Self-Selection Models,” *The Review of Economics and Statistics*: 69(1), Feb. 1987, pp. 42–49.

Blundell, Richard W. and Richard J. Smith (1986) “An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply,” *Econometrica*: 54(3), May 1986, pp. 679-685.

Blundell, Richard W., Pierre-André Chiappori, Thierry Magnac and Costas Meghir (2005) "Collective Labour Supply: Heterogeneity and Nonparticipation," *IZA* Discussion Paper No. 1785, September 2005, 58 p.

Blundell, Richard W. and James L. Powell (2004) "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*: 71(3), July 2004, pp. 655-679.

Blundell, Richard W. and James L. Powell (2007) "Censored regression quantiles with endogenous regressors," *Journal of Econometrics*: 141(1), November 2007, pp. 65-83.

Bound, J., D. Jaeger, and R. Baker (1995) "Problems with Instrumental Variables Estimation when the Correlation between Instruments and the Endogenous Explanatory Variable is Weak", *Journal of the American Statistical Association*, 90.

Bourguignon, F., Fournier, M. and Gurgand, M. (2007) "Selection bias corrections based on the multinomial logit model: Monte Carlo comparisons, " *Journal of Economic Surveys*: 21(1), Blackwell Publishing, Feb. 2007, pp. 174-205.

Brock, W.A. and Durlauf, S.N. (2001) "Discrete Choice with Social Interactions " *Review of Economic Studies*: 68(2), pp. 235-260.

Brock, W.A. and S.N. Durlauf (2007) "Identification of binary choice models with social interactions," *Journal of Econometrics*: 140(1), September 2007, pp. 52-75.

Buchinsky, M. (1998) "The dynamics of changes in the female wage distribution in the USA: a quantile regression approach," *Journal of Applied Econometrics*: 13(1), pp. 1-30.

Card, D. (1993) "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," *National Bureau of Economic Research* Working Paper: 4483, October 1993, pg. 26.

Card, D. and Alan B. Krueger (1994) "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania" *American Economic Review*: 84(4), Sep., 1994, pp. 772-793.

Carneiro, Hansen, and Heckman (2003) "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College," *National Bureau of Economic Research* Working Paper: 9546, March 2003, pg. 62.

Carneiro P., E. Vytlacil, J.J. Heckman (2002) "Estimating The Return to Education When It Varies Among Individuals": http://www.cerge.cuni.cz/events/seminars/ sem-fall01/011108_a.asp

Chamberlain G. (1984) "Panel Data," in *Handbook of Econometrics* vol. II, pp. 1247-1318. Amsterdam, North-Holland.

Chamberlain, G. (1980) "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*; 47(1), Jan. 1980, pages 225-38.

Chamberlain, G. (1982) "Multivariate Regression Models for Panel Data," *Journal of Econometrics* 18(1), Jan. 1982, pages 5-46.

Cameron, A. Colin, Gelbach, Jonah B., and Douglas L. Miller (2008) "Bootstrap-based improvements for inference with clustered errors," The *Review of Economics and Statistics*, August 2008, 90(3): 414–427.

Chay, Kenneth Y. and James L. Powell (2001) "Semiparametric Censored Regression Models," *The Journal of Economic Perspectives*: 15(4), Autumn 2001, pp. 29-42.

Chen, S. (1999) "Distribution-free estimation of the random coefficient dummy endogenous variable model," *Journal of Econometrics*: 91(1), July 1999, pp. 171-199.

Chernozhukov, V. and Hansen, C. (2008) "The reduced form: a simple approach to inference with weak instruments " *Economics Letters*: 100(1), pp. 68-71.

Chioda, L. and M. Jansson (2006) "Optimal Conditional Inference for Instrumental Variables Regression," unpublished manuscript, department of economics, UC Berkeley.

Cochran, W. G. and Rubin, D. B. (1973) "Controlling Bias in Observational Studies: A Review " *Sankhya*: Ser. A 35, pp. 417–446.

Conley and Taber (2005) "Inference with "Difference in Differences" with a Small Number of Policy Changes," *NBER* Technical Working Paper No. 312, July 2005, 51 p.

Cosslett S. (1981) "Maximum Likelihood Estimator for Choice-Based Samples", *Econometrica*, 49(5), Sept. 1981, pages 1289-1316.

Costa Dias, Monica, Hidehiko Ichimura and Gerard J. van den Berg (2008) "The Matching Method for Treatment Evaluation with Selective Participation and Ineligibles" *IZA* DP No. 3280, January 2008, 36 p.

Cox, D.R. (1972) "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2), 1972, pp. 187-220.

Cox, D.R. (1975) "Partial Likelihood," *Biometrica*: 62(2), 1975, p.269.

Crump, R.K., Hotz, V.J., Imbens, G.W. and Mitnik, O.A. (2009) "Dealing with limited overlap in estimation of average treatment effects " *Biometrika*: Oxford University Press, 96(1), pp. 187-199.

Cruz, L.M. and M.J. Moreira (2005) "On the Validity of Econometric Techniques with Weak Instruments: Inference on Returns to Education Using Compulsory School Attendance Laws," *Journal of Human Resources*, 40(2), 393-410.

Cutler D., J. Gruber (1995) "Does Public Insurance Crowd Out Private Insurance?" *National Bureau of Economic Research*, Working Paper: 5082, April 1995, pages 33

Davidson, R. and J.G. MacKinnon (1993): *Estimation and Inference in Econometrics*, Oxford University Press, 1993.

Deaton, A. (1985) "Panel data from time series of cross-sections," *Journal of Econometrics*: 30(1-2), pp. 109-126.

Deaton, A. (1997) *The analysis of household surveys: A microeconometric approach to development policy*, Baltimore and London: Johns Hopkins University Press for the World Bank, 1997, pg: 67-72.

Deaton, A. (2009) "Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development, " *NBER* WP No. 14690.

Dehejia, Rajeev H. and Sadek Wahba (1998) "Casual effects in non-experimental studies: re-evaluating the evaluation of training programs," *NBER* WP No. 6586, June 1998, 24 p.

Donald, Stephen G. and Kevin Lang (2007) "Inference with Difference in Differences and Other Panel Data," *The Review of Economics and Statistics*, May 2007, 89(2): 221–233.

Durlauf, S.N. (2002) "On the empirics of social capital," *Economic Journal*: 112(483), pp. F459-F479.

Eckstein Z. and K. Wolpin (1989) "The Specification and Estimation of Dynamic Stochastic Discrete Choice Models: A Survey", *Journal of Human Resources*, 24(4), Fall 1989, pages 562-98.

Elbers, Chris and Geert Ridder (1982) "True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Model," *The Review of Economic Studies*: 49(3), July 1982, pp. 403-409.

Engberg, J. (1990) "Structural Estimation of the Impact of Unemployment Benefits on Job Search," *University of Wisconsin*, Ph.D. 1990

Engberg J., S.L. Kim (1995) "Efficient Siplification of Bisimulation Formulas," in *TACAS*, pages 58-73.

Epple D., H. Sieg (1999) "Estimating Equilibrium Models of Local Jurisdictions," *Journal of Political Economy* 107(4), August 1999, pages 645-81.

Fan J., I. Gijbels (1996): *Local Polynominal Modelling and its Applications*, New York: Chapman & Hall, 1996.

Flores-Lagunes, A. (2007) "Finite sample evidence of IV estimators under weak instruments, " *Journal of Applied Econometrics*: 22(3), pp. 677-694.

Frölich, M. (2007), "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics*, 139, 35-75.

Galton, F. (1886) "Regression towards mediocrity in hereditary stature, " *Journal of the Anthropological Institute*: 15, pp. 246-263.

Geman, S., D. Geman (1984) "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," Polson,-Nicholas; Tiao,-George-C., eds. *Bayesian inference*. Volume 2. Elgar Reference Collection. International Library of Critical Writings in Econometrics, vol. 7.Aldershot, U.K.: Elgar; distributed in the U.S. by Ashgate, Brookfield, Vt., 1995, pages 149-69. Previously published: [1984].

Geweke, John, Michael Keane and David Runkle (1994) "Alternative Computational Approaches to Inference in the Multinomial Probit Model," *The Review of Economics and Statistics*: 76(4), Nov. 1994, pp. 609-632.

Godfrey, L.G. (1988) "Misspecification tests in econometrics: The Lagrange multiplier principle and other approaches," *Econometric Society Monographs series*, no. 16, Cambridge; New York and Melbourne: Cambridge University Press, 1988, pages xii, 252.

Greene, William H. (2005): *Econometric Analysis*, fifth edition, Prentice Hall.

Griliches Z. (1977) "Estimating the Returns to Schooling: Some Econometric Problems," *Econometrica* 45(1), Jan. 1977, pages 1-22.

Griliches Z., J. Hausman (1986) "Errors in Variables in Panel Data," *Journal of Econometrics* 31: pp. 93-118.

Haavelmo, T. (1994) "Statistical Testing of Business-Cycle Theories," Poirier, D.J., ed. *The methodology of econometrics.* Volume 1. Elgar Reference Collection. International Library of Critical Writings in Econometrics, vol. 6. Aldershot, U.K.: Elgar; distributed in the U.S. by Ashgate, Brookfield, Vt., 1994, pages 75-80. Previously published: [1943].

Hahn J. and J. Hausman (2002a) "Notes on bias in estimators for simultaneous equation models," *Economics Letters* 75, pp. 237–41.

Hahn J. and J. Hausman (2002b) "A New Specification Test for the Validity of Instrumental Variables," *Econometrica* 70 (1): 163-189.

Hall, Peter (1992) *The Bootstrap and Edgeworth Expansion*, New York, Springer, 1992.

Ham and LaLonde (1996) "The Effect of Sample Selection and Initial Conditions in Duration Models: Evidence from Experimental Data on Training", *Econometrica* 64(1), January 1996, pages 175-205.

Hansen, C.B. (2007) "Asymptotic properties of a robust variance matrix estimator for panel data when T is large," *Journal of Econometrics*: 141(2), December 2007, pp. 597-620.

Härdle, W., (1989): *Applied Nonparametric Regression*, Cambridge University Press, 1989.

Hausman, J.A. (1978) "Specification Tests in Econometrics," *Econometrica* 46: 1251-1272.

Hausman, J.A. *et al.* (1991) "Identification and Estimation of Polynomial Errors-in-Variables Models", *Journal of Econometrics*, 50(3), December 1991, pages 273-95.

Hausman, J.A., D. McFadden (1984) "Specification Tests for the Multinomial Logit Model", *Econometrica* 52(5), September 1984, pages 1219-40.

Heckman, J.J. (1979) "Sample Selection Bias as a Specification Error," *Econometrica* 47: 153-161.

Heckman, J. J. (1980) "Addendum to sample selection bias as a specification error, " E. Stromsdorfer, E. and Farkas, G. (Eds.) *Evaluation Studies Review Annua.* Volume 5. Beverly Hills: Sage Publications.

Heckman, J.J. (1990) "A Nonparametric Method of Moments Estimator for the Mixture of Geometrics Model," Hartog, J.; Ridder,G.; Theeuwes, J., eds. *Panel data and labor market studies*. Contributions to Economic Analysis, vol. 192, Amsterdam; Oxford and Tokyo: North-Holland; distributed in the U.S. and Canada by Elsevier Science, New York, 1990, pages 69-79.

Heckman, J.J. (2000) "Causal Parameters and Policy Analysis in Econometrics: A Twentieth Centurtury Perspective," *Quarterly Journal of Econometrics*, Feb. 2000.

Heckman, J.J., H. Ichimura, J. Smith, P. Todd (1995) "Nonparametric characterization of selection Bias Using Experimental Data: A Study of Adult Males in JTPA," Presented at the *Latin American Econometric Society Meeting*, Caracas, Venezuela, 1994.

Heckman, J., Ichimura, H., Smith, J. and Todd, P. (1998) "Characterizing selection bias using experimental data, " *Econometrica*: 66(5), pp. 1017-1098.

Heckman, J.J., H. Ichimura and P. Todd (1997) "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *The Review of Economic Studies*: 64(4), Special Issue: Evaluation of Training and Other Social Programmes, Oct. 1997, pp. 605-654.

Heckman J.J., B. Singer (1984) "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica* 52(2), March 1984, pages 271-320.

Heckman, J.J., Tobias, J.L. and Vytlacil, E. (2000) "Simple Estimators for Treatment Parameters in a Latent Variable Framework with an Application to Estimating the Returns to Schooling," *NBER* WP No. 7950, October 2000, 36 p.

Heckman,J.J. and Urzua, S. (2009) "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify, " *IZA*Discussion Paper no. 3980. IZA, 2009.

Heckman J.J., E. Vytlacil (1998) "Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return Is Correlated with Schooling", *Journal of Human Resources* 33(4), Fall 1998, pages 974-87.

Heckman J.J., E. Vytlacil (2000) "The Relationship between Treatment Parameters within a Latent Variable Framework," *Economics Letters* 66(1), January 2000, pages 33-39.

Heckman, J.J. and Vytlacil, E. (2005) "Structural equations, treatment effects and econometric policy evaluation, " *NBER* WP No. 11259.

Hellerstein, J.K., G.W. Imbens, (1999) "Imposing Moment Restrictions from Auxiliary Data by Weighting," *Review of Economics and Statistics* 81(1), February 1999, pages 1-14.

Holland P. (1986) "Statistics and Causal Inference", *Journal of the American Statistical Association* 81(396), December 1986, pages 945-60.

Honoré, B. (1992) "Trimmed Lad and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica*: 60(3), May 1992, pp. 533–565.

Honoré, B., S. Khan and J.L. Powell (2002) "Quantile regression under random censoring," *Journal of Econometrics*: 109(1), July 2002, pp. 67-105.

Honoré, B. and A. Lleras-Muney (2006) "Bounds in competing risks models and the war on cancer," *Econometrica*: 74(6), November 2006, pp. 1675-1698.

Horowitz J. (1992) "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60(3), May 1992, pages 505-31.

Hotz V., R. Miller (1993) "Conditional Choice Probabilities and the Estimation of Dynamic Models," *Review of Economic Studies* 60(3), July 1993, pages 497-529.

Hoxby, C. (2000) "Does competition among public schools benefit students and taxpayers?," *American Economic Review*: 90(5), pp. 1209-1238.

Hsiao, C. (1986) *Analysis of Panel Data*, Cambridge U. Press, 1986.

Hsiao, C. and M.H. Pesaran (2004) "Random Coefficient Panel Data Models," *IZA* Discussion Paper no. 1236. IZA, 2004.

Ichimura H. (1993) "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, 58(1-2), July 1993, pages 71-120.

Ichimura, H. and C. Taber (2000) "Direct Estimation of Policy Impacts," *NBER* Technical WP No. 254, 41 p.

Imbens, Guido W. (1992) "An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-Based Sampling," *Econometrica*, 60(5), September 1992, pages 1187-214.

Imbens, Guido W. (2004) "Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *The Review of Economics and Statistics*, February 2004, 86(1): pp. 4–29.

Imbens, Guido W. and J.K. Hellerstein (1993) "Raking and Regression," *Harvard Institute of Economic Research* Working Paper No. 1657, 15 p.

Imbens, Guido W. and Thomas Lemieux (2007) "Regression Discontinuity Designs: A Guide to Practice," *National Bureau of Economic Research* Working Paper: 13039, April 2007, pg. 37.

Imbens, Guido W. and Jeffrey M. Wooldridge (2007) "What's New in Econometrics," Unpublished Manuscript, National Bureau of Economic Research, Summer 2007.

Inoue, Atsushi and Gary Solon (2005) "Two-Sample Instrumental Variables Estimators," *NBER* Technical Working Paper 311, June 2005, 19 p.

Jakubson, G. (1991) "Estimation and Testing of the Union Wage Effect Using Panel Data," *Review of Economic Studies,* 58: 971-991.

Jurajda, S. (2002) "Estimating the Effect of Unemployment Insurance Compensation on the Labor Market Histories of Displaced Workers," *Journal of Econometrics,* June 2002, Vol. 108, No. 2.

Kelejian, Harry H. (1971) "Two-Stage Least Squares and Econometric Systems Linear in Parameters but Nonlinear in the Endogenous Variables," *Journal of the American Statistical Association*: 66(334), June 1971, pp. 373-374.

Khan S., S. Chen (2000) "Estimating Censored Regression Models in the Presence of Nonparametric Multiplicative Heteroskedasticity," *Journal of Econometrics* 98, 283-316.

Khan S., S. Chen (2001) "Semiparametric Estimation of a Partially Linear Censored Regression Model," *Econometric Theory* 17, pg. 567-590.

Khan and Powell (2001) "Two-step estimation of semiparametric censored regression models," *Journal of Econometrics*: 103(1-2), July 2001, pp. 73-110.

Kiefer N. (1988) "Economic Duration Data and Hazard Function," *Journal of Economic Literature* 26(2), June 1988, pg: 646-79.

Kiefer, D. (1988) "Interstate Wars in the Third World: A Markov Approach," *Conflict Management and Peace Science*; 10(1), Spring 1988, pages 20-36.

Klein R., R. Spady (1993) "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61(2), March 1993, pages 387-421.

Kling, J. (1999) "Interpreting IV Estimates of the Returns to Schooling," *Princeton University Industrial Relations Section WP* 415, 1999.

Kmenta, J. (1986): *Elements of Econometrics*, 2nd Ed., New York: Macmillan, 1986.

Koenker, R. (2004) "Quantile Regression for Longitudinal Data," *Journal of multivariate analysis*: 91(1), pp. 74-89.

Koenker, R., and G. Bassett (1982) "Robust Tests for Heteroscedasticity Based on Regression Quantiles," *Econometrica,* 50(1), Jan. 1982, pages 43-61.

Kyriazidou, Ekaterini (1997) "Estimation of a Panel Data Sample Selection Model," *Econometrica*: 65(6), Nov. 1997, pp. 1335-1364.

Lafontaine, F., J.K. Shaw (1995) "Firm-Specific Effects in Franchise Contracting: Sources and Implications", mimeo.

Lafontaine, F., J.K. Shaw (1999) "The Dynamics of Franchise Contracting: Evidence from Panel Data," *Journal of Political Economy,* 107(5), October 1999, pages 1041-80.

LaLonde, Robert J. (1986) "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *The American Economic Review*: 76(4), Sep. 1986, pp. 604-620.

Lancaster T., (1990): *The Econometric Analysis of Transition Data*, Cambridge U. Press, 1990.

Leamer, E. E. (1978) "Least-Squares versus Instrumental Variables Estimation in a Simple Errors in Variables Model," *Econometrica*, 46(4), July 1978, pages 961-68.

Lee, L.F. (1981) "Fully Recursive Probability Models and Multivariate Log-Linear Probability Models for the Analysis of Qualitative Data," *Journal of Econometrics*, 16(1), May 1981, pages 51-69.

Lee, L.F. (1983) "Generalized econometric models with selectivity, " *Econometrica*: 51, pp. 507-512.

Lee, Myoung-jae (1996) *Methods of Moments and Semiparametric Econometrics for Limited Dependent Variable Models.* Springer, 1996, 296 p.

Lee, D. and Lemieux, T. (2009) "Regression Discontinuity Designs in Economics," *NBER* Working Paper: 14723.

Lechner, M. (2001) "The empirical analysis of East Germany fertility after unification: An update, " *European Journal of Population*: 17, pp. 61-74.

Lewbel A. (2004) "Simple Estimators For Hard Problems: Endogeneity in Discrete Choice Related Models," *Boston College Working Papers in Economics* No. 604, 34 p.

Loeb, Susanna and John Bound (1996) "The Effect of Measured School Inputs on Academic Achievement: Evidence from the 1920s, 1930s and 1940s Birth Cohorts," *The Review of Economics and Statistics*: 78(4), Nov. 1996, pp. 653-664.

Machado, José A.F. and José Mata (2005) "Counterfactual decomposition of changes in wage distributions using quantile regression," *Journal of Applied Econometrics*: 20(4), pp. 445-465.

Maddala G.S., (1983): *Limited-dependent and Qualitative Variables in Econometrics*, Cambridge U. Press, 1983.

Manning, A. (2004) "Instrumental Variables for Binary Treatments with Heterogeneous Treatment Effects: A Simple Exposition," *Contributions to Economic Analysis & Policy*: 3(1), Article 9.

Manski, Charles F. (1975) "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3(3), Aug. 1975, pages 205-28.

Manski, Charles F. (1985) "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27(3), March 1985, pages 313-33.

Manski, Charles F. (1995) *Identification Problems in the Social Sciences*, Harvard University Press, 194 p.

Manski, Charles F. and Daniel McFadden (1981) *Structural Analysis of Discrete Data and Econometric Applications*, <elsa.berkley.edu/users/mcfadden/discrete.html> Cambridge: The MIT Press, 1981.

Marschak, J. (1953) "Economic Measurements for Policy and Prediction," Hood, W., Koopmans,T. and Hendry, D. F., eds. *Studies in Econometric Method.* Wiley, New York, 1953.

Marschak, J. (1995) "Economic Interdependence and Statistical Analysis," Hendry, D. F.; Morgan, M. S., eds. *The foundations of econometric analysis*. Cambridge; New York and Melbourne: Cambridge University Press, 1995, pages 427-39. Previously published: [1942].

Matsudaira, J.D. (2008) "Mandatory summer school and student achievement," *Journal of Econometrics*: 142(2), February 2008, pp. 829-850.

Matzkin, R. (1992) "Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and the Binary Choice Models," *Econometrica*, 60(2), March 1992, pages 239-70.

McCleary, R. and Barro, R. (2006) "Religion and Economy," *Journal of Economic Perspectives*: 2(2), Spring 2006, pp. 49-72.

McCulloch, R. E., P.E. Rossi (1994) "An Exact Likelihood Analysis of the Multinomial Probit Model," *Journal of Econometrics*; 64(1-2), Sept.-Oct. 1994, pages 207-40.

McFadden, Daniel (1984) "Econometric Analysis of Qualitative Response Models" in *Handbook of Econometrics*. Volume II, Griliches,Z. , Intriligator M. ed. Handbooks in Economics series, book 2. Amsterdam; New York and Oxford: North-Holland; distributed in the U.S. and Canada by Elsevier Science, New York, 1984.

McFadden, Daniel (1989) "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration", *Econometrica*, 57(5), September 1989, pages 995-1026.

McFadden, Daniel, and Kenneth Train (2000) "Mixed MNL models for discrete response," *Journal of Applied Econometrics*: 15(5), pp. 447-470.

Miguel,E., Satyanath, S. and Sergenti, E. (2004) "Economic Shocks and Civil Conflict: An Instrumental Variables Approach, " *Journal of Political Economy* : 112(41), pp. 725-753.

Miller, R. A. (1984) "Job Matching and Occupational Choice," *Journal of Political Economy*, 92(6), December 1984, pages 1086-120.

Moffitt (2007) "Estimating Marginal Returns to Higher Education in the UK," *NBER* WP no. 13534, October 2007, 41 p.

Moreira, Marcelo J. (2003) "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*: 71(4), July 2003, pp. 1027-1048.

Moreira, Marcelo J. and Brian P. Poi (2003) "Implementing Conditional Tests with Correct Size in the Simultaneous Equations Model," *Stata Journal*: 1(1), pp. 1-15.

Münich, Daniel, Jan Svejnar and Katherine Terrell (2005) "Returns to Human Capital under the Communist Wage Grid and during the Transition to a Market Economy," *The Review of Economics and Statistics*, February 2005, 87(1): 100–123.

Murray, Michael P. (2006) "Avoiding Invalid Instruments and Coping with Weak Instruments" *Journal of Economic Perspectives*: 20(4), Fall 2006, pp. 111-132.

Nevo, A. and Rosen, A. (2008) "Identification with Imperfect Instruments," *NBER* Working Paper: 14434.

Newey, W. (1985) "Generalized Method of Moments Specification Tests," *Journal of Econometrics,* 29: 229-238.

Newey, W. and J. Powell (1990) "Efficient Estimation of Linear and Type I Censored Regression Models under Conditional Quantile Restrictions," *Econometric Theory*: 6(3), Sep. 1990, pp. 295–317.

Olsen, R. J. (1987) "Note on the Uniqueness of the Maximum Likelihood Estimator for the Tobit Model," *Econometrica*, 46(5), Sept. 1978, pages 1211-15.

Pagan, A., F. Vella (1989) "Diagnostic Tests for Models Based on Individual Data: A Survey," *Journal of Applied Econometrics,* 4(0), Supplement, December 1989, pages S29-59.

Pakes, A.S. (1986) "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," *Econometrica,* 54(4), July 1986, pages 755-84.

Pakes, A.S. and D. Pollard (1989) "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57(5), pp. 1027-1057.

Pesaran, M.H., Y. Shin, R.J. Smith (2000) "Structural Analysis of Vector Error Correction Models with Exogenous I(1) Variables," *Journal of Econometrics,* 97(2), August 2000, pages 293-343..

Popper, K. (1959) *The logic of scienrifc discovery.* London: Hutchinson.

Powell, J.L. (1983) "The Asymptotic Normality of Two-Stage Least Absolute Deviations Estimators," Econometrica: 51(5), Sep. 1983, pp. 1569-1575.

Powell, J.L. (1984) "Least Absolute Deviation Estimation for the Censored Regression Model," *Journal of Econometrics*, 25(3), July 1984, pages 303-25.

Powell, J.L. (1986) "Symmetrically Trimmed Least Squares Estimation for Tobit Models," *Econometrica*, 54(6), November 1986.

Pudney, S. (1989) *Modelling Individual Choice*, Basil Blackwell, 1989.

Rivers, Douglas and Quang H. Vuong (1988) "Limited information estimators and exogeneity tests for simultaneous probit models" *Journal of Econometrics*: 39, November 1988, pp. 347-366.

Robinson, P.M. (1988) "Using Gaussian Estimators Robustly," *Oxford Bulletin of Economics and Statistics*; 50(1), February 1988, pages 97-106.

Rosenbaum, P.R. (2002) "Covariance Adjustment in Randomized Experiments and Observational Studies," *Statistical Science*: 17(3), 2002, pp. 286–327.

Rosenbaum, P.R. and D.B. Rubin (1983) "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society. Series B (Methodological)*: 45(2), 1983, pp. 212-218.

Rosenbaum P.R. and D.B Rubin (1984) "Estimating the Effects Caused by Treatments: Comment [On the Nature and Discovery of Structure]", *Journal of the American Statistical Association*, 79(385), March 1984, pages 26-28.

Roy P.N. (1970) "On the Problem of Measurement of Capital," *Economics Affairs*, 15(6-8 156), Aug. 1970, pages 269-76.

Rust, J. (1987) "A Dynamic Programming Model of Retirement Behavior," *National Bureau of Economic Research* Working Paper: 2470, December 1987, pages 64.

Rust, J. (1992) "Do People Behave According to Bellman's Principle of Optimality?," *Hoover Institute Working Papers in Economics*: E-92-10, May 1992, pages 76.

Rust J. (1994) "Structural Estimation of Markov Decision Processes," Engle,-Robert-F.; McFadden,-Daniel-L., eds. *Handbook of econometrics.* Volume 4. Handbooks in Economics, vol. 2. Amsterdam; London and New York: Elsevier, North-Holland, 1994, pages 3081-3143.

Schennach, S (2004) “Nonparametric regression in the presence of measurement error, ” *Econometric Theory*: 20(6), December, 2004, pp. 1046-1093.

Scott, A.J. and Wild, C.J. (1997) “Fitting regression models to case-control data by maximum likelihood, ” *Biometrika*: 84 (1), pp. 57-71.

Sianesi, B. (2001) “Differential effects of Swedish active labour market programs for unemployed adults during the 1990s,” *IFS* Working Papers W01/25, Institute for Fiscal Studies, 2001.

Silverman (1986): *Density Estimation for Statistics and Data Analysis*, London, Chapman & Hall, 1986.

Staiger, Douglas and James H. Stock (1997) “Instrumental Variables Regression with Weak Instruments,” Econometrica: 65(3), May 1997, pp. 557-586.

Stock, J. H., and Yogo, M. (2005) “Testing for Weak Instruments in Linear IV Regression ” in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. D.W. Andrews, D.W. ed. and Stock, J.H. Cambridge University Press. 2005.

Taber, Chr.R. (2000) “Semiparametric Identification and Heterogeneity in Discrete Choice Dynamic Programming Models,” *Journal of Econometrics,* 96(2), June 2000, pages 201-29.

Tanner, M. A (1993) *Tools for Statistical Inference*, 2nd. edition. New York: Springer-Verlag.

Tibshirani, Robert and Trevor Hastie (1987) “Local Likelihood Estimation,” *Journal of the American Statistical Association*: 82(398), June 1987, pp. 559-567.

Todd, P. E., and Wolpin, K.I. (2006) “Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility, ” *American Economic Review*: 96(5), pp. 1384-1417.

Urquiola, M. and Verhoogen, E. (2007).“Class size and sorting in market equilibrium: theory and evidence, ” *NBER* Working Papers: 13303.

Van der Klaauw (2008) “Regression-Discontinuity Analysis: A Survey of Recent Developments in Economics” (forthcoming).

White, H. (1980) “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity,” *Econometrica,* 48(4), May 1980, pages 817-38.

Willis R. J., S. Rosen (1979) "Education and Self-Selection," *Journal of Political Economy*, 87(5), Part 2, Oct. 1979, pages S7-36.

Wooldridge, Jeffrey M. (2002) *Econometric Analysis of Cross Section and Panel Data*, MIT Press, 2002.

Wooldridge, Jeffrey M. (2003) "Cluster-Sample Methods in Applied Econometrics," *The American Economic Review*, Vol. 93, No. 2, Papers and Proceedings of the One Hundred Fifteenth Annual Meeting of the American Economic Association, Washington, DC, January 3-5, 2003 (May, 2003), pp. 133-138

Wolpin, K. I. (1984) "An Estimable Dynamic Stochastic Model of Fertility and Child Mortality," *Journal of Political Economy*, 92(5), October 1984, pages 852-74.

Yatchew A., (1998) "Nonparametric Regression Techniques," *Journal of Economic Literature* 1998.

Zanella, Giulio (2007) "Discrete Choice with Social Interactions and Endogenous Memberships," *Journal of the European Economic Association*, March 2007, 5(1): 122–153.

Zellner, Arnold (1996) *An Introduction to Bayesian Inference in Econometrics.* Wiley-Interscience, 448 p.

Ziliak, S.T. and McCloskey, D.N.(2008) *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor, MI: University of Michigan Press, 320 p.