CERGE-EI

Charles University
Center for Economic Research and Graduate Education
Academy of Sciences of the Czech Republic
Economics Institute

Lecture Notes Series

# ECONOMETRICS OF PANEL DATA AND LIMITED DEPENDENT VARIABLE MODELS

Štěpán JURAJDA

April 2003

CERGE-EI LECTURE NOTES 3

P.O. Box 882, Politických vězňů 7, 111 21 Praha 1, Czech Republic
http://www.cerge-ei.cz

# Econometrics IV

Štěpán Jurajda

May 9, 2003

## Contents

## Preamble

These lecture notes were written for a 2nd-year Ph.D. course in econometrics of panel data and limited-dependent-variable-models. The primary 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 cross-section and panel data analysis, (ii) to focus on situations where linear regression models are not appropriate and to study alternative methods. 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 textbooks** for the course are:

1. *Econometric Analysis of Cross Section and Panel Data*, [W], Jeffrey M. Wooldridge, MIT Press 2002.
2. *Econometric Analysis*, [G], William H. Green.
3. *Analysis of Panel Data*, [H], Cheng Hsiao, Cambridge U. Press, 1986.
4. *Limited-dependent and Qualitative Variables in Econometrics*, [M], G.S. Maddala, Cambridge U. Press, 1983.

Other useful references are:

1. *Advanced Econometrics*, [A], Takeshi Amemiya, Harvard U. Press, 1985

2. Gary Chamberlain (1984) “Panel Data”, [C] in *Handbook of Econometrics vol. II*, pp. 1247-1318. Amsterdam North- Holland.

3. *Modelling Individual Choice,* [P], S. Pudney, Basil Blackwell, 1989.

4. *The Econometric Analysis of Transition Data*, [L], Tony Lancaster, Cambridge U. Press, 1990.

5. *Estimation and inference in econometrics* [DM] Davidson, R., and J.G. MacKinnon, Oxford University Press, 1993.

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

7. *Applied Nonparametric Regression*, [N] Wolfgang Härdle, Cambridge U. Press, 1989.

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

Below find a simplified course outline including *selected* readings:

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. Review of basic linear regression model and Introduction to Maximum Likelihood Estimation and Hypothesis testing ([G])

3. Cases where residuals are correlated ([G] 14 ,[A] 6)

    - GLS

      White, H. (1980) “A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity, ” *Econometrica* 48:817-838.

    - Panel data analysis ([H] 3.3, 6)

4. Cases where residuals and regressors are correlated ([H] 6-7, [A] 7-8)

- 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.

  Jacubson (1991) "Estimation and Testing of the Union Wage Effect Using Panel Data," *Review of Economic Studies* 58:971-991.

- Misspecification ([H] 3.4, 3.5, 3.8, [C])

  Hausman, J. (1978) "Specification Tests in Econometrics," *Econometrica* 46:1251-1272

  Newey, W. (1985) "Generalized Method of Moments Specification Tests," *Journal of Econometrics* 29:229-238.

- 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 ([G] 20)

5. Cases where linear regression models are not appropriate

- Maximum Likelihood Estimation ([A] 3-4)
- Qualitative response models ([M] 2-3, [A] 9, [H] 7, [G] 21)
- Tobit models ([A] 10, H [6], [G] 22)

  Amemiya T. (1984) "Tobit Models: A Survey," *Journal of Econometrics* 24(1-2).

- Self selection models ([M] 9)

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

- Duration analysis ([L], [G] 22)

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

6. Introduction to nonparametric methods

- Kernel estimation and Local Linear Regression

*Density Estimation for Statistics and Data Analysis* (1986), B.W. Silverman, Chapman and Hall.

*Local Polynomial Modelling and its Applications* (1996), J. Fan and I. Gijbels, Chapman and Hall.

- Discrete choice models

  Matzkin R. (1992) “Nonparametric and Distribution Free Estimation of Threshold crossing and Binary Choice Models,” *Econometrica* 60(2).

- Selection bias

  Heckman J., H. Ichimura, J. Smith and P. Todd (1995) “Nonparametric Characterization of Selection Bias Using Experimental Data: A Study of Adult Males in JTPA”

- Trimmed LS and Censored LAD Estimators

  Powell, J.L. (1984) “Least Absolute Deviation Estimation for the Censored Regression Model,” *Journal of Econometrics* 25(3).

  Powell, J.L. (1986) “Symmetrically trimmed Least Squares Estimation for Tobit Models,” *Econometrica* 54(6).

# Part I
# Introduction

## 1. Causal Parameters and Policy Analysis in Econometrics

Econometrics[1] 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.

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.

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, nonparametrics (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 incredible 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 descrip-

[1] This introductory class is based on a recent survey by J.J. Heckman (2000).

tion 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.

- 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 to identify causal effects.

  (+) transparency: credible identification.

  (−) theory remains only at an intuitive level; causal parameters are relative to IV (LATE[2]); it is hard to cumulate knowledge and the estimates to not render counterfactual policy predictions.

[2] See Section 11.

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).

Finally, note that today, the Cowless commission paradigm (Haavelmo, 1944; Popper, 1959) is partly abandoned in favor of more interaction with data (learning) so that it is merely used as a reporting style (Leamer, 1978).

In this course, we will mostly remain within the classical paradigm and discuss parametric reduced-form econometric models. We will also occasionaly touch on non-parametric and natural-experiment research and return to discussing causal inference when introducing the program evaluation literature in Section 11.

## 2. Reminder

This section aims at reminding ourselves with some of the basic econometric issues. See also [W]1. First, in subsection 2.1, we make the link between a linear regression and the true regression function $E[y \mid x]$. Second, we survey the main principles of hypothesis testing. Finally, we remind ourselves about the sensitivity of extremum estimators to distributional assumptions and preview the issues important in cross-sectional data: measurement error, sampling (endogenous sampling and consistency, multiple-stage sampling and inference), combining data sets, etc.

### 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. 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 2.1.** *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 $X_1$.*

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

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

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

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 $X_2 \mid X_1 \sim N(\mu_{2.1} + \Delta_{2.1}X_1, \Sigma_{22.1})$. $\blacksquare$

**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).*

### 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 miss-specification.

- Specification tests: preview of Hansen and Hausman.

- Sequential testing, data mining: 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, 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.

- Non-nested testing.

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).

**Exercise 2.2.** *Prove or provide a counterexample for the following statements:*

- $Y \perp X \Longleftrightarrow COV(X,Y) = 0$. See also Exercise 2.1.
- $E[X \mid Y] = 0 \Longleftrightarrow E[XY] = 0 \Longleftrightarrow COV(X,Y) = 0$
- $E[X \mid Y] = 0 \Longrightarrow E[Xg(Y)] = 0 \; \forall g(\cdot)$. Is $COV(X \mid Y) = 0$ ?
- $E[Y] = E_X[E_Y(Y \mid X)]$ and $V[Y] = \underbrace{E_X[V_Y(Y \mid X)]}_{\text{residual variation}} + \underbrace{V_X[E(Y \mid X)]}_{\text{explained variation}}$.

## 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 variance-covariance matrix is not full of 1s: (a) e.g., linear prediction vs. $E[y \mid x]$ or heteroscedasticity,[3] (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[4], converges to the true matrix with all of the $V[\varepsilon|x]$ so that

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

[3] 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 $V(y_r) = \frac{1}{N_r}V(y_{ir})$.

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

(Here we also preview the Hausman test by comparing the OLS and Huber/White variance-covariance matrix. See [G]14.6, [W]4.2.3.

**(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 $\beta$ (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 $\beta$, (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 $\epsilon$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

$$\begin{aligned} y_{it} &= x_{it}'\beta_{it} + \epsilon_{it} \text{ with } i = 1, ..., N \text{ and } t = 1, ..., T \text{ , or} \qquad (4.1) \\ \underset{T\times 1}{y_i} &= \underset{T\times k}{X_i}\underset{k\times 1}{\beta_{it}} + \epsilon_i \text{ with } i = 1, ..., N \text{ or} \\ \underset{NT\times 1}{y} &= \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix} \beta_{it} + \epsilon \text{ ,} \end{aligned}$$

where the covariance structure of $\epsilon_{it}$ 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 2.** *$N$ 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 3.** *The number of time periods $T$ may differ for each person. This is often referred to as unbalanced panel.*

**Remark 4.** *$T$ is usually smaller than $N$ and most asymptotic results rely on $N \rightarrow \infty$ with $T$ fixed.*

The first question is whether we constrain $\beta$ to be the same across either dimension. We cannot estimate $\beta_{it}$ 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 $\beta$s 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 $\beta_i$ 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 $\beta_i$s. We have $V[\varepsilon] = E[\varepsilon\varepsilon'] = \Sigma \otimes I_T \neq \sigma^2 I_{NT}$, i.e. the $\epsilon$ '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. See [G].

### 4.2. Random Coefficients Model

What if we still want to allow *parameter* flexibility across cross-sectional units, but some of the $\beta_i$s are very uninformative. Then one solution may be to combine the estimate of $\beta_i$ from each time series regression 4.2 with the 'composite' estimate of $\beta$ from the pooled data in order to improve upon an imprecise $\widehat{\beta}_i$ using information from other equations.[5] In constructing $\beta$, each $\beta_i$ 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 $V[\widehat{\beta}_i] = \sigma_i^2 (X_i' X_i)^{-1} + \Gamma = V_i + \Gamma$

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

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

**Remark 5.** *$V_i$ tells us how much variance around $\beta$ 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

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

$\Gamma$ can be estimated from the sample variance in $\widehat{\beta}_i$ 's ([G] p460). 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 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.*[6] *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. If the parameter vectors differ for each sampling strata $s = 1, ..., S$ so that $\beta_s \neq \beta$, a 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 $\beta_s$.*[7]

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

[7] The WLS estimator is consistent for $\beta$ 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.

**Remark 7.** *As usual we need asymptotics to analyze the behavior of $\widehat{\beta}$ since weights are nonlinear.*

**Remark 8.** $\widehat{\Gamma}$ *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[8] $A_i = (\Gamma^{-1} + V_i^{-1})^{-1}\Gamma^{-1}$.

**Remark 9.** *If $E[\nu_i] = f(X_i) \Longrightarrow E[\nu_i \mid X_i] \neq 0 \Longrightarrow \widehat{\beta}_i$ is not consistent for $\beta_i$.*

### 4.3. Random Effects Model

Assuming $\beta_{it} = \beta\ \forall i, t$ in Equation 4.1 one can impose a covariance structure on $\epsilon$ 's 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 $E[\alpha_i\alpha_j] = \sigma^2_\alpha$ if $i = j$ and is 0 otherwise. Other than that the only covariance is between $u_{it}$ and $u_{it}$ which is $\sigma^2_u$. We could also add a time random effect $\lambda_t$ to $\epsilon_{it}$.

Given this structure $V \equiv V(\underset{T\times 1}{\epsilon_i}) = \sigma^2_u I_T + \sigma^2_\alpha e_T e_T^{'}$, 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 the fixed effect model to describe this within estimator. Then we return to the above GLS formula, reparametrize $V^{-1}$ 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 $\alpha_i$ as random (and uncorrelated with $x$) provides us with an intermediate solution between treating $\alpha_i$ as being the same ($\sigma^2_\alpha = 0$) and as being different ($\sigma^2_\alpha \to \infty$). We combine both sources of variance: (i) over time within $i$ units and (ii) over cross-sectional units.

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

**Remark 10.** *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 11.** *However, with panel data one does not have to impose so much structure as in REM: (i) can estimate the person specific covariance using $\widehat{\epsilon}_{it}^{OLS}$ , $t = 1, ..., T$ (we will come to this later in one empirical example, see example 8.6), (ii) we can use minimum distance methods and leave the structure of error terms very flexible (see section 6.2.2).*

# 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 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 $\alpha_i$ 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 $\alpha$ 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 $\beta$ appropriately into $\beta^w$ and $\beta^z$. In this case note that we cannot separately identify $\beta^z$ from $\alpha_i$. 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*

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

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

**Remark 14.** *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 15.** *While effects of time-constant variables are not identifies in fixed effects models, one can estimate the change in the effect of these variables. Angrist (1995) AER.*

**Remark 16.** *Bertrand et al. (2001) suggest that a fixed effect estimation using state-time changes in laws etc. such as*

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

*may have the 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 oposed to 5%. 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. If not, they suggest the use of randomization inference tests: Use the distribution of estimated placebo laws to form the test statistic. However, recently Kézdi (2002) suggests that using option cluster() in Stata is fine.*[9]

## 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' $\sigma_x^2$ to 'noise' $\sigma_\nu^2$. 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 17.** *In the case of miss-clasification of a binary variables $E[\nu\varepsilon] = 0$ 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.*

**Remark 18.** *Within estimators (differencing) will typically make the measurement error bias worse. The signal-to-noise ratio will depend on $\sigma_x^2$ and on $\sigma_x^2 +$*

[9] http://www.econ.lsa.umich.edu/~kezdi/FE-RobustSE-2002-feb.pdf

*$\sigma_\nu^2[(1-\tau)/(1-\rho)]$ where $\tau$ is the first-order serial correlation in the measurement error and $\rho$ 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 $\beta$.*

**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.*

**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.*

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.

## 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 $\alpha_i$ (see [H] 3).*

### 6.1. Hausman test

- Basic idea is to compare two estimators[10]: 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 $\widehat{q}^{'} V_q^{-1} \widehat{q} \longrightarrow \chi^2(k)$.*

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.** *Hausman 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.*

[10] But it is not really LR test as the two hypotheses are non-nested.

**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 $\beta$s 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. 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 $\alpha_i$s. When $T$ is small we can test the validity of those restrictions using the MD methods. The same technique allows for estimation of $\beta$ with a minimal structure imposed on $\alpha$, allowing for correlation between the unobservable $\alpha$ and the regressors $x$. We will first understand the MD method and then apply it to panel data problems.

#### 6.2.1. The Minimum Distance Method

Suppose we have a model which implies restrictions on parameters which 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 $\theta$. Next, maintain the assumption that at the true value of $\theta$ 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:[11]

$$\widehat{\theta}_N = \arg\min\{S_N\} \text{ where } S_N = N[\widehat{\pi}_N - f(\theta)]' A_N[\widehat{\pi}_N - f(\theta)], \tag{6.1}$$

and where $A_N \xrightarrow{p} A$ is a weighting matrix and $\sqrt{N}[\widehat{\pi}_N - f(\theta)] \xrightarrow{D} N(0,\Delta)$.[12]

**Remark 26.** *The minimization problem 6.1 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 27.** *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 $\Gamma$ and $B$.*

**Remark 28.** *MD is efficient only among the class of estimators which do not impose apriori restrictions on the error structure.*

**Remark 29.** *The MD method can be used to pool two data sets to create an IV estimator (Arellano and Meghir 1991) if instruments are in both data sets, while one of the data sets includes the dependent variable and the other includes the explanatory variable of interest.*

[11] 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.

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

### 6.2.2. 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 $\alpha_i$ and time constant $x$ variables.

However, when $T$ is fixed and $N \longrightarrow \infty$,[13] one can allow $\alpha$ to have a general expectations structure given $x$ and estimate this structure together with our main parameter of interest: $\beta$ (Chamberlain 1982, [H] 3.8). That is we will not eliminate $\alpha_i$ (and its correlation with $x$) by first differencing. Instead, we will control for (absorb) the correlation between $\alpha$ 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 $\beta$ with as little structure on the omitted latent random variable $\alpha$ as possible.[14] The technique of estimation will be the MD method.

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

$$y_i = e_T \alpha_i^* + \underset{T \times K}{X_i} \beta + \varepsilon_i \tag{6.2}$$

and let $x_i = \text{vec}(X_i^{'})$.[15] To allow for possible correlation between $\alpha_i$ and $X_i$ , assume $E[\alpha_i^* \mid X_i] = \mu + \lambda^{'} x_i = \sum_{t=1}^{T} \lambda_t^{'} x_{it}$ (note $\mu$ and $\lambda$ do not vary over $i$) and plug back into 6.2 to obtain

$$y_i = e_T \mu + (I_T \otimes \beta^{'} + e_T \lambda^{'}) x_i + [y_i - E[y_i \mid x_i]] = e_T \mu + \underset{T \times KT}{\Pi} x_i + \upsilon_i \tag{6.3}$$

We can obtain $\widehat{\Pi}$ by gigantic OLS and impose the restrictions on $\Pi$ using MD.[16] We only need to assume $x_{it}$ are *iid* for $t = 1, \ldots, T$. 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 $\upsilon_i$ 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.*

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

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

[15] 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.

[16] How many underlying parameters are there in $\Pi$? Only $K + KT$.

**Remark 30.** *Hsiao's formulae (3.8.9.) and (3.8.10.) do not follow the treatment in (3.8.8.), but use time varying intercepts.*

### 6.2.3. 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 $\alpha_i$ is potentially correlated with $x_i$[17]. 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:[18]

  - Is differencing valid? Substitute for $\alpha_i$ 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 30), test $(\frac{\gamma_t}{\gamma_{t-1}}) = 1$ (does it make sense to use $\Delta y_{it}$ on the left-hand side?), if valid test $\beta_t = \beta \ \forall t$.

  - Is the effect "symmetric"?

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

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

[18] 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.

  - Does the effect vary with other $X$s?

**Remark 31.** *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

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

Solution: IV/GMM to find variation in the $X$ with simultaneity bias which is not related to the variation in the $\epsilon$s, i.e., use $\widehat{X}$ instead. Theory or intuition is often used to find an "exclusion restriction" postulating that a certain variable (a potential instrument) does not belong to the equation in question. We can also 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.** *To illustrate this, consider the demand and supply system from Econometrics I:*

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

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

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

*and note that one can identify $\beta_1$ 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 $\widehat{\beta_1} = \frac{\widehat{\pi_1}}{\widehat{\pi_2}}$; this is called indirect least squares and demasks IV. To identify $\alpha_1$ estimate $\Omega$ , 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 $\alpha_1$ as a function of $\beta_1$.*

A valid instrument $Z$ must be correlated with the endogenous part of $X$ (in the first-stage regression controlling for all exogenous explanatory variables!) and not correlated with $\varepsilon$.

**Remark 32.** *For testing the validity of exclusion restrictions (overidentification), that is testing* $COV(Z,\varepsilon)=0$, *see remark 27.*

**Example 7.2.** *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 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.*

**Example 7.3.** *Aside from econometric tests for IV validity (overidentification), one can also conduct intuitive tests when the exogenous variation (IV) comes from some quasi-experiment. For example, one can ask whether there is an association between the instrument and outcomes in samples where there should be none. For example Angrist in the Vietnam draft paper asks if earning vary with draft-eligibility status for the 1953 cohort, which had a lottery, but was never drafted.*

**Remark 33.** *Other then testing for* $COV(Z,\varepsilon)=0$, *one should also consider the* weak instrument *problem (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* $\widehat{\beta}$ *if* $COV(X,Z)$ *is small because* $p\lim\widehat{\beta}=\beta_0+COV(Z,\varepsilon)/COV(X,Z)$. *See [W]5.2.6.*

*IV is an asymptotic estimator, unlike OLS which is unbiased in small samples.*[19] *IV needs large samples to invoke consistency. Finite sample bias is larger when there are more instruments, samples are smaller, and instruments are weaker. Bound et al. (1995) suggest the use of F tests in the first stage. Also see Staiger and Stock (1997) who suggest that an F statistic below 5 suggests weak instruments. Alternatively, use LIML which is median-unbiased. Or use exactly identified models.*

**Exercise 7.1.** *Consider an endogenous dummy variable problem. Do you put in the predicted outcome or probability?*

## 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.

[19] IV is consistent but not unbiased because it features a ratio of two random variables.

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 $\chi^2$ 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:

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

To reach $\chi^2$ 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 6.2.1 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 34.** *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 $\widehat{\beta}_{GMM} = \widehat{\beta}_{TSLS}$.*

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

**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 $\varepsilon$ 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.** *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.*

**Remark 35.** *GMM does not require strong distributional assumptions on $\varepsilon$ like MLE. Further, when $\varepsilon$s are not independent, the MLE will not piece out nicely, but GMM will still provide consistent estimates.*

**Remark 36.** *GMM is consistent, but biased in general. It is a* large *sample estimator. In small samples it is often biased downwards (Altonji and Segal 1994).*

**Remark 37.** *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. See section 5. of the GMM handout by George Jakubson in the library.*

**Example 8.6.** *The GMM analogue to TSLS 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 conditional 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.*

**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 6.2.2.*

**Example 8.7.** *Nonlinear system of simultaneous equations. Euler equations.*

McFadden (1989) and Pakes (1989) allow the moments to be simulated: SMM (see Remark 44). Imbens and Hellerstein (1993) propose a method to utilize exact knowledge of some *population* moments while estimating $\theta$ from the *sample* moments: reweight the data so that the transformed sample moments would equal the population moments.

# Part III
# Qualitative and Limited Dependent Variables

## 9. Qualitative response models

Reading assignment: [M] 1,2, [G] 21.

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 $\varepsilon$ 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.[20] 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![21] Since MLE crucially hinges on distributional assumptions, recent literature focuses on estimation methods not requiring specification of any parametric distribution.

### 9.1. Binary Choice Models

### 9.1.1. 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

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

[21] For example, Arabmazar and Schmidt (1981), give some examples of the asymptotic biases for the Tobit model, see section 10.1.

probability the event occurs breaks down when $\widehat{\beta}^{'} x_i \notin [0,1]$.

**Exercise 9.1.** *Show that given $E[\varepsilon_i] = 0$, the residuals $\varepsilon_i$ which can take on only two values are heteroscedastic.*

The advantage of the LPM is in its ability to handle IV estimation easily. Applying the LPM is valid in large samples when the empirical $\widehat{y_i}$s are not close to 0 or 1. One should also allow the $x$s to enter as finite order polynomials, allowing for non-linearity.

**Example 9.1.** *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.*

### 9.1.2. 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

$$P[y_i = 1 \mid x_i] = P[u_i > -\beta^{'} x_i] = 1 - F(-\beta^{'} x_i) = F(\beta^{'} x_i). \tag{9.1}$$

Two common choices 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.

**Remark 38.** *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. $\theta$. 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 39.** *Both models are suitable for non-linear optimization using the Newton-Raphson methods as the Hessian is always n.d.*

**Remark 40.** $\widehat{\beta}'$*s from logit and probit are not directly comparable* ($\widehat{\beta}_{Logit} \simeq 1.6\widehat{\beta}_{\Pr obit}$ *, 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.* $\widehat{\beta}_{Logit} \neq \frac{\partial P[y_i=1|x_i]}{\partial x_i} = \Lambda(-\beta' x_i)[1 - \Lambda(-\beta' x_i)]\widehat{\beta}_{Logit}$.

**Remark 41.** *Parametric methods (e.g. probit and logit) force strict monotonicity and homoscedasticity.*[22]

**Remark 42.** *There are bivariate extensions in the SURE spirit ([G] 21.6). 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 9.2.** *Show that in Probit, one can only estimate $\beta/\sigma$.*

**Exercise 9.3.** *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 as a function of $x'\beta$.*

**Remark 43.** *See Davidson and MacKinnon (1993) textbook, chapter 15.4 for a useful auxiliary regression connected to qualitative response models.*

[22] There is a heterogeneity test for probit ([G]p.649).

### 9.1.3. The WLS-MD for Multiple Observations

([M] 2.8, [G] 21.4.3) 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 9.2.** *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 9.4.** *Show the WLS is a genuine MD (see HW#4).*

### 9.1.4. 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 $\beta'$s, we need to integrate the random effects $\alpha$ out. For MLE we must assume a particular distribution for $\alpha$, say $g(\alpha)$ depending on parameters $\delta$. Then allowing for correlation of $\alpha$ over time for the same person we can maximize the following with respect to both $\beta$ and $\delta$ :

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \int \prod_{t=1}^{T} \Phi(\beta' x_{it} + \alpha)^{y_{it}} \left[1 - \Phi(\beta' x_{it} + \alpha)\right]^{1-y_{it}} dG(\alpha|\delta) d\alpha \tag{9.2}$$

Notice the multidimensional integral (each $\Phi$ is an integral inside the $T$-dimensional integral over $\alpha$). We can simplify matters by assuming that the correlation of $\alpha$ 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 44.** *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 9.3):*

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

*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 $\varepsilon_i$. 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 45.** *To allow (flexible) correlation between $x_i$ and $\alpha_i$ we may follow Chamberlain (1980), but we now need the true regression function (see section 6.2.2) and a distributional assumption on the $\alpha$ equation error term.*

**Remark 46.** *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.*

**Conditional Fixed Effect Logit** The motivation for a fixed effect model is similar as in panel data linear regression. In MLE the $\alpha_i$s are again consistent only with $T \longrightarrow \infty$. Since $T$ is usually fixed and since MLE relies on consistency, the $\alpha_i$s 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 9.2.1) of $y_i$ given $\sum_{t=1}^{T} y_{it}$ which turns out to remove the heterogeneity. Conditional on $\alpha_i$s we have independence over both $i$ and $t$.

**Exercise 9.5.** *To verify this, write down the conditional likelihood contribution of $y_i' = (0,1)$ when $T = 2$.*

**Remark 47.** *Again, use Hausman test to compare the fixed effect model with the $\alpha_i = \alpha$ simple pooled-data logit.*

**Remark 48.** *The conditional fixed effect logit is computationally cumbersome for $T \geq 10$.*

**Exercise 9.6.** *Explain why $y_i$s with no change over time are not used in the estimation and show that observations with time constant $x$s are not used either.*

### 9.1.5. Choice-based sampling

[A]9.5[23] To analyze a rare event when population probabilities $p(y_i) = P[y_i = 1]$ are tiny (training treatment, violent crime), sample randomly within each group to obtain $f(x_i \mid y_i)$. Then 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

$$L(\cdot) = \prod_{i \in S_1} f(x_i \mid y_i = 1) \prod_{i \in S_2} f(x_i \mid y_i = 0) \tag{9.3}$$

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

**Remark 49.** $P[y_i = 1], P[y_i = 0]$ *usually come from a different data set (are known), but can be estimated as part of the problem.*

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:

$$L(\theta) = \sum_{i=1}^{N} \ln \frac{p(y = y_i | x_i, \theta) H_{y_i} / p(y = y_i)}{\sum_{j=1}^{M} p(y_i = j | x_i, \theta) H_j / p(y_i = j))}, \tag{9.4}$$

where $H_j$ is the probability of sampling from a strata $j$ (can be unknown to the researcher). 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 $\theta$ better. But then we do not get rid of $f(x)$, which was the beauty of 9.4.

[23] Se also Pudney (1989), chapter 3.2.3.

**Remark 50.** *Replacing $H$ with $\widehat{H}$ actually improves efficiency. Counterintuitive. Only works with an inefficient estimator.*

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 9.3.** *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 $\delta$ 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 $\delta$ does not have a discrete distribution.*

### 9.1.6. Relaxing the distributional assumptions of binary choice models

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 $\beta$ 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 $\varepsilon$ 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 $s = 0$.[24] Then $\beta_0^{'} x_i \geq (\leq) 0$ iff $E[y_i \mid x_i] \geq (\leq).5$ and we use sgn$(\beta^{'} x_i)$ −sgn$(E[y_i \mid x_i] - .5) = 0$ as a moment condition.[25] Then

$$\widehat{\beta}_{MRE} = \underset{s.t.\beta^{'}\beta=1}{\arg\max} \frac{1}{n} \sum_{i=1}^{n} \Big[ (2y_i - 1) sgn(\beta^{'} x_i) \Big] \tag{9.5}$$

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 (*1*992) 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 idea of comparing pairs of people.[26] Assume the model $y_i = d_i = 1\{x_i\beta + \epsilon_i > 0\}$ and assume $\epsilon_i$ 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 $\beta$ by maximum rank estimator such as

$$\max_{\beta} \sum_{i<j} sign(d_i - d_j) sign((x_i - x_j)\beta) \tag{9.6}$$

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.

Finally, Sognian Chen (1999) uses the additional assumption of symmetry of the distribution of $\epsilon$ 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

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

[25] Note that sgn$(\cdot)$ is not invertible.

[26] See the discussion on selection the Powell's way below in Section 10.6.4

sum does not get rid of the intercept. So, estimate something like

$$\max_{\beta} \sum_{i<j} sign(d_i - d_j - 1) sign((x_i + x_j)\beta). \tag{9.7}$$

## 9.2. Multinomial Choice Models

See McFadden (1984) for a summary of pioneering research, the Limdep v.7 (1995) manual for recent references and examples of application, and S. 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)

### 9.2.1. Unordered Response Models

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

**Example 9.4.** *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:

$$\frac{p_j}{p_j + p_m} = F(\beta_j' x) \Rightarrow \frac{p_j}{p_m} = \frac{F(\beta_j' x)}{1 - F(\beta_j' x)} = G(\beta_j' x) \Rightarrow p_j = \frac{G(\beta_j' x)}{1 + \sum_{k=1}^{m-1} G(\beta_k' x)} \tag{9.8}$$

**Remark 51.** *Note that in our binary choice example ($m = 2$), we also started by defining $p_j(p_j + p_m)^{-1} = p/(p + 1 - p) = p = F(\beta_j' x)$.*

**Multinomial Logit** 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

$$\max_{\beta_1, \beta_2, \ldots, \beta_{m-1}} \log L = \sum_{i=1}^{N} \sum_{j=1}^{m} y_{ij} \log p_{ij}, \text{ where} \tag{9.9}$$

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

**Remark 52.** *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 9.3). 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 $\alpha_j$.*

**Exercise 9.7.** *Verify that the benchmark likelihood equals $\prod_{j=1}^{m}\left(\frac{N_j}{N}\right)$ where $N_j=\sum_{i=1}^{N}y_{ij}$.*

**Exercise 9.8.** *What happens in the commuting choice example when all males choose to drive?*

**Remark 53.** *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 54.** *There is a utility maximization model of individual choice which leads to the multinomial logit, assuming additivity of disturbances $y_{ij}^{*}=V_j(x_i)+\epsilon_i$ with $y_{ik}=1$ iff $\forall j\neq k$ $y_{ik}^{*}>y_{ij}^{*}$ and assuming $\epsilon_i$'s are iid type I extreme value distribution.*[27] *$P[y_k=1|x_i]$ corresponds to the joint occurrence of $V_k(x_i)+\epsilon_k>V_j(x_i)+\epsilon_j$ $\forall j\neq k$, 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 $\exp(V_k(x_i))/\sum_{j=1}^{m}\exp(V_j(x_i))$.*

**Exercise 9.9.** *Verify the derivatives in [M] p.36 and show $\frac{p_j}{p_k}=\exp[(\beta_j-\beta_k)'x]$.*

[27] 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.

The joint distribution of extreme value $\varepsilon$s does not involve any unknown parameters and is therefore not capable of approximating a wide range of stochastic structures. Furthermore, the multinomial logit model (MNL) assumes that disturbances are *independent* (see Remark 54). 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 9.9 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.[28]

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 $\beta_j$ 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 55.** *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.*

**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 $x_i$.

$$P[y_i = k] = \frac{\exp(\beta' z_{ik})}{\sum_{s=1}^{m} \exp(\beta' z_{is})} \tag{9.10}$$

**Remark 56.** *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 9.10.** *Show* $\frac{p_j}{p_k} = \exp[(z_j - z_k)'\beta]$.

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

**Remark 57.** *The elimination by aspect model ([M]3.4) represents another way how to account for similarities between alternative choices.*

**Multinomial Probit and GEV** Unlike MNL, the multivariate probit allows for a full correlation structure with $\epsilon \sim N(0,\Sigma)$ and requires $m-1$ dimensional numerical integration. One has to impose normalization and identification restrictions on the $m(m-1)$ free elements $\sigma$ of the $m \times m$ matrix $\Sigma$.The likelihood requires $m-1$ dimensional numerical integration, numerical 1st and 2nd derivatives and is therefore potentially messy.

**Example 9.5.** *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 $\eta_{12}$ and $\eta_{13}$, the 2x2 matrix $\widetilde{\Sigma}$, from the original $\sigma$ elements. Finally,*

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

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 $\varepsilon$ correlation across choices:

$$F(\varepsilon_1, \varepsilon_2, ..., \varepsilon_m) = \exp\left[-G(\exp(-\varepsilon_1), ..., \exp(-\varepsilon_m))\right], \tag{9.11}$$

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 9.2.2).

**Example 9.6.** *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 $\sigma$ 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 54).*

### 9.2.2. 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 prefferences 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 9.7.** *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* $\beta_H$ *and* $\beta_C$ *– 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. Below, we will focus on a simpler case where the set of choices has a nested structure (not necessarily corresponding to time sequence).

**Nested Multinomial Logit Model** Here, 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 9.8.** *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.*

(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 $p_{ij} = p_i p_{j|i}$,[29] where as it turns out $p_{j|i}$ only

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

involves $\beta$ but not $\alpha$:

$$p_{j|i} = \frac{\exp(\beta' x_{ij} + \alpha' z_i)}{\sum_{n=1}^{N_i} \exp(\beta' x_{in} + \alpha' z_i)} = \frac{\exp(\beta' x_{ij})}{\sum_{n=1}^{N_i} \exp(\beta' x_{in})}. \tag{9.12}$$

Similarly,

$$p_i = \frac{\exp(I_i + \alpha' z_i)}{\sum_{m=1}^{C} \exp(I_m + \alpha' z_m)}, \text{ where } I_i = \log \left[ \sum_{n=1}^{N_i} \exp(\beta' x_{in}) \right] \tag{9.13}$$

is the so-called inclusive value (the total contribution of each house in a neighborhood). One can therefore first estimate $\beta$ off the choice within neighborhoods (based on $p_{j|i}$) and then use the $\widehat{\beta}$ to impute $\widehat{I_i}$ and estimate $\alpha$ 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.[30]

**Remark 58.** *The assumed forms of utility functions can differ across branches and decisions.*

**Remark 59.** *The NMNL gives identical fits to data as the hierarchical elimination by aspect model.*

(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 54 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):

$$p_i = \frac{\exp[(1-\sigma)I_i + \alpha' z_i]}{\sum_{m=1}^{C} \exp[(1-\sigma)I_m + \alpha' z_m]} \tag{9.14}$$

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

**Remark 60.** *Nlogit is in* Limdep.

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

### 9.2.3. Ordered Response Models

**Example 9.9.** *Ratings, opinion surveys, attained education level.* $0 < 1 < 2$ *but* $1 - 0 \neq 2 - 1$.

Use threshold "constants" to split the range of $\epsilon$s. A common $\beta$ affects the decision among many alternatives.

**Example 9.10.** *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.674).

## 9.3. Models for Count Data

**Example 9.11.** *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 $\lambda_i$ and that these processes are independent across plants. To bring in $x'\beta$ assume $\ln \lambda_i = \beta^{'} x_i$ and maximize the likelihood:

$$\max_{\beta} L = \prod_i \exp(-\lambda_i)\lambda_i^{y_i}(y_i!)^{-1}. \tag{9.15}$$

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:

$$E[y_i \mid x_i] = V[y_i \mid x_i] = \lambda_i = \exp(\beta^{'} x_i) \Longrightarrow \frac{\partial E[y_i \mid x_i]}{\partial x_i} = \lambda_i \beta \tag{9.16}$$

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 $\epsilon$ out of likelihood before maximization (as in Random Effect Probit) and maximize w.r.t. both $\beta$ and $\alpha$, the overdispersion parameter.[31]

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 10.4).

### 9.4. 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

$$\begin{aligned} P[0|x_i] &= P[Z_i = 1|x_i] + P[Z_i = 0|x_i]P[y_i^* = 0|x_i] = \\ &= F(\gamma' x_i) + (1 - F(\gamma' x_i)) \exp(-\lambda_i)\lambda_i^{y_i}(y_i!)^{-1}. \qquad (9.17) \end{aligned}$$

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

**Example 9.12.** *Use a binary choice model to estimate the probability of having any children or not and combine this decision with the ordered or count model of how many children you have given you decided to have some.*

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

# 10. Limited Dependent Variables

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

**Example 10.1.** *Zero expenditure and corner solutions: labor force participation, smoking, demand. A change in the $x$s affects both the usual intensive margin and the extensive margin of the corner solution.*

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

### Technical Preliminaries

**(a)** Means of truncated distributions:

$$E[\varepsilon \mid \varepsilon \geq c] = \int_c^\infty \frac{\varepsilon f(\varepsilon)}{1-F(c)} d\varepsilon \tag{10.1}$$

**Exercise 10.1.** *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.*[32] *Also find $V[\varepsilon \mid \varepsilon \geq c]$.*

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

$$E[\varepsilon^c] = F(c)c + [1 - F(c)]E[\varepsilon \mid \varepsilon \geq c] \tag{10.2}$$

## 10.1. Censored Models

**Example 10.2.** *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 $y_i^*$ :

$$\begin{aligned} y_i^* &= \beta' x_i + u_i \text{ with } u_i \sim N(\mu, \sigma^2) \qquad (10.3) \\ y_i &= y_i^* \text{ iff } y_i^* > c \\ y_i &= c \text{ iff } y_i^* \leq c \end{aligned}$$

[32] 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. As a result, OLS will typically result in coefficients biased towards zero.

WLOG[33] 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 10.2.** *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.*

Therefore we use MLE:

$$L = \prod_{y_i^* > c} \frac{1}{\sigma}\varphi\left(\frac{y_i - x_i'\beta}{\sigma}\right) \prod_{y_i^* \leq c} \Phi\left(\frac{c - x_i'\beta}{\sigma}\right), \tag{10.4}$$

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

**Remark 61.** *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.*

The model has many extensions allowing for multiple and variable thresholds $c_i$, heteroscedasticity,[34] panel data random effects, sample selection (see section 10.4), SEM, nested structures, and non-normality (see LIMDEP manual, Ch. 27).[35]

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

[34] The available methods use a parametric assumption on the form of heteroscedasticity. Semiparametric estimators are a focus of much current research, see section 10.3.

[35] 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.

How do we interpret the coefficients? There are 3 types of predictions we can consider, using the definitions of $E[y^* \mid x]$, $E[y \mid x]$ and $E[y \mid x, y^* > 0]$.

**Exercise 10.3.** *Find the expressions for these 3 conditional mean functions and their derivatives w.r.t. $x$.*

**Remark 62.** *There is little theoretical justification for Tobit in rational choice models (see [P]p.141).*

**Remark 63.** *The estimator is biased in 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 10.3 for recent heteroscedasticity-robust alternatives to Tobit such as CLAD.*

**Remark 64.** *The likelihood is only piece-wise continuous.*

**Remark 65.** *First think: Are you really interested in the intensive and extensive margin separately?*

**Remark 66.** *Of course the model can be easily extended to censoring from above* and *below.*

**Remark 67.** *Up to now the $c$ threshold was exogenous. There is a wide variety of models where the classification criteria are endogenous.*

**Remark 68.** *Under joint normality of error terms, one probably could instrument in a Tobit. Of course, heteroscedasticity kills it, so one should use CLAD or Ahn and Powell, etc., but then can you do IV there?*

**Grouped Data**

**Example 10.3.** *Wages reported only in ranges, i.e. $w_i \in [\$10000, \$20000)$, i.e. $w_i \in [c_{j-1}, c_j)$ $j = 1, \ldots J$*

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 $y_i$s in those ranges is

$$\ln L_{HL} = \sum_{i=1}^{N} \left\{ \ln[\Phi(\eta H - x_i^{'}\gamma) - \Phi(\eta L - x_i^{'}\gamma)] \right\}, \tag{10.5}$$

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

$$E[y_i^* \mid x_i] = x_i^{'}\beta + \sigma \frac{\varphi_{iL} - \varphi_{iH}}{\Phi_{iH} - \Phi_{iL}} \tag{10.6}$$

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

### 10.2. Truncated Models

**Example 10.4.** *Only have data on low income households when studying the impact of variable $x$ on income $y$.*

$$L = \prod_{y_i^* > c} \frac{1}{\sigma} \varphi \left( \frac{y_i - x_i^{'}\beta}{\sigma} \right) \left[ 1 - \Phi \left( \frac{c - x_i^{'}\beta}{\sigma} \right) \right]^{-1} \tag{10.7}$$

Tobit-type model is not feasible here as we do not observe $x$s for the $y = 0$ 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 10.6 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 $x$s affecting the extensive and intensive margins to be the same ([G]p.700).

**Example 10.5.** *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 69.** *Assuming the $x$s 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 a LR test:*

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

**Remark 70.** *But the disturbances from the two separate equations are likely dependent, which is why we need a sample selection model!*

### 10.3. 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 $\varepsilon$ 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 $\varepsilon$. 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 $\beta$. *Assume* that $f_\varepsilon(s|x)$ is symmetric around zero and unimodal. Then for $x_i'\beta > 0$, the $\varepsilon$ is truncated at $0 - x_i'\beta$ so a symmetric truncation of $\varepsilon$ 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

$$\sum I\{x_i'\beta > 0\} I\{y_i < 2x_i'\beta\} \left[y - x_i'\beta\right]^2 . \tag{10.8}$$

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

$$\sum I\{x_i'\beta > 0\} \left[\min(y_i, 2x_i'\beta) - x_i'\beta\right]^2 . \tag{10.9}$$

**Censored Least Absolute Deviation** 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 median is not affected by censoring. (This is true 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, again, we work only with variation in the $x_i'\beta > 0$ area.)

The main assumption of the estimator is zero median of $F_\epsilon(s|x)$. We consider $median(y_i^*|x_i) = \max\{x_i'\beta, 0\}$ and note that medians are estimated using LAD.[36] The CLAD is found by minimizing

$$\sum \left| y_i - \max\{0, x_i'\beta\} \right| . \tag{10.10}$$

It can be used in a Tobit or Heckman's $\lambda$ setup (see below) in presence of heteroscedasticity and it is even more robust than STLS. CLAD is programmed into Stata.[37] 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.[38]

[36] Estimate $\delta$,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 14.6).

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

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

### 10.4. Introduction to 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 10.6.** *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 10.7.** *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.*

**Example 10.8.** *College wage effect (remember Card IV for ability bias?).*

**Example 10.9.** *Past unemployment predicts future unemployment.*

**Example 10.10.** *Average wage over the business cycle: seems flatter due to selective drop out of work.*

### 10.5. Endogenous Stratified Sampling

It occurs when the probability that an individual is observed in the sample depends on $y$s.[39]

**Remark 71.** *Stratification (over/undersampling) based on $x$ variables presents no problem for OLS, as long as there is no parameter heterogeneity across strata (see Remark 6).*

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

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 $x$s. Here $n_j$ is the number of observations in data from strata $j$ and $N_j$ 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[40,41]

$$\begin{aligned} L(y_i|x_i) &= L_i^{-1} p_1 \phi((y_i - x_{ij}^{'}\beta)/\sigma) \text{ if } y_i < L \text{ and} \qquad (10.11) \\ L(y_i|x_i) &= L_i^{-1}(1-p_1)\phi((y_i - x_{ij}^{'}\beta)/\sigma)) \text{ if } y_i > L \text{ where} \\ L_i &= p_1 \Phi[(L - x_{ij}^{'}\beta)/\sigma] + (1-p_1)(1 - \Phi[(L - x_{ij}^{'}\beta)/\sigma]). \end{aligned}$$

**Example 10.11.** *Another example was the choice based sampling method of section 9.1.5.*

In the next subsection, we will consider cases when truncation or censoring occurs with stochastic or unobservable thresholds.

## 10.6. Models with Self-selectivity

**Example 10.12.** *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 $\lambda$), second, when we observe $y$ under all *chosen* alternatives (union wages, switching regression).

[40] The formulas in [M]6.10 are conditional on $y_{ij}$ actually being drawn from a given strata $j$.

[41] See [Mp.173] for the asymptotic justification of WLS based on this MLE.

### 10.6.1. 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.[42]

Assuming that

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

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 then 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 72.** *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 70).*

**Remark 73.** *Note that $\sigma_H^2 - \sigma_{HF} < 0$ and $\sigma_F^2 - \sigma_{HF} < 0$ cannot happen due to Cauchy-Schwartz inequality.*

[42] For purposes of policy evaluation we will need to deal with estimating the counterfactual. See subsection 11.

### 10.6.2. Heckman's $\lambda$

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

$$\begin{aligned} y_{i1} &= x_{i1}^{'}\beta_1 + u_{i1} \\ y_{i2} &= x_{i2}^{'}\beta_2 + u_{i2}, \end{aligned} \tag{10.12}$$

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

**Example 10.13.** *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:

$$\begin{aligned} E[y_{i1}|x_i, y_{i2} &> 0] = x_{i1}^{'}\beta_1 + E[u_{i1}|\textit{selection rule}] = \\ &= x_{i1}^{'}\beta_1 + E[u_{i1}|y_{i2} > 0] = x_{i1}^{'}\beta_1 + E[u_{i1}|u_{i2} > -x_{i2}^{'}\beta_2]. \end{aligned} \tag{10.13}$$

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 70 and $\sigma_{HF}$ in the Roy's model).

If we assume that $u_{i1}$ and $u_{i2}$ are jointly normal with correlation $\sigma_{12}$ and variances $\sigma_1^2$ and $\sigma_2^2$ 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:

$$E[y_{i1}|x_i, y_{i2} > 0] = x_{i1}^{'}\beta_1 + \frac{\sigma_{12}}{\sigma_2}\frac{\phi(x_{i2}^{'}\beta_2/\sigma_2)}{\Phi(x_{i2}^{'}\beta_2/\sigma_2)} = x_{i1}^{'}\beta_1 + \sigma_\lambda\lambda(x_{i2}^{'}\beta_2). \tag{10.14}$$

While we can numerically identify $\sigma_\lambda$ from $\beta_1$ even when $x_{i2} = x_{i1}$ because $\lambda$ 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 $x$s).

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 ran 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 $\Gamma$). 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, you need to correct your standard errors!). Hence, the variance-covariance of the error term in the second-step regression is composed of $\Gamma$ 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 10.6.4 below).

**Example 10.14.** *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 10.15.** *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 74.** *Asymptotic distribution: two stage methods are efficient in one iteration.*

**Remark 75.** *The $\lambda$ method is also applicable in unbalanced panel data, see Lafontaine and Shaw (1995) for an example. Franchisees who go out of business have shorter $T_i$. Their fixed effect estimates eliminate most of the selection bias suggesting that within-firm variation in selection has little effect.*

**Remark 76.** *Estimating parametric Limdep models with a $\lambda$ on the RHS is a big problem, especially with heteroscedasticity, which kills consistency. The $\lambda$ problem is that selection affects the whole distribution and $\lambda$ only fixes the expectation (centering).*

**Remark 77.** *If the unobservable selection threshold is time constant we can use a fixed effect panel data model to deal with it.*

### 10.6.3. Switching Regression

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

**Example 10.16.** *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:

$$y_i = x_i'\beta + \alpha D_i + \varepsilon_i, \tag{10.15}$$

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 78.** *The consistency is contingent on correct specification of the error distribution $F_\nu(\cdot)$.*

**Remark 79.** *Again, the standard errors of $\beta$ need to be corrected for the estimation of $\gamma$, see Lee (1981), and the discussion of Delta method above for $\beta_2$.*

**Remark 80.** *The restrictions of the model can be somewhat relaxed, see Robinson (1988).*

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 $\lambda$. See HW#6.

**Example 10.17.** *Engberg and Kim (1995) study the intra-metropolitan earnings variation: is it caused by person or place effects? People $i$ choose their location $j$ (inner city/poor suburb/rich suburb) based on their (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 $\lambda$ 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:*

### 10.6.4. Semiparametric Sample Selection

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 not want to assume a particular form for the selection term (i.e., I am not willing to assume a particular distribution $f$), follow Powel 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

$$\left[\sum K\left(\frac{(x_i - x_j)\widehat{\gamma}}{n}\right)(x_i - x_j)(x_i - x_j)'\right]^{-1}\left[\sum K\left(\frac{(x_i - x_j)\widehat{\gamma}}{n}\right)(x_i - x_j)(y_i - y_j)\right] \tag{10.16}$$

**Example 10.18.** *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.*

**Index Sufficiency (Matching on Propensity Score)** Ahn and Powel further suggest that one needs to match on the probability of selection, as there is no need for any $\gamma$ here an all can be done non-parametrically. See Rosenbaum and Rubin for early theoretical work on matching using the propensity score. More recent work arguing that a useful way of controlling for selection bias is to compare the outcome for those with similar probability of selection includes Heckman, Ichimura, Smith, and Todd (1995) and Angrist (1995). Matching is practical when the causing variable takes on two values (union status, military service, see Angrist 1998). We return to this work below within a general setup.

**Remark 81.** *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 $x$.*

## 11. Program Evaluation

This is where the above discussion (and this whole course) naturally anchors, but a thorough treatment goes beyond the scope of the course. Still, some leads are given below:

After all, evaluation of social programs is what much of true micro-econometrics is all about. (Otherwise we simply use regressions as a statistical data description tool, not to estimate causal parameters.) We ask how to estimate the effect of a social program (policy) (i.e., participation in a training program, change in college tuition) in absence of controlled experiments ([M]9.2.). How can one create the counterfactual? (Return to the first class of the course for a broad introduction to **causal inference**; also, see again the section on sample selection. Further reading starts with Holland (1986) and continues with many of the recent Heckman's articles. See also Angrist and Krueger (1999), or Card (1993).

**Setup**[43] Effect of a training program: $y_{1i}$ are earnings with training, $y_{0i}$ are earnings without training (think Roy model). We only look at 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). We want to know $E[y_{1i} - y_{0i}]$ – under random assignment. We also want to know $E[y_{1i} - y_{0i}|D_i = 1]$ and $E[y_{1i} - y_{0i}|D_i = 0]$. Think of the first parameter (effect of treatment on treated, TT): the data only provides $E[y_{1i}|D_i = 1]$ but $E[y_{0i}|D_i = 1]$ is the counterfactual.

**Experimental Solution** Almost ideal for measuring causal effects (think of medical trials using placebo). 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

[43] Based on Heckman, Ichimura and Todd (1995).

can be used as a benchmark for the accuracy of sample selection techniques that we need when we have no experiment.

**Remark 82.** *However, experiments are costly, often socially unacceptable (in Europe), and people may behave differently knowing they are in an experiment (think of expanding medical coverage).*

**Remark 83.** *Even with experimental data, there are often problems of selectivity. See Ham and LaLonde (1996) for a duration study.*

**Non-experimental Solution** 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 we can (this is the matching method): $E[y_0|D=1]=E\{E[y_0|D=1,X]|D=1\}=E\{E[y_0|D=0,X]|D=1\}$ so estimate RHS by

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

that is estimate a regression using $D=0$ but predict outcome using $D=1$. But what if the $X$ support of $E[y_0|D=1,X]$ and $E[y_0|D=1,X]$ does not coincide? We need to predict out of the sample of $X$ for $D=0$. This is hard non-parametrically especially given that the dimension of $X$ may be high. So here index sufficiency enters the picture: Instead of conditioning on $X$, it is enough to condition on $P(X)$, the probability of selection (see above). So we condition on $P(X)$ over the **common support** – compare the outcome for those (pairs of) individuals (ENPs compared to Treatments) with similar probability of participation in the program.

**Big Picture** 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. (See, e.g., Heckman's Nobel lecture.[44]) So focus on (1).

[44] For example at http://www.nobel.se/economics/laureates/2000/heckman-video.html

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!

The literature makes clear the key need to properly define the **policy parameters of interest**: What do we want to know? The effect of the program treatment on the treated (TT; useful for cost-benefit analysis), the effect of the program on untreated (whom we could make participate), the average treatment effect in the population (ATE), or a treatment effect related to a specific new policy.

A recent set of papers by Heckman and others (1998, 2000) shows when IV estimates uninteresting policy parameters. The comparison is between IV (LATE) and sample selection:

First, consider the **Local Average Treatment Effect** interpretation of IV estimates.[45] Oversimplifying: Suppose that the effect of $x$ on $y$ differs across groups of the population (parameter heterogeneity). 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. So the IV estimate is the treatment effect on specific groups–it is a "local" effect.

**Example 11.1.** *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 11.2.** *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 84.** *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.*

Second, consider the research work by Heckman and coauthors, who use the concept of the marginal treatment effect (MTE) to unify the IV and sample selection literature (see example below for estimating the returns to education).

[45] For an introduction to LATE, see <http://www.irs.princeton.edu/pubs/pdfs/415.pdf>.

**Example 11.3.** *Carneiro, Heckman and Vytlacil (2002): There are two general approaches to estimating the return to schooling (college/high school, $S = 1/0$).*

*(a) Griliches (1977) Human capital is homogenous. There is one "true" effect of schooling on wages $\beta$. Schooling may be correlated with unobservables through ability bias or measurement error (see first half of this course) so we need to run IV for $S$ in estimating:*[46]

$$\ln Y = \alpha + \beta S + U.$$

*(b) Roy (1951), Willis and Rosen (1979) Human capital is heterogeneous. People know (estimate) their* varying *returns from education and* act upon the size of the return*:*

$$\begin{aligned} \ln Y_0 &= \alpha + U_0 \\ \ln Y_1 &= \alpha + \overline{\beta} + U_1, \end{aligned}$$

*so that the causal effect $\beta = \ln Y_1 - \ln Y_0 = \overline{\beta} + U_1 - U_0$. There is a* distribution *of returns (random coefficients, ex post causal effects) that cannot be summarized by one number $\beta$ as in (a). There is a range of policy parameters (TT, UT, ATE, etc.) that can be expresses as differentially weighted averages (integrals over population) of the marginal treatment effects. MTE is the effect of $S$ on a person with $X = x$ and $U = u$ that is just indifferent between taking up college. IV and OLS can also be expressed as weighted averages of MTE, but the weights are not those of TT, 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.*[47] *Also note that there are more econometric problems here compared to (a): $COV(S, U_0) \neq 0$ as before, but also $COV(\beta, U_0) \neq 0$ and crucially $COV(\beta, S) \neq 0$.*

[46] Alternatively, in terms of potential outcomes write the common effec model as $\ln Y_0 = \alpha + U$, $\ln Y_1 = \alpha + \beta + U$.

[47] In the IV literature on returns to schooling, we worry about (a) upward ability bias ($COV(S, U) \neq 0$), (b) downward measurement error bias, and (c) the weak instrument bias, where $\frac{COV(U,S)}{COV(IV,S)}$ is large because $COV(IV, S)$ is small. Card uses college proximity as an IV (see example 7.2). Typically in the IV literature $\beta_{IV} > \beta_{OLS}$. Now think of the LATE IV interpretation: $\beta_{IV}$ 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 $\beta$,but don't attend college because of credit constraints. Heckman et al. say this interpretation is wrong. They say that there is a large positive sorting gain (comparative advantage), Willis and Rosen prevail, and true $\beta >$ IV>OLS.

## 12. Duration Analysis

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).[48]

**Example 12.1.** *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, etc.*

The advantage of duration models is in their ability to handle time changing $x$s (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].

### 12.1. 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 $x$s 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 12.1.** *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.*

[48] Think of how we built a model from Poisson distribution as the natural model for count processes.

**Remark 85.** *A first approximation to the hazard, ignoring both observed and unobserved differences is the so called Kaplan-Meier statistic (also called empirical hazard):*

$$\lambda(t) = \frac{\#[exit(t)]}{\#[risk(t)]} \text{ with } \sigma_\lambda(t) = \sqrt{\frac{\lambda(t)(1-\lambda(t))}{\#[risk(t)]}}. \tag{12.1}$$

**Exercise 12.2.** *Verify the formula for $\sigma_\lambda$. 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

$$\lambda(t) = \lim_{h \to 0} \frac{1}{h} \Pr(t \leq t^* < t + h \mid t^* \geq t) \tag{12.2}$$

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 86.** *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 12.2.** *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.*[49]

**Exercise 12.3.** *Can the logit model be interpreted as an approximation to a proportional hazard model?*

**Remark 87.** *One can trick* LIMDEP *or other software to estimate the logit duration model.*

[49] For proportional hazard models, Elbers and Ridder show that the heterogeneity distribution and the duration dependence can be separately identified.

### 12.2. 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 12.1 and below).[50]

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.[51]

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). 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 $\beta$s. There is no known way of correctly constructing the asymptotic standard errors, since the dimension of the parameter space depends on the sample size. So assume the number of points of support is fixed to invoke standard asymptotics.

**Remark 88.** *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 89.** *Note that if there is no variation in the $x$s independent of duration, identification will be difficult.*

#### 12.2.1. 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.

[50] 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.

[51] 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.

To use the "random effects" approach, estimate a discrete mixing distribution $p(\theta)$ of an unobserved heterogeneity term $\theta$ 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 $\theta_k^j$, $k = 1, 2, ..., N_\theta^j$. 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 $\lambda$ 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

$$L_e(t) = \lambda(t, x_t) \prod_{v=1}^{t-1} [1 - \lambda(v, x_v)]. \tag{12.3}$$

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

$$L_e^q(t) = \lambda_q(t, x_t) \prod_{v=1}^{t-1} [1 - \lambda_q(v, x_v)][1 - \lambda_l(v, x_v)], \tag{12.4}$$

where $\lambda_q$ and $\lambda_l$ 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

$$L_e^l(t) = \lambda_l(t, x_t) \prod_{v=1}^{t-1} [1 - \lambda_q(v, x_v)][1 - \lambda_l(v, x_v)]. \tag{12.5}$$

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

$$S_e(T) = \prod_{v=1}^{T} [1 - \lambda_q(v, x_v)][1 - \lambda_l(v, x_v)]. \tag{12.6}$$

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

$$L_u^n(t) = \sum_{k=1}^{N_\theta^n} \sum_{m=1}^{N_\theta^r} p(\theta_k^n, \theta_m^r) L_u^n(t|\theta_k^n, \theta_m^r), \tag{12.7}$$

where $p(\theta_k^n, \theta_m^r)$ is the probability of having the unobserved components $\theta_k^n$ and $\theta_m^r$ in the new job and recall hazards respectively, and where

$$L_u^n(t|\theta_k^n, \theta_m^r) = \lambda_n(t, x_t|\theta_k^n) \prod_{v=1}^{t-1} [1 - \ \lambda_n(v, x_v|\theta_k^n)] \ [1 - \ \lambda_r(v, x_v|\theta_m^r)]. \tag{12.8}$$

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 12.9 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 $w - r - s - t$).

$$L(s, w) = \sum_{k=1}^{N_\theta^q} \sum_{m=1}^{N_\theta^l} p(\theta_k^q, \theta_m^l) L_e^l(s|\theta_k^q, \theta_m^l) L_e^q(w|\theta_k^q, \theta_m^l) \tag{12.9}$$

Here $\theta^q$ and $\theta^l$ denote the unobserved terms entering quit and layoff hazards respectively and

$$L_e^l(s|\theta_k^q, \theta_m^l) = \lambda_l(s, x_s|\theta_m^l) \prod_{v=t+1}^{s-1} [1 - \ \lambda_q(v, x_v|\theta_k^q)] \ [1 - \ \lambda_l(v, x_v|\theta_m^l)] \ , \tag{12.10}$$

$$L_e^q(w|\theta_k^q,\theta_m^l) = \lambda_q(w,x_w|\theta_m^l) \prod_{v=r+1}^{w-1} [1 - \ \lambda_q(v,x_v|\theta_k^q)]\ [1 - \lambda_l(v,x_v|\theta_m^l)]\ .$$

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

$$L^{u,l,u}(t,s,T) = \sum_{k=1}^{N_\theta^u}\sum_{m=1}^{N_\theta^q}\sum_{n=1}^{N_\theta^l} p(\theta_k^u,\theta_m^q,\theta_n^l)L_u(t|\theta_k^u)L_e^l(s|\theta_m^q,\theta_n^l)S_u(T|\theta_k^u), \quad (12.11)$$

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 12.10 . 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 ($N_\theta^u$, $N_\theta^q$ and $N_\theta^l$) is determined from the sample likelihood (using Schwarz or Akaike criterion).[52] Note the assumption of $\theta^u$, $\theta^q$ and $\theta^l$ 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:

|  | $\theta_1^l$ | $\theta_2^l$ | ... | $\theta_N^l$ |
|---|---|---|---|---|
| $\theta_1^q$ | $p(\theta_1^q, \theta_1^l)$ | $p(\theta_1^q, \theta_2^l)$ | ... | $p(\theta_1^q, \theta_N^l)$ |
| $\theta_2^q$ | $p(\theta_2^q, \theta_1^l)$ | $p(\theta_2^q, \theta_2^l)$ | ... | $p(\theta_2^q, \theta_N^l)$ |
| ... | ... | ... | ... | ... |
| $\theta_M^q$ | $p(\theta_M^q, \theta_1^l)$ | $p(\theta_M^q, \theta_2^l)$ | ... | $p(\theta_M^q, \theta_N^l)$ |

3. One factor loading:

| | |
|---|---|
| $p(\Theta_1)$ | $\Theta_1 = \{\theta_1^l, c\theta_1^q\}$ |
| $p(\Theta_2)$ | $\Theta_2 = \{\theta_2^l, c\theta_2^q\}$ |
| ... | ... |
| $p(\Theta_N)$ | $\Theta_N = \{\theta_N^l, c\theta_N^q\}$ |

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 heterogeneity. A typical stayer model would then parametrize an individual's contribution to the likelihood as follows:

$$L(t) = p_s + (1 - p_s)\{\lambda_q(t, x_t) \prod_{v=1}^{t-1} [1 - \lambda_q(v, x_v)]\ \},$$

[52] See Baker and Melino (2000).

where $p_s$ is the probability of never leaving employment and $\lambda_q$ is a quit hazard. See Jurajda (2002) for details on estimation.

6. Continuous parametric distributions of heterogeneity, for example Weibull.

### 12.2.2. 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 $\beta$ from the fresh spells. See also exercise 12.1.

### 12.2.3. 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 $x$s. Interpret the difference between expected durations as the magnitude of the particular $\beta$. The expected duration is computed as

$$E(t|X) = \sum_{i=1}^{I} \frac{\sum_{t=1}^{\infty} t f_i(t)}{I}, \tag{12.12}$$

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.[53] 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:

$$\begin{aligned} f_i(t) &= \sum_{k=1}^{N} p(\theta_k^r, \theta_k^n) f_i(t|\theta_k^r, \theta_k^n), \text{ where} \\ f_i(t|\theta_k^r, \theta_k^n) &= \{\lambda_r(t, x_{it}|\theta_k^r) + \lambda_n(t, x_{it}|\theta_k^n) - \lambda_r(t, x_{it}|\theta_k^r)\lambda_n(t, x_{it}|\theta_k^n)\} \times \\ & \prod_{v=1}^{t-1} [1 - \ \lambda_r(v, x_v|\theta_k^r)] \ [1 - \ \lambda_n(v, x_v|\theta_k^n)] \ . \end{aligned}$$

**Remark 90.** *In multiple-state multiple-spell data, single-spell duration simulations do not provide a full picture. See, e.g., Jurajda (2002).*

[53] A simpler (biased) approach is to evaluate the expected duration at a mean individual $\overline{x}$.

### 12.2.4. Partial Likelihood

Cox (1972, 1975): estimate $\beta$ 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

$$\frac{\lambda(t_{(1)}, x_{(1)})}{\sum_{i=1}^{n}\lambda(t_{(1)}, x_{(i)})} = \frac{\exp(x_{(1)}'\beta)}{\sum_{i=1}^{n}\exp(x_{(i)}'\beta)}. \tag{12.13}$$

In the absence of information about the form of duration dependence, only the information about the *order* of the spell durations is used.

**Remark 91.** *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* $\exp(x_i'\widehat{\beta})$. *See [P].*

# Part IV
# Some Recent Topics in Econometrics

## 13. 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 recent surveys see Eckstein and Wolpin (1989) and Rust (1992, 1994).

**Example 13.1.** *Engberg (1992) estimates a structural model of job search allowing for unobserved heterogeneity.*

**Example 13.2.** *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 13.3.** *Other applications include equilibrium models of unemployment (e.g. van den Berg and Ridder 1993) or local jurisdictions*[54] *(Epple and Sieg 1996).*

## 14. Nonparametrics

The very opposite of the structural models. We already mentioned the use of semiparametric methods in the estimation of discrete choice models (section 9.1.6) and in the selection bias models (section 10.6.4). Here, we will discuss the basic nonparametric methods underlying the semi-parametric applications.

[54] Look at an equilibrium distribution of households by income across communities.

### 14.1. 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[55], see [N] and Silverman (1986).

### 14.2. K-th Nearest Neighbor

Define $J(x) = \{i : x_i$ is one of the $K$ 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.

### 14.3. 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.[56]

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)

[55] The bandwidth can be also data-dependent.

[56] 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.

- 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 92.** STATA *has a kernel smoother and does local linear regression. Advanced programs are available on the www for* S-PLUS *(http:\\lib.stat.cmu.edu\).*

**Remark 93.** *Kernel estimation is basically a LLR on just the constant term.*

**Remark 94.** *There are also extensions of the localization idea to the MLE framework, see Tibshirani and Hastie (1987).*

### 14.4. Multidimensional Extensions

The curse of dimensionality is severe. To have a reasonable speed of convergence need very large samples. There are a few ways how to proceed (the last two have applications in econometrics):

- 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:

$$m(x) = g_1(x\beta_1) + g_2(x\beta_2) + \dots.$$

- Index sufficiency. It was mentioned in the semiparametric sample selection literature.
- Average derivative estimation: If I am interested in $\theta = E\left\{\frac{\partial m(x_i)}{\partial x_i}\right\}$ then $\theta$ can be estimated with $\sqrt{n}$ speed of convergence. Example: binary choice or tobit models.

### 14.5. Partial Linear Model

For a model $y = z\beta + f(x) + \varepsilon$, where both $z$ and $x$ are scalars, estimators of $\beta$ can be constructed which are asymptotically normal at the $\sqrt{n}$ speed of convergence. See Yatchew, A. (1998).

### 14.6. Quantile Regression

Also related to other non-least-squares regressions, like the LAD estimator. One goal of these robust regression methods is to reduce the influence of outliers. Minimize

$$\sum_{i=1}^{n} \{\mid t_i \mid + (2\alpha - 1)t_i\} \, K\left(\frac{x_i - x_0}{a_n}\right) \text{ with } t_i = y_i - \sum_{j=0}^{k} \beta_j (x_i - x_0)^j$$

For example with $\alpha = 0.5$ it would be a median regression. See Fan and Gijbels book (p.201) for details and Chamberlain (1982). Also see the LOWESS procedure in S-PLUS.

## 15. 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. Also, they may turn more accurate than asymptotic theory because they are constructed based on the right sample size (see Hall's book from 1992).

  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 $\widehat{\epsilon}$.

- **Empirical Process Method of Showing Consistency of an Extremum Estimator with a Non-smooth Objective Function.** Does not require continuity or differentiability. Among the books on the topic are van der Vaart and Wellner (1996), Pollard (1994), and Pollard (1990).

- **Gibbs Sampler**. A new 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). Further, just about every issue of JASA has a paper with Gibbs sampler these days. For an example of multinomial probit estimation see McCulloch and Rossi (1994).

- **Combining Data Sets**. See Arelano and Meghir (1992) and Angrist (1990). If $y$s and instruments are in one sample and $x$s and instruments are in the other sample, combine using MD.

- **Censored panel-data models with a lagged dependent variable.** See [AH] and note that this course covered only the (easier) part of panel-data econometrics where we assume regressors strictly exogenous as opposed to predetermined.

# References

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, pages 3-29.

Altonji J. and 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.

Angrist J. (1995) "Conditioning on the Probability of Selection to Control Selection Bias", *National Bureau of Economic Research Technical Paper*: 181, June 1995.

Angrist J. (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 J. and A. 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. (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.

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 M., Bo Honoré: *Panel Data Models: Some Recent Developments*, <ftp://ftp.cemfi.es /wp/00/0016.pdf>

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

Baker, Melino (2002) "Duration Dependence and Nonparametric Heterogeneity: A Monte Carlo Study," *Journal of Econmetrics* 2002.

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.

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.

Carneiro P., E. Vytlacil, J. Heckmen (2002): http://www.cerge.cuni.cz/events/ seminars/semfall01/011108_a.asp

Cosslett S.(1981) "Maximum Likelihood Estimator for Choice-Based Samples", *Econometrica*, 49(5), Sept. 1981, pages 1289-1316.

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., J.G. MacKinnon, (19993): *Estimation and Inference in Econometrics*, Oxford University Press, 1993.

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.

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.

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.

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].

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.

Green, W.H.(1997): *Econometric Analysis*, third 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: 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].

Hall, Peter (1992): *The Bootstrap and Edgeworth Expansion*, New York, Springer, 1992.

Ham, 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.

Härdle, W., (1989): *Applied Nonparametric Regression*, Cambridge University Press, 1989.

Hausman, J. (1978) “Specification Tests in Econometrics,” *Econometrica* 46: 1251-1272.

Hausman, J. *et al.* (1991) “Identification and Estimation of Polynomial Errors-in-Variables Models”, *Journal of Econometrics*, 50(3), December 1991, pages 273-95.

Hausman, J., D. McFadden (1984) “Specification Tests for the Multinomial Logit Model”, *Econometrica* 52(5), September 1984, pages 1219-40.

Heckman 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., 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., 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., (1979) “Sample Selection Bias as a Specification Error,” *Ecopnometrica* 47: 153-161.

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.J., (2000) “Causal Parameters and Policy Analysis in Econometrics: A Twentieth Centurtury Perspective,” *Quarterly Journal of Econometrics* 2000.

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.

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.

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.

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.

Cheng H., (1986): *Analysis of Panel Data*, Cambridge U. Press, 1986.

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.

Imbens G. (1992) “An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-Based Sampling,” *Econometrica*, 60(5), September 1992, pages 1187-214.

Jakubson (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.

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.

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.

Koenker, R., Jr.G. Bassett (1982) "Robust Tests for Heteroscedasticity Based on Regression Quantiles," *Econometrica*, 50(1), Jan. 1982, pages 43-61.

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.

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.

Maddala G.S., (1983): *Limited-dependent and Qualitative Variables in Econometrics*, Cambridge U. Press, 1983.

Manski C., McFadden: *Structural Analysis of Discrete Data and Econometric Applications*, <elsa.berkley.edu/users/mcfadden/discrete.html>

Manski C. (1975) "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3(3), Aug. 1975, pages 205-28.

Manski C.(1985) "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27(3), March 1985, pages 313-33.

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].

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.

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 D. (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 D. (1989) "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration", *Econometrica*, 57(5), September 1989, pages 995-1026.

Miller, R. A. (1984) "Job Matching and Occupational Choice," *Journal of Political Economy,* 92(6), December 1984, pages 1086-120.

Newey, W. (1985) "Generalized Method of Moments Specification Tests," *Journal of Econometrics,* 29: 229-238.

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.

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..

Powel, 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, pages

Pudney, S. (1989): *Modelling Individual Choice*, Basil Blackwell, 1989.

Robinson, P.M. (1988) "Using Gaussian Estimators Robustly," Oxford Bulletin of Economics and Statistics; 50(1), February 1988, pages 97-106.

Rosenbaum P., D. Rubin (1984) "Estimating the Effects Caused by Treatments: Comment [On the Nature and Discovery of Structure]", J*ournal 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. (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.

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.

Silverman (1986): *Density Estimation for Statistics and Data Analysis*, London, Chapman & Hall, 1986.

Taber, Chr.R. (2000) "Semiparametric Identification and Heterogeneity in Discrete Choice Dynamic Programming Models," *Journal of Econometrics,* 96(2), June 2000, pages 201-29.

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.

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.