Advanced Econometrics

Phu Nguyen-Van

ECONOMIX, CNRS & University of Paris Nanterre

Contact: pnguyenvan@parisnanterre.fr

M2 EEET, Saclay, 2024

References

- *Econometrics textbooks:*
  - **Angrist J.D., Pischke J.-S. (2009), *Mostly Harmless Econometrics - An Empiricist's Companion*, Princeton University Press.**
  - Baltagi B. (2008), *Econometric Analysis of Panel Data*, John Wiley & Sons.
  - Cameron A.C., Trivedi P.K. (2005), *Microeconometrics: Methods and Applications*, Cambridge University Press.
  - **Cunningham S. (2021), *Causal Inference: The Mixtape*, Yale University Press.**
  - Franses P.H., Paap R. (2001), *Quantitative Models in Marketing Research*, Cambridge University Press.
  - Greene W.H. (2011), *Econometric Analysis*, 7th edition, Prentice Hall.
  - Lee M.-j. (2016), *Matching, Regression Discontinuity, Difference in Differences, and Beyond*, Oxford University Press.
  - Maddala G.S. (1999), *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press.
  - **Verbeek M. (2008), *A Guide to Modern Econometrics*, John Wiley & Sons, 3rd edition.**
  - Wooldridge J.M. (2006), *Introductory Econometrics: A Modern Approach*, Thomson South-Western, 3rd edition.
  - **Wooldridge J.M. (2010), *Econometric Analysis of Cross Section and Panel Data*, MIT Press.**

References

- *Stata books*
  - Baum C.B. (2006), *An Introduction to Modern Econometrics Using Stata*, Stata Press, Texas.
  - Cameron A.C., Trivedi P.K. (2010), *Microeconometrics Using Stata*, Stata Press, Texas.
  - Kohler U., Kreuter F. (2012). *Data Analysis Using Stata*. Stata Press, 3rd edition.
  - Mehmetoglu M., Jakobsen T.G. (2017), *Applied Statistics Using Stata. A Guide for the Social Sciences*. Sage, Los Angeles.
  - Mitchell M.N. (2015), *Stata for Behavioral Sciences*, Stata Press, Texas.

References

- *R books*
  - Crawley M.J. (2007), *The R Book*, John Wiley & Sons.
  - Dalgaard P. (2002), *Introductory Statistics with R*, Springer.
  - Everitt B.S., Hothorn T. (2006), *A Handbook of Statistical Analyses Using R*, Chapman & Hall.
  - Fox J. (2002), *An R and S-Plus Companion to Applied Regression*, Sage Publications.
  - Kleiber C., Zeileis A. (2008), *Applied Econometrics with R*, Springer.

Foreword

This document will be served as the basis of the course, which is rather intensive within about 20 hours. The mathematical tools are reduced to minimum and are used when they are necessary. This course only requires basic knowledge on addition, substraction, and some basic matrix operations. When we cannot discard them, a remark is added in order to explain their meaning and utility.

Students can read the textbooks given in the list of references for additional notions on mathematics and statistical and probability theory.

Knowledge on Stata and R softwares is required. The handout is based on Stata examples but the corresponding R examples will be also presented during the course.

Contents

Chapter 2: Regression Diagnostics

1. Outliers
2. Functional forms
3. Multicolinearity
4. Heteroscedasticity and autocorrelation
5. Normality of residuals

Chapter 3: Endogeneity and Instrumental Variables

1. Endogeneity
2. IV estimator
3. GIV estimator
4. GMM estimator

Chapter 4: Maximum Likelihood

1. Maximum likelihood estimator
2. Tests based on maximum likelihood

Chapter 5: Limited Dependent Variable Models

1. What is a limited dependent variable?
2. Binary models
   - Probit model
   - Logit model
   - Marginal effects
   - Goodness-of-fit and tests
3. Multiresponse models
   - Ordered response models
   - Multinomial models
4. Censored and truncated models
   - Standard censored model
   - Truncated model
   - Sample selection model

Chapter 6: Panel Data Models

1. What are panel data?
2. Fixed effects
   - Individual fixed effects
   - Least squares dummy variable model
   - Time fixed effects
   - Twoway fixed effects
3. Random effects
   - Individual random effects model
4. Tests
   - Existence of fixed effects
   - Existence of random effects
   - Fixed effects or random effects
5. Further developments

Chapter 7: Causal Inference

1. Correlation and causality
2. Randommisation
3. Regression and causality
4. Matching
5. Differences-in-Differences
6. Regression discontinuity

Background and Notations

Background and Notations

- *Econometrics* is the interaction of economic theory, observed data, and statistical methods, which aims : (i) to quantify the relationships between different (*economic*) quantities on the basis of available data and using statistical techniques, (ii) to interpret, use or exploit the resulting outcomes appropriately.
- The data can be cross-section (observations on several individuals collected at a certain point of time), time series (observations of an individual collected over a period of time), or panel (observations on several individuals collected over time).

- Consider the following linear simple (or *univariate*) regression model

$$y_i = \alpha + \beta x_i + u_i, \tag{1}$$

where y_i is the dependent variable, x_i is the explanatory variable, u_i is the residual (also known as error or disturbance) of the regression corresponding to individual i, $i = 1, 2, ..., N$ (N being the sample size). The purpose is to estimate coefficients α (the intercept) and β.

- In case of *multiple* (or *multivariate*) regression, the model includes K explanatory variables x_1, x_2,...,x_K:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_K x_{Ki} + u_i, \tag{2}$$

or equivalently

$$y_i = \alpha + \sum_{k=1}^{K} \beta_k x_{k,i} + u_i, \tag{3}$$

- *In most part of the course, I use the univariate regression model, but it is straightforward to generalize it to the multivariate case.*

Example

Explaining individual wages: wage depends on the education level attained by the individual:

$$wage_i = \alpha + \beta educ_i + u_i. \tag{4}$$

Example

The Capital Asset Pricing Model (CAPM) states that expected returns on individual assets are linearly related to the expected return on the market portfolio. The CAPM for asset j is

$$\underbrace{r_{jt} - r_f}_{\text{excess asset return}} = \beta_j \underbrace{(r_{mt} - r_f)}_{\text{excess market return}} + u_{jt}. \tag{5}$$

Remark

Expected value: *the expected value of a random variable X, denoted as $E(X)$, is a weighted average of all possible values of X. If X is discrete,*

$$\mu \equiv E(X) = \sum_{i=1}^{N} x_i f(x_i).$$

If X is continuous,

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx.$$

The sample analog (or estimator) of expectation corresponds to the sample mean:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

If $\{x_i\}_{i=1}^{N}$ are classified following M modalities (categories), i.e. $\{x_m\}_{i=1}^{M}$ $(M < N)$, then

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{M} p(x_m) x_m$$

Remark

Conditional mean (conditional expectation):

$$E(Y|X) = \int_{-\infty}^{+\infty} yf(y|x)dy. \tag{6}$$

*If Y is **conditional mean independent** of X, it means $E(Y|X) = 0$.*

Moreover, $E(Y|X) = 0$ implies that $E(Yg(X)) = 0$ for any function g.

In particular, when g is the identity function

$$\begin{aligned} E(Y|X) &= 0 \text{ (**conditional mean independence**)} \\ \Rightarrow E(YX) &= 0 \text{ (**zero correlation**).} \end{aligned}$$

Remark

Variance: *The variance is the expected distance from X to its mean:*

$$V(X) = E\left[(X - E(X))^2\right]$$

The sample analog is

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}\left[(x_i - \bar{x})^2\right]$$

Remark

Covariance: *The covariance between X and Y is:*

$$Cov(X, Y) = E\left[(X - E(X))(Y - E(Y))\right]$$

The sample analog is

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}\left[(x_i - \bar{x})(y_i - \bar{y})\right].$$

If $E(Y)$ or $E(X) = 0$,

$$Cov(X, Y) = E(XY). \tag{7}$$

Chapter 1

Linear Models

Estimation

- The linear model

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2,i} + ... + \beta_K x_{K,i} + u_i \tag{8}$$

- The Ordinary Least Squares (OLS) criterion for the linear model is

$$\min_{\alpha,\beta_1,...,\beta_K} \sum_{i=1}^{N} u_i^2. \tag{9}$$

- For the univariate case, $y_i = \alpha + \beta x_i + u_i$, the OLS estimator is

$$\hat{\beta}_{OLS} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})y_i}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \quad \hat{\alpha}_{OLS} = \bar{y} - \bar{x}\hat{\beta}_{OLS} \tag{10}$$

  where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$.
- For the multivariate case, the OLS estimator is

$$\hat{\theta}_{OLS} = \left(\sum_{i=1}^{N} w_i w_i'\right)^{-1} \sum_{i=1}^{N} w_i y_i \tag{11}$$

  where $w_i \equiv (1, x_{1i}, x_{2i}, ..., x_{Ki})'$, $\theta \equiv (\alpha, \beta_1, \beta_2, ..., \beta_K)'$.
- Use Stata command `regress` for OLS regression.

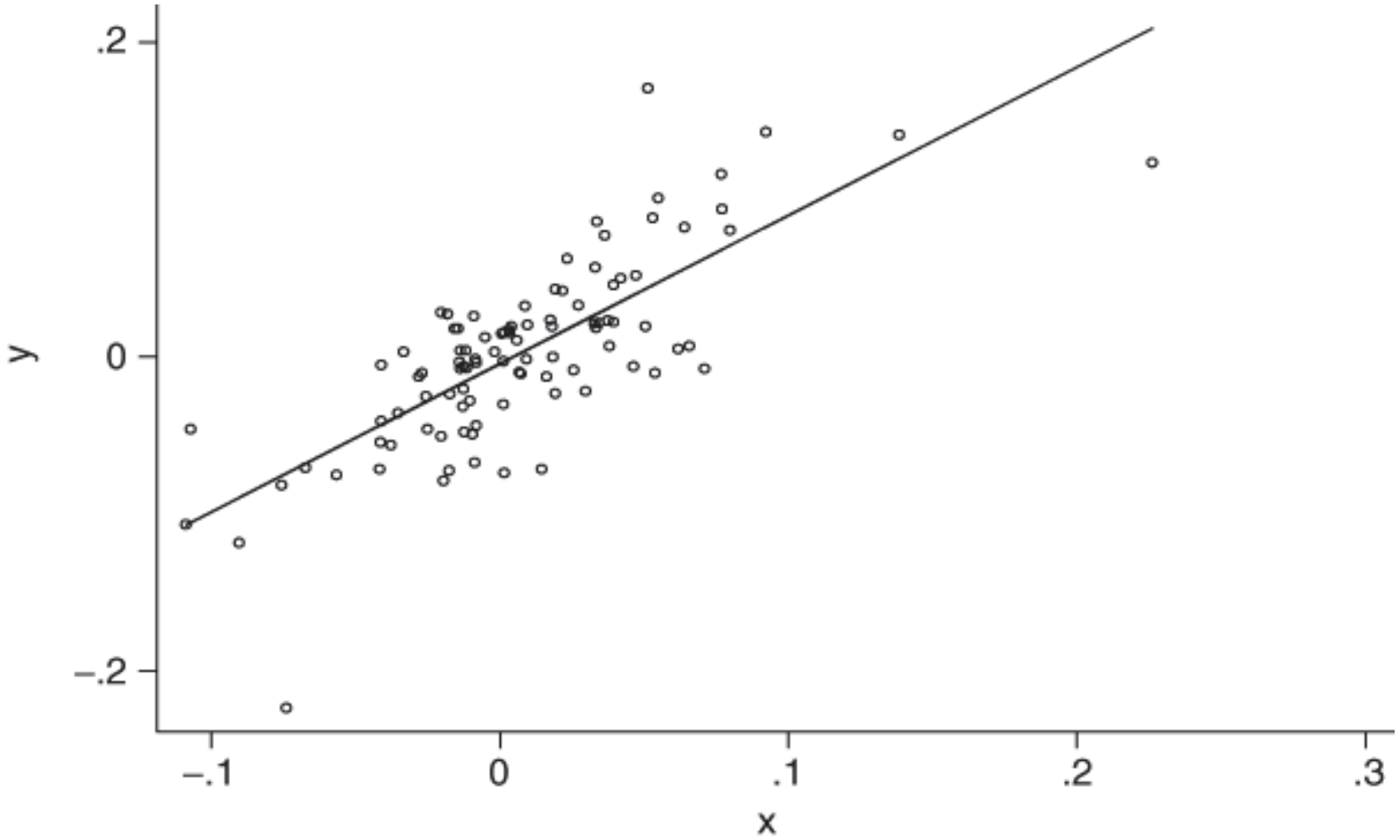


Figure: Simple linear regression: fitted line and observation points. Source: Verbeek (2008).

Assumptions of the OLS

The OLS estimator is the best linear unbiased estimator (BLUE) if the following assumptions are satisfied:

Assumption

$$\textit{Zero mean:} \quad E(u_i) = 0 \tag{A1}$$

Assumption

$$\textit{Independence:} \quad E(u_i|x_i) = 0 \tag{A2}$$

Assumption

$$\textit{Homoscedasticity (constant variance):} \quad V(u_i) = \sigma^2 \tag{A3}$$

Assumption

$$\textit{Non autocorrelation:} \quad Cov(u_i, u_j) = 0, \quad i \neq j \tag{A4}$$

Goodness-of-fit: How well does the estimated regression line fit the data?

- *Total Sum of Squares* (TSS) = *Explained Sum of Squares* (ESS) + *Residual Sum of Squares* (RSS)
- R^2: the proportion (or the percentage) of the (sample) variance of y (TSS) explained by the model (ESS).

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{12}$$

and when the model includes an intercept term (α),

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{N}\hat{u}_i^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{13}$$

where $\hat{y}_i = \hat{\alpha} + x_i'\hat{\beta}$ and $\hat{u}_i = y_i - \hat{y}_i$.

- As R^2 will never decrease if the number of regressors (K) is increased, one can use the **adjusted** R^2:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-K-1} = 1 - \frac{\frac{1}{N-K-1}\sum_{i=1}^{N}\hat{u}_i^2}{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{14}$$

Hypothesis testing

- When testing a hypothesis (called the **null hypothesis**, H_0), ones can commit two types of error:
  - **Type I error**: H_0 is rejected while it is actually true (i.e. rejecting a true H_0)
  - **Type II error**: H_0 is not rejected while it is false (i.e. keeping a false H_0)
- The probability of a type I error, α, (referred as **significance level** or **size** of the test) is controlled by the researcher and conventionally fixed equal to $\alpha = 5\%$.
- The probability of a type II error, β, is not controlled by the researcher. The **power** of the test, $1 - \beta$, is the probability of rejecting H_0 when it is false.
- *p*-**value** is the marginal significance level for which H_0 would still be rejected. If *p*-value is smaller than the significance level (say 5%), then H_0 is rejected.

Test for significance

- The aim is to test the null hypothesis, H_0: $\beta_k = \beta_k^0$. Note that H_0: $\beta_k = 0$ corresponds to testing the significance of β_k.
- The statistic of the ***t*-test** is

$$t = \frac{\hat{\beta}_k - \beta_k^0}{\sqrt{V(\hat{\beta}_k)}}. \tag{15}$$

  It follows, under H_0, a Student distribution with $N - K - 1$ degrees of freedom (maybe approximated by the standard normal distribution).
- For the **two-sided test** (or two-tailed test), i.e. the null hypothesis H_0: $\beta_k = \beta_k^0$ against the alternative hypothesis H_1: $\beta_k \neq \beta_k^0$, the null is rejected at the 5% level if

$$t < -1.96 \quad \text{or} \quad t > 1.96 \tag{16}$$

- For the **one-sided test** (or one-tailed test), H_0: $\beta_k = \beta_k^0$ against H_1: $\beta_k > \beta_k^0$, the null is rejected at the 5% level if $t > 1.64$.
- Conversely, when testing H_0: $\beta_k = \beta_k^0$ against H_1: $\beta_k < \beta_k^0$, the null is rejected at the 5% level if $t < -1.64$.

Testing linear restrictions (test for joint significance)

- One can test J linear restrictions (i.e. H_0: $R\theta = q$ where R is a $J \times (K+1)$ matrix and q is a $J \times 1$ vector)
- One uses ***F*-test** which follows, under the null, an F distribution with J and $N - K - 1$ degrees of freedom.

Example

We want to test 2 linear restrictions: $\beta_1 + ... + \beta_K = 1$ and $\beta_1 = \beta_2$. Hence, the statistic F has under H_0 an F distribution with 2 and $N - K - 1$ degrees of freedom.

Test for overall significance (or model's significance)

- The model's significance corresponds to the significance of all regressors, H_0: $\beta_1 = \beta_2 = ... = \beta_K = 0$.
- The F statistic is given by

$$F = \frac{R^2/K}{(1-R^2)/(N-K-1)} \tag{17}$$

and has an F distribution with K and $N-K-1$ degrees of freedom.

Standardized coefficients and relative importance

- When explanatory variables are measured using the same metric, the slope coefficients (β_k) are comparable $\Rightarrow$ Ones can identify the relative importance of each of the explanatory variables.
- If explanatory variables are expressed in different measurement units, using standardized coefficients (denoted $\hat{b}_k$) helps determine their relative importance:

$$\hat{b}_k = \hat{\beta}_k \frac{\hat{\sigma}_{x_k}}{\hat{\sigma}_y} \tag{18}$$

- These standardized coefficients result from the linear regression of standardized dependent variable on standardized explanatory variables (without the intercept): i.e.

$$\frac{y_i - \bar{y}}{\sigma_y} = \frac{x_{1,i} - \bar{x}_1}{\sigma_{x_1}} b_1 + ... + \frac{x_{K,i} - \bar{x}_K}{\sigma_{x_K}} b_K + \varepsilon_i \tag{19}$$

- Interpretation: the mean-y changes by an amount of $\hat{b}_k$ standard deviations for a standard deviation increase in x_k, having controlled for other variables.

Standardized coefficients and relative importance

- Stata command: `regress ... , beta`
- Example on house size

```
. regress size hhinc hhsize east owner, beta noheader
```

size	Coef.	Std. Err.	t	P>|t|	Beta
hhinc	.0046358	.0001956	23.70	0.000	.2692651
hhsize	83.56691	4.396585	19.01	0.000	.2115789
east	-106.6267	10.88597	-9.79	0.000	-.1001275
owner	366.1249	9.889078	37.02	0.000	.3972494
_cons	550.58	12.39905	44.41	0.000	.

Figure: Slope coefficients and standardized coefficients.

Margins plot

- Stata command: `margins` and `marginsplot`

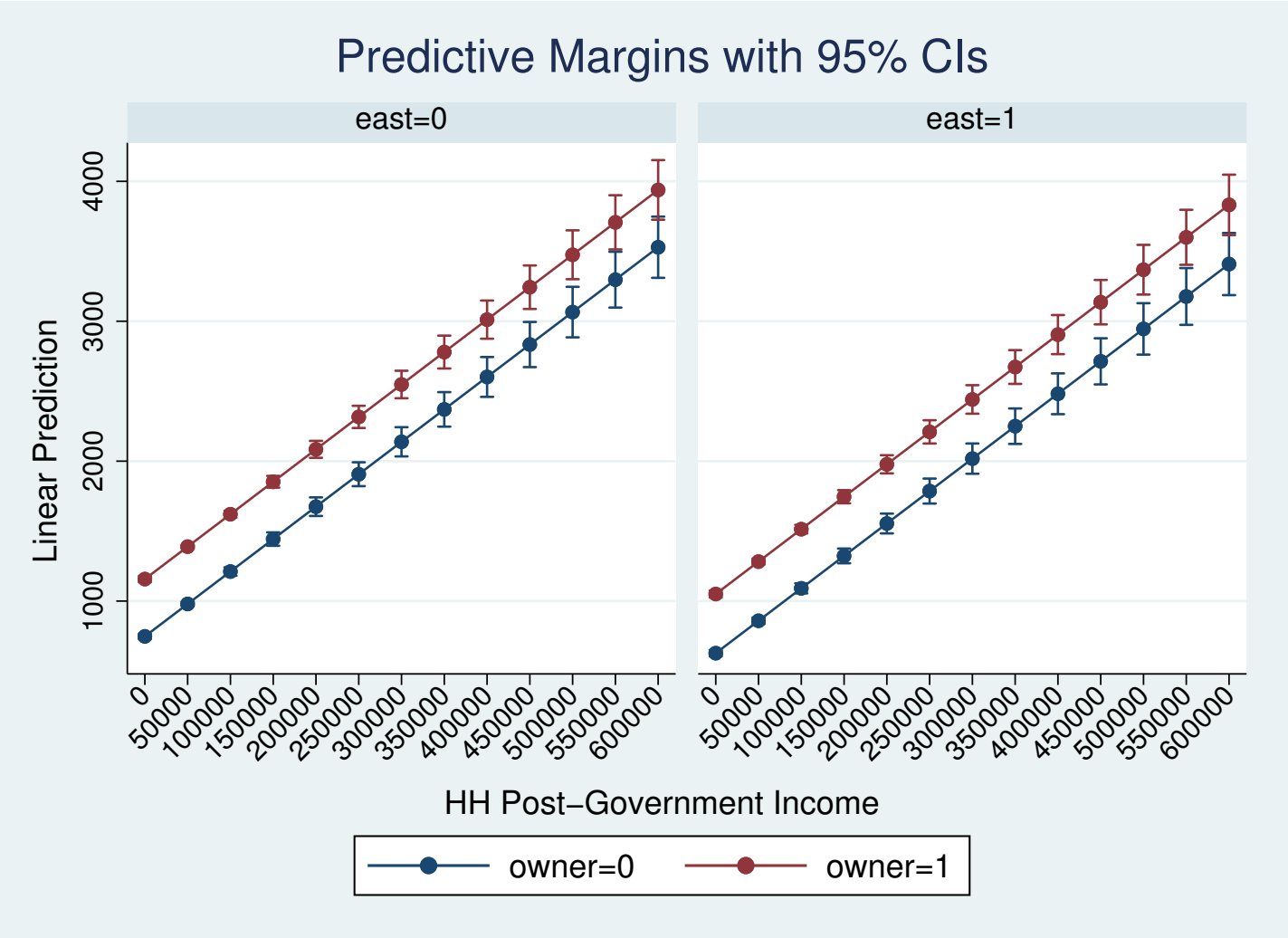


Figure: Coefficients and Confidence Intervals.

Regressions with dummy variables and interaction terms

- Consider the following model with a *dummy*

$$y_i = \alpha + \beta_1 x_i + \beta_2 D_i + \varepsilon_i \tag{20}$$

  where D_i is a dummy variable ($D_i = 1$ if i belongs to some category, =0 otherwise). For example, D indicates gender (=1 if female, 0 if male).

- This model is equivalent to

$$\begin{aligned} y_i &= \alpha + \beta_1 x_i + u_i && \text{for males, when } D_i = 0 && (21) \\ y_i &= (\alpha + \beta_2) + \beta_1 x_i + u_i && \text{for females, when } D_i = 1 && (22) \end{aligned}$$

Regressions with dummy variables and interaction terms

- Regression with a dummy variable graphically visualized:

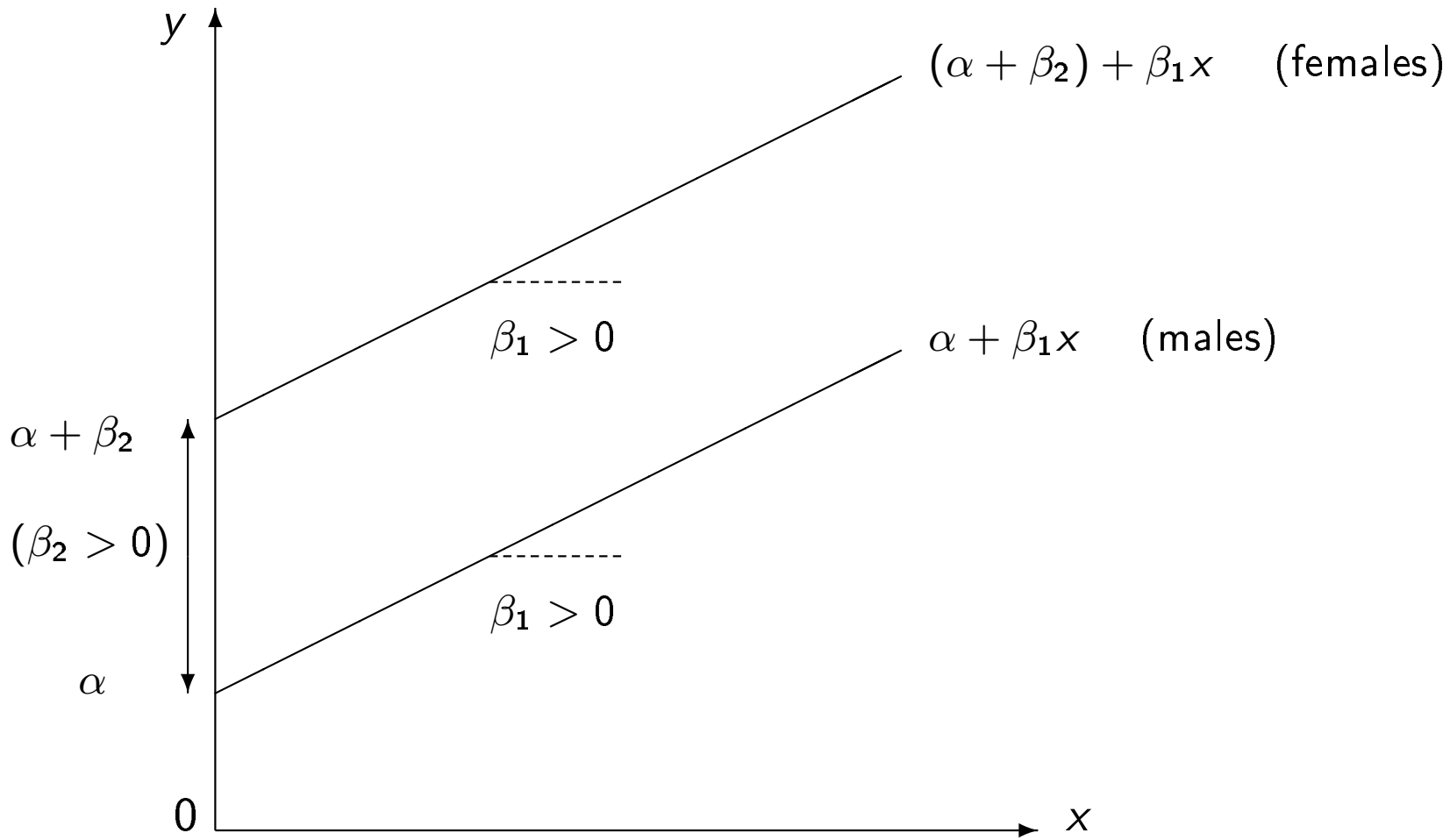


Figure: Linear regression with a dummy variable.

Regressions with dummy variables and interaction terms

- Consider the following model with *dummy* D_i and the *interaction between* D_i *and* x_i:

$$y_i = \alpha + \beta_1 x_i + \beta_2 D_i + \beta_3 D_i * x_i + \varepsilon_i \tag{23}$$

- This model is equivalent to

$$y_i = \alpha + \beta_1 x_i + u_i \quad \text{for males, when } D_i = 0 \tag{24}$$

$$y_i = (\alpha + \beta_2) + (\beta_1 + \beta_3) x_i + u_i \quad \text{for females, when } D_i = 1 \tag{25}$$

Regressions with dummy variables and interaction terms

- Regression with a dummy variable and an interaction term graphically visualized:

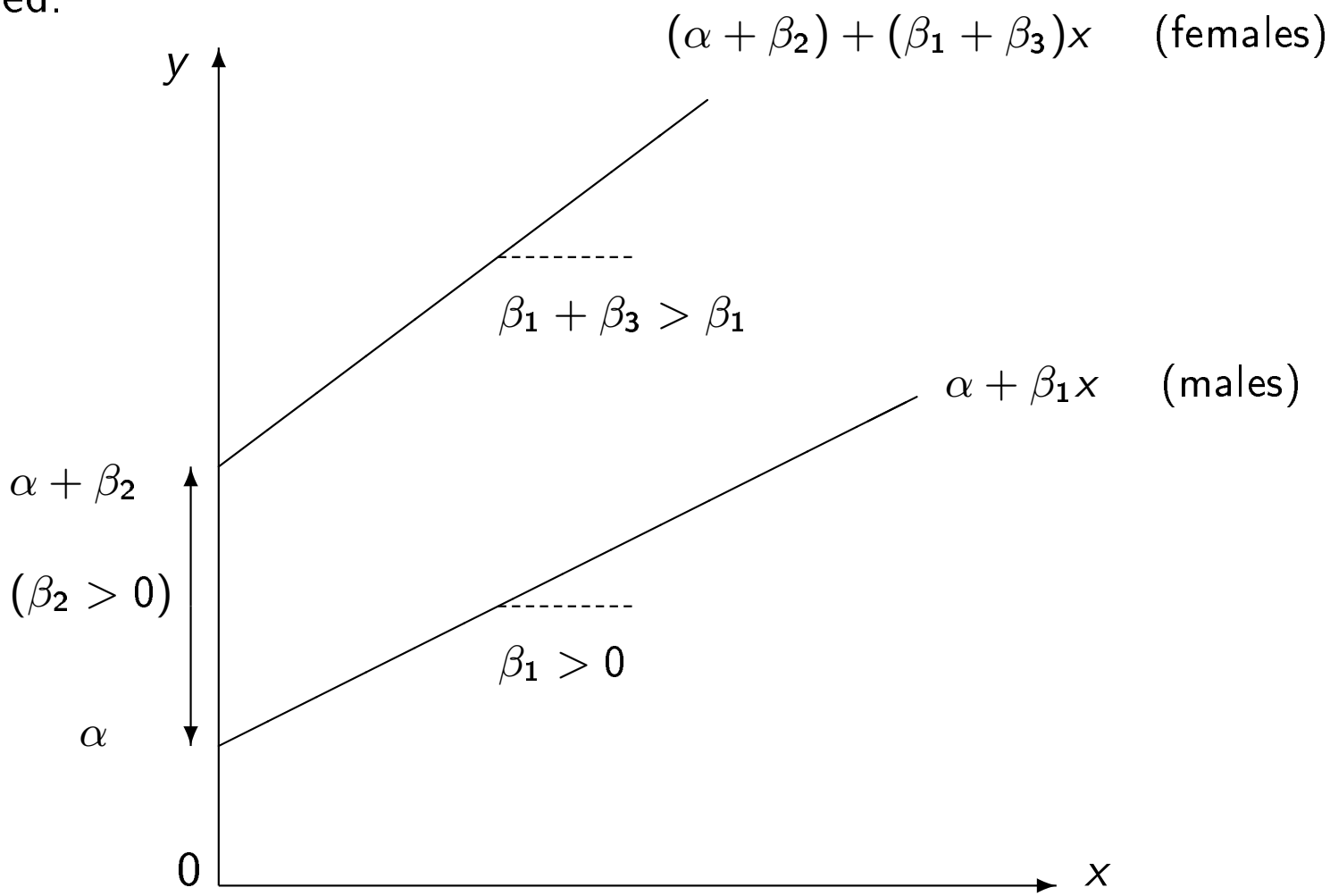


Figure: Linear regression with dummy and interaction term.

Regressions with dummy variables and interaction terms

Example

Consider the following hypothetical data about the impact of an optimism therapy (OT) on individuals' optimism. Participants to this research are split into two groups: individuals receiving the optimism therapy (treated group) and individuals receiving nothing (control group). The initial depression of participants (*depscore*) is measured by the simple depression scale (SDS) that can range from 0 to 60. The optimism score at the end of the study (*opt*) range from 0 to 100. The research question is: "Does the effect of optimism therapy (compared with the control group) depend on the level of depression?"
The regression model is

$$opt_i = \alpha + \beta_1 depscore_i + \beta_2 treat_i + \beta_3 treat_i \times depscore_i + \varepsilon_i \tag{26}$$

where *treat* is the dummy variable that indicates the treatment status (= 1 if individual receives optimism therapy, 0 otherwise).

Regressions with dummy variables and interaction terms

Example

(*continued*) The OLS results are

Table: Impact of optimism therapy on optimism score, OLS results.

Variables	Coef.	Std.err.
OT	14.690*	1.831
depscore	-0.268*	0.057
$OT \times$ *depscore*	-0.392*	0.077
intercept	48.448*	1.293

*Notes. Data source: Mitchell (2015). * indicate the significance at the 5% level.*

Regressions with dummy variables and interaction terms

Example

(*end*) The OLS results are obtained using `margins treat, at(depscore=(0 40))` and `marginsplot`, the effect of the treatment (*OT*) is displayed as follows.

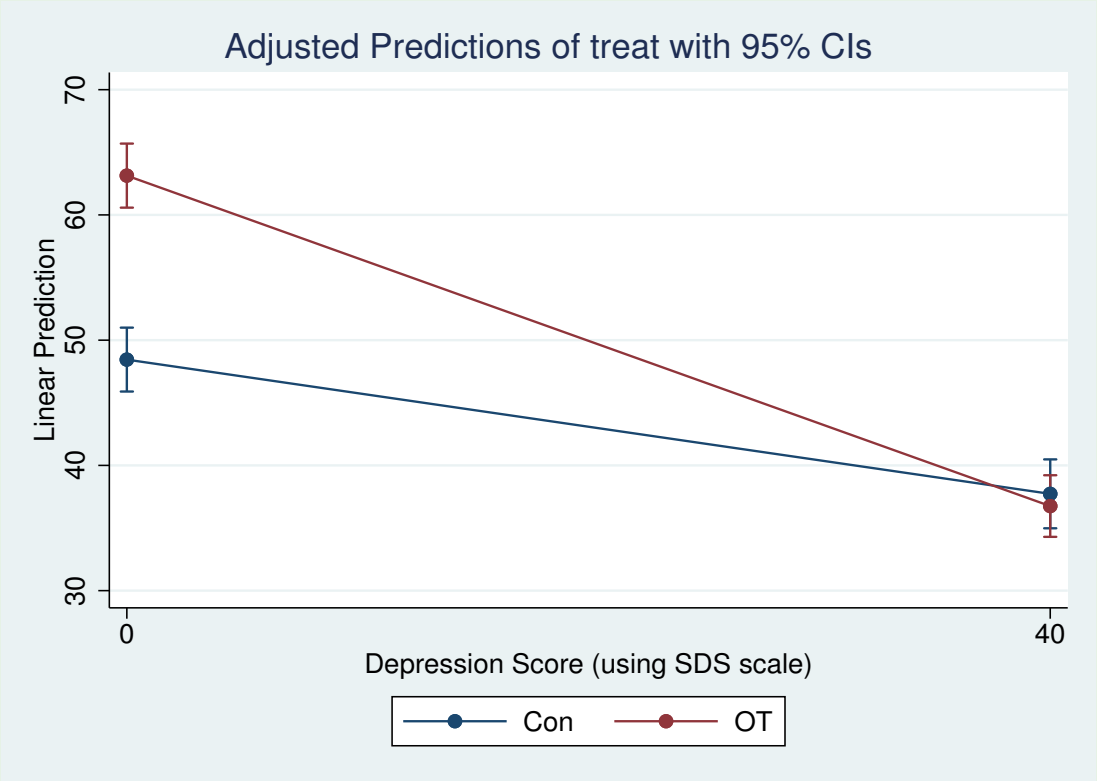


Figure: Predicted means of optimism by depression score and treatment group.

Functional forms

- Coefficients β of the linear model $y_i = \alpha + \beta_1 x_{1,i} + ... + \beta_K x_{K,i} + u_i$, is generally interpreted as marginal effects of x on y:

$$\frac{\partial y_i}{\partial x_{ki}} = \beta_k. \tag{27}$$

  Hence, β_k represents the change in in y when x_k varies by 1 unit.

- When a **loglinear model** is used,

$$\ln y_i = \alpha + \beta_1 \ln x_{1,i} + ... + \beta_K \ln x_{K,i} + u_i, \tag{28}$$

  β represent the elasticities of y with respect to x:

$$\frac{\partial \ln y_i}{\partial \ln x_{ki}} = \frac{\partial y_i}{\partial x_{ki}} \frac{x_{ki}}{y_i} = \beta_k \tag{29}$$

  Hence, β_k represents the relative change (measured in %) in y when x_k varies by 1%.

Functional forms

Example

The Cobb-Douglas production function. Consider the following stochastic Cobb-Douglas production function

$$Y_i = AK_i^{\alpha} L_i^{\beta} e^{\varepsilon_i} \tag{30}$$

where Y_i, K_i, L_i are respectively output, capital stock, and labor of firm i. A is the technological level (or productivity) and ε is a random variable.
Taking log to both sides of the equation gives the log linear model (note that $a \equiv \ln A$):

$$\ln Y_i = a + \alpha \ln K_i + \beta \ln L_i + \varepsilon_i \tag{31}$$

Chapter 2

Regression Diagnostics

Outliers

- Outlier: an observation that does not fit the trend shown by the data
- Outliers may correspond to
  - Erroneous data: data should be corrected
  - Violation of the model assumptions: an alternative model should be considered
  - Unusual values occurred by chance: they should be retained and analyzed accordingly
- Diagnosis:
  - Graphical analysis scatter plot (*pre-estimation*)
  - Compute the *leverage of observation* h_i to identify influential (or high leverage) observations:

$$h_i = \frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum_{i=1} N(x_i - \bar{x})^2} \tag{32}$$

    If $h_i > 3(K+1)/N$: influential observation.
  - Use the Cook's distance measure

$$D_i = \frac{(y_i - \hat{y}_i)^2 h_i}{(K+1)\hat{\sigma}^2(1-h_i)^2} \tag{33}$$

    where $\hat{\sigma}^2 = \frac{1}{N-1}\sum_i u_i^2$. If $D_i > 1$ (or $D_i > 4/N$): influential observation.
  - Plot leverages against residuals squared.

Outliers

- Use the following Stata post-estimation commands to compute the leverage and the Cook's distance, and plot the leverage values:

```
predict , leverage
predict , cooksd
lvr2plot
```

Example

Consider the example on return to education (4) above.

Variable	Obs	Mean	Std. Dev.	Min	Max
leverage	935	.0053476	.0046327	.0017298	.038133
cooksd	935	.0010866	.0027325	6.93e-16	.0520885

We have $3(K+1)/N = .01604278$ and $4/N = .00427807$.

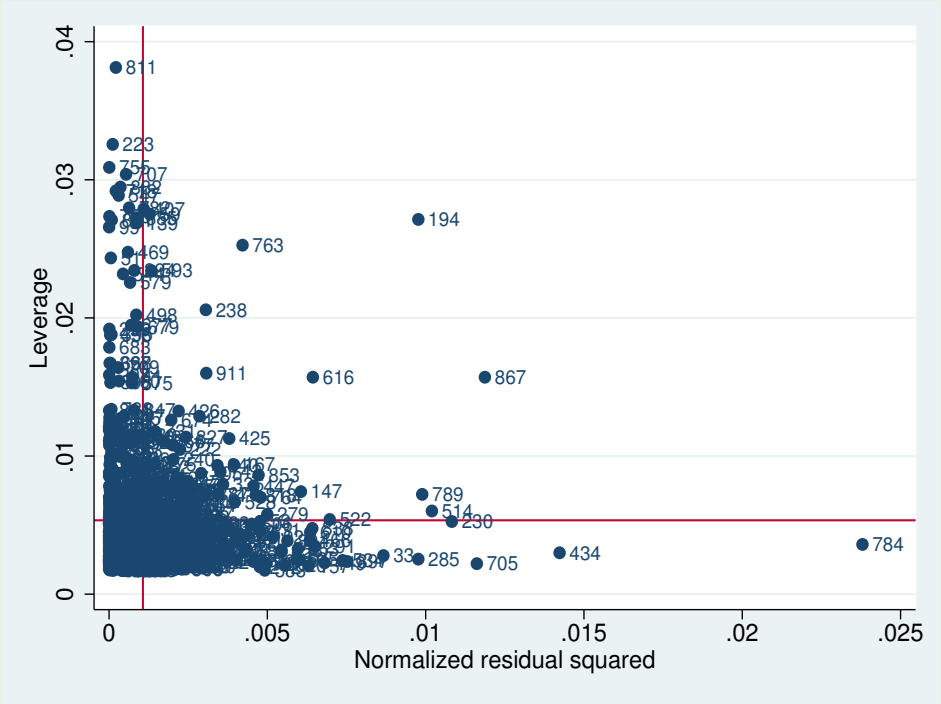


Figure: Leverages against residuals squared.

Multicolinearity

- Correlations between explanatory variables are generally not harmful.
- However, a too high correlation between explanatory variables may result in multicolinearity problems (the OLS estimator is not uniquely defined). Examples: too many dummy variables, regressions with the linear, squared and cubic terms of a variable, etc.
- Solutions:
  - Excluding one of these variables
  - Imposing some a priori restrictions on the vector of parameters
  - Extending the sample size.

Diagnosis for multicolinearity

- Correlation coefficients
- Variance inflation factor (VIF)

$$VIF(x_k) = \frac{1}{1 - R_k^2} \tag{34}$$

where R_k^2 is the coefficient of determination of x_k on all remaining explanatory variables in the model.

- $VIF(x_k) \to 1$ if $R_k^2 \to 0$. $VIF(x_k)$ very large if $R_k^2 \to 1$.
- If $VIF > 10$: multicolinearity.
- Use `correlate` to calculate the pairwise correlation coefficients.
- Use the post-estimation command `estat vif` to identify multicolinearity.

Example

Consider the estimation of the model on return to education on a sample of 663 obs.:

$$\ln wage = \alpha + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 sibs + u_i$$

where ln *wage*, *educ*, *exper*, and *sibs* represent respectively the log of the monthly wage, years of education, years of experience, and number of siblings.

Example

(*continued*) The correlation matrix is

	educ	exper	exper2	sibs
educ	1.0000			
exper	-0.4556	1.0000		
exper2	-0.4670	0.9761	1.0000	
sibs	-0.2393	0.0643	0.0724	1.0000

The VIF coefficients for the model are

Variable	VIF	1/VIF
exper2	21.48	0.046550
exper	21.21	0.047149
educ	1.35	0.739464
sibs	1.06	0.939929
Mean VIF	11.28	

Specification

Selecting the set of regressors

- The set of potentially relevant variables (included in x) should be chosen on the basis of *economic arguments* rather than statistical ones. Using *statistical arguments* (e.g., including or not an explanatory variable by using the t-test), referred to as **data snooping/data mining**, would encounter a high probability of Type I errors due to the accumulation of several tests.
- A good strategy is the **general-to-specific approach**, known as the **LSE methodology**.
- In practice, most researchers will start somewhere "in the middle" and then test (i) whether restrictions imposed by the model are correct (autocorrelation, heteroscedasticity, etc.) and (ii) whether restrictions not imposed by the model could be included (parametric restrictions such as zero coefficients for some variables, functional forms, etc.).

- Some recommendations:
  - It is not useless to have insignificant variables included in your specification: An insignificant variable is also a result!
  - Keep the intercept term even if it is insignificant!
- Some criteria for selecting the set of regressors:
  - Use $\bar{R}^2$
  - **Akaike's Information Criterion** (AIC):

$$AIC = \ln \frac{1}{N} \sum_{i=1}^{N} \hat{u}_i^2 + \frac{2(K+1)}{N}. \tag{35}$$

  - **Schwarz Bayesian Information Criterion** (BIC):

$$BIC = \ln \frac{1}{N} \sum_{i=1}^{N} \hat{u}_i^2 + \frac{(K+1)}{N} \ln N. \tag{36}$$

  - Use the F-test for testing parametric restrictions (e.g., zero coefficients for some variables).

Specification

Functional forms

- Use some graphics to detect the form of the relation between variables
- Use log transformation to reduce dispersion ($\ln x_i$)
- Use polynomial terms, such as x, x^2, x^3, etc. for nonlinear functions:

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + u_i \tag{37}$$

- Use interaction terms:

$$y_i = \alpha + \beta_1 x_i + \beta_2 z_i + \beta_3 (x_i * z_i) + u_i \tag{38}$$

- Use dummy variables, etc.

Specification

Structural break

- The original model (including the intercept)

$$y_i = \alpha + \beta_1 x_{1,i} + ... + \beta_K x_{K,i} + u_i \tag{39}$$

- Coefficients can be different across two (or more) subsamples. The general model is

$$y_i = \alpha + \beta_1 x_{1,i} + ... + \beta_K x_{K,i} + G_i \left(\alpha^{S2} + \beta_1^{S2} x_{1,i} + ... + \beta_K^2 x_{K,i}\right) + u_i \tag{40}$$

  where $G_i = 0$ if individual i belongs to the first subsample and $G_i = 1$ if i belongs to the second subsample.

- This model is equivalent to

$$\begin{aligned} y_i &= a^{S1} + b_1^{S1} x_{1,i} + ... + b_K^{S1} x_{K,i} + u_i \quad \text{for subsample 1} & (41)\\ y_i &= a^{S2} + b_1^{S2} x_{1,i} + ... + b_K^{S2} x_{K,i} + u_i \quad \text{for subsample 2} & (42) \end{aligned}$$

  with $a^{S1} = \alpha$, $b_k^{S1} = \beta_k$ and $a^{S2} = \alpha + \alpha^{S2}$, $b_k^{S2} = \beta_k + \beta_k^{S2}$, for all k.

Specification

Structural break

- **Chow test** for structural break: H_0: $\alpha^{S2} = \beta_1^{S2} = ...\beta_K^{S2} = 0$ (or equivalently, $a^{S1} = a^{S2}$ and $b_k^{S1} = b_k^{S2}$, for all k). This is also an F-test:

$$F = \frac{(RSS_R - RSS)/(K+1)}{RSS/(N-2(K+1))}. \tag{43}$$

  $K+1$ = number of parameters (including the intercept), i.e. number of components in w in the restricted model (model under H_0). RSS = residual sums of squares of the unrestricted model (equation 40), and RSS_R is the residual sums of squares of the restricted model (equation 39).
- This statistic has an F distribution with $K+1$ and $N-2(K+1)$ degrees of freedom.

What happens when assumption A1 is violated ?

- Assumption A1 ($E(u_i) = 0$) required that the average value of the regression errors is zero.
- If the regression includes a constant (i.e. α), this assumption will never be violated.
- If the regression does not include a constant, R^2 may be negative, and estimates of other coefficients (β_1, β_2, ...) may be biased.

  $\Rightarrow$ Always include the intercept (unless otherwise indicated)!

Example

Consider the Capital Asset Pricing Model (CAPM) for the food industry.

Table: OLS results, food industry. Dependent variable: excess industry portfolio returns.

	with intercept		without intercept	
Variables	Coef.	Std.err.	Coef.	Std.err
excess market return	0.774**	0.028	0.781**	0.028
intercept	0.294**	0.123	–	–
$\bar{R}^2$	0.576		0.581	
$F(1, 562)$	765.25			
Chow test, 2003 break, $F(2, 560)$	2.34			
Number of observations	564		564	

*Notes. Data source: Verbeek (2008). * and ** indicate the significance at the 5% and 1% levels respectively.*

What happens when assumption A3 is violated ? Heteroscedasticity

- Assumption A3 means that the model is homoscedastic.
- We plot the estimated residuals $\hat{u}_i = y_i - \hat{\alpha} - \hat{\beta}x_i$ against x_i or against the fitted values ($\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$) in order to detect heteroscedasticity.
- Tests for heteroscedasticity: Breusch-Pagan test, White's test, etc.
- **The White test** is a general test for heteroscedasticity in the residuals. The test statistic is distributed as $\chi^2(p)$ under the null hypothesis of homoscedasticity, where p is the number of regressors used in the auxiliary regression (given by Stata).
- One way to reduce heteroscedasticity is to use variables expressed in logarithm (e.g., ln(y) instead of y).
- If the form of heteroscedasticity is known, we can use **Generalized Least Squares (GLS)** to estimate the model.
- When heteroscedasticity is suspected, we should use the **White**'s procedure which produces the estimation with **robust standard errors**.
- Use `regress , robust` for robust estimation; `bpagan` for the Breusch-Pagan test, `whitetst` for the White's test.

Example

Consider the model in the previous example (CAPM for the food industry): $foodrf_t = \alpha + \beta\, rmrf_t + u_t$. We use the Stata command `rvfplot`.

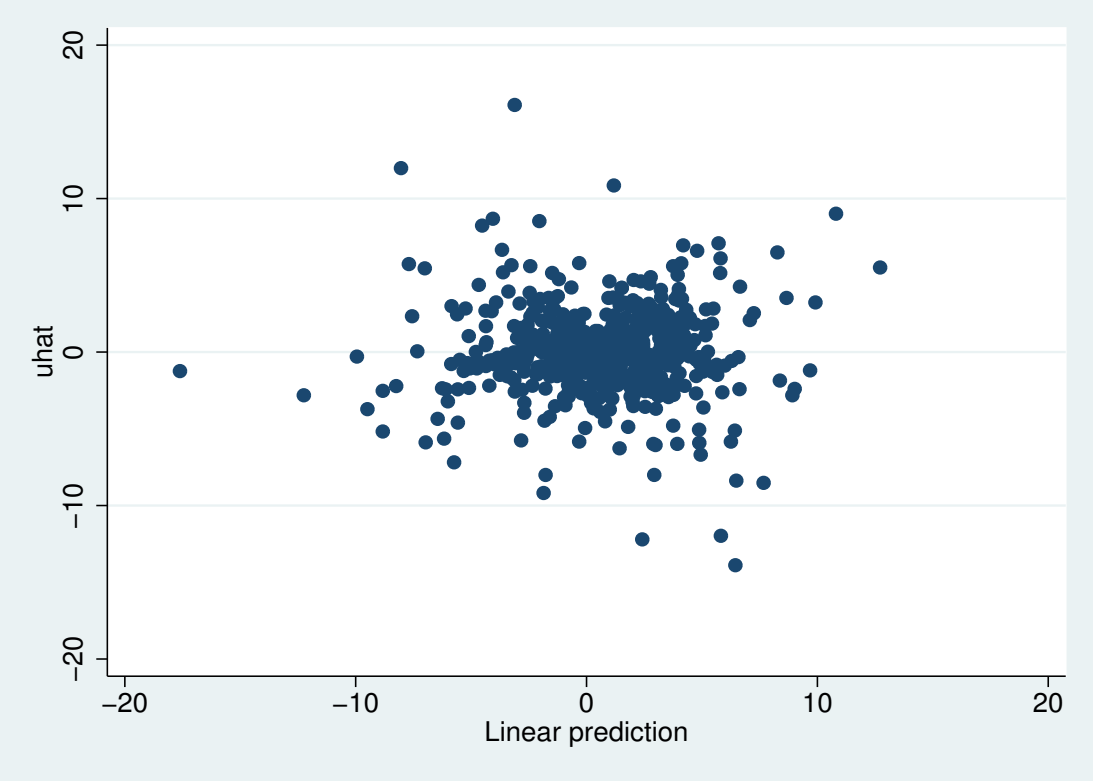


Figure: Residuals ($\hat{u}$) versus fitted values (linear prediction, $\hat{y} = \hat{\alpha} + \hat{\beta}\, rmrf$).

Result for the White test: $\chi^2(2) = 18.656$, p-value = 8.9e-05, which rejects the null hypothesis of homoscedasticity.

Example

Table: OLS results with robust standard errors, food industry. Dependent variable: excess industry portfolio returns.

Variables	Coef.	Std.err.	Robust std.err
excess market return	0.774	0.028*	0.038*
intercept	0.294	0.123*	0.122*
$\bar{R}^2$		0.576	
Number of observations		564	

*Notes. Data source: Verbeek (2008). * significance at the 5% level.*

What happens when assumption A4 is violated? Autocorrelation

- A4 means that the residuals are not autocorrelated.
- Usually, this assumption makes sense in the models with time series data.
- We plot $\hat{u}_t$ over time and $\hat{u}_t$ against $\hat{u}_{t-1}$ in order to detect the pattern of **autocorrelation** (or **serial correlation**).
- Linear model with first-order autocorrelation ($t = 1, 2, ..., T$):

$$y_t = \alpha + \beta x_t + u_t \quad (44)$$

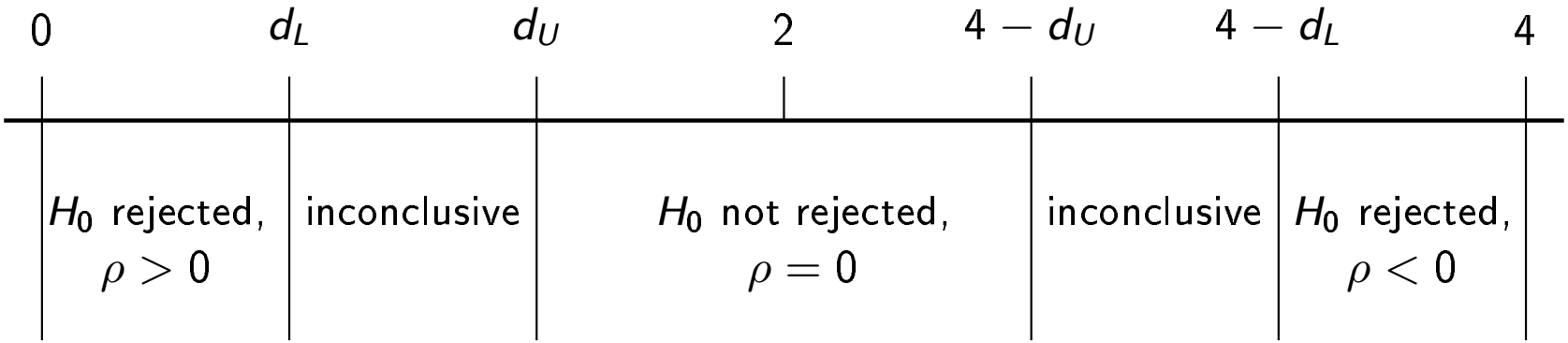
$$u_t = \rho u_{t-1} + v_t \quad (45)$$

First-order autocorrelation in (45) is also referred to as $AR(1)$.

- The **Durbin-Watson test** for first-order autocorrelation, H_0: $\rho = 0$, is

$$DW = \frac{\sum_{t=2}^{T}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T}\hat{u}_t^2} \approx 2 - 2\hat{\rho} \tag{46}$$

  where $\hat{\rho}$ is the OLS estimator of the model in (45) where u_t is replaced by $\hat{u}_t$.
- We can use the critical values (lower limit d_L and upper limit d_U) of the test (provided in many softwares and textbooks) to assess H_0.
- The alternative hypothesis can be H_1: $\rho > 0$ (positive autocorrelation) or H_1: $\rho < 0$ (negative autocorrelation). The rule is



- The model with first-order autocorrelation can be transformed as

$$y_t - \rho y_{t-1} = (1-\rho)\alpha + \beta(x_t - \rho x_{t-1}) + v_t,\ t = 2, 3, ..., T \quad (47)$$

$$\sqrt{1-\rho^2}\, y_1 = \beta\sqrt{1-\rho^2}\, x_1 + \sqrt{1-\rho^2}\, u_1 \quad (48)$$

- Estimation of this model can be performed by using GLS-type estimators such as **Cochrane-Orcutt** estimator (which does not use the first transformed observation $t = 1$) or **Prais-Winsten** estimator (which uses all transformed observations).
- When autocorrelation (even of higher-order) of unknown form is suspected, we use the **Newey-West** procedure to obtain the OLS estimator with **Heteroscedasticity and Autocorrelation Consistent (HAC) standard errors**.
- Use Stata commands `prais` for the Prais-Winsten estimator; `prais , corc` for the Cochrane-Orcutt estimator (including the Durbin-Watson test); `newey` for the Newey-West estimator.

Example

Always with the CAPM for the food industry: $foodrf_t = \alpha + \beta\, rmrf_t + u_t$.

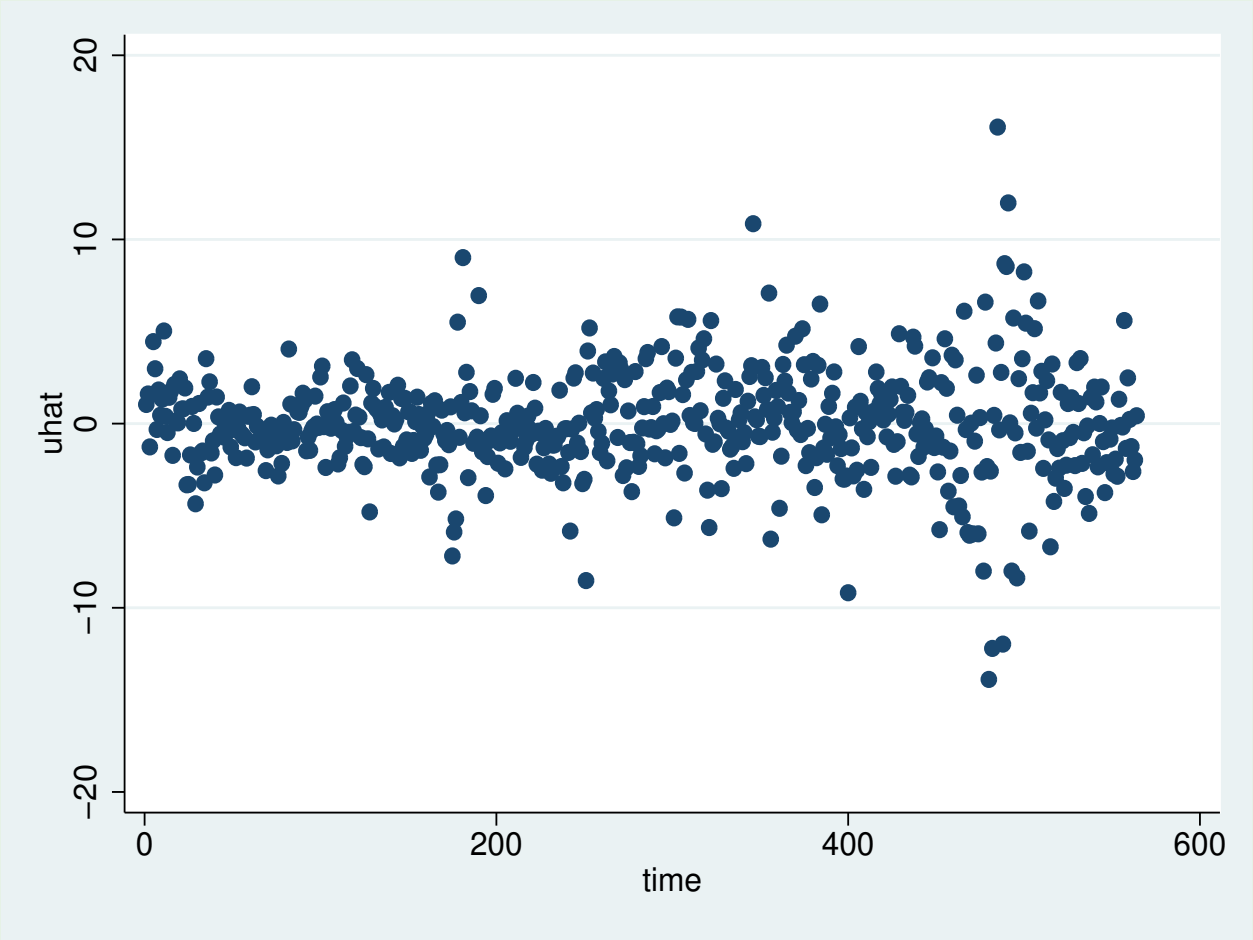


Figure: Residuals ($\hat{u}$).

Example

Table: OLS results with first-order autocorrelation, food industry. Dependent variable: excess industry portfolio returns.

	Cochrane-Orcutt		Prais-Winsten		Newey-West	
Variables	Coef.	Std.err.	Coef.	Std.err.	Coef.	Std.err.
excess market return	0.774*	0.0280	0.774*	0.0279	0.774*	0.071
intercept	0.291*	0.1361	0.295*	0.1361	0.294	0.151
$\bar{R}^2$	0.577		0.578		–	
Number of obs.	563		564		564	

*Notes. Data source: Verbeek (2008). * significance at the 5% level.*

Chapter 3

Endogeneity and Instrumental Variables

What happens when assumption A2 of OLS is violated?

- Assumption A2 (also known as *zero conditional mean*) means that the error terms u_i are uncorrelated with explanatory variables x_i. This means that x_i is *exogenous regressors*.
- A2 is crucial for the consistency of the Ordinary Least Squares (OLS) estimator.
- In particular, independence between u_i and x_i implies zero correlation between them, i.e.

$$E(u_i|x_i) = 0 \ (\textit{condtl. mean indep.}) \Rightarrow E(x_i u_i) = 0 \ (\textit{zero corr.}) \tag{49}$$

- When this assumption fails, i.e. the error terms are correlated with some or all explanatory variables (which are known as **endogenous regressors**), the OLS estimator is inconsistent and alternative estimators should be considered.
- Two alternative methods: **Instrumental Variables (IV)** and **Generalized Method of Moments (GMM)**.

Cases where OLS cannot be saved

In the following cases, assumption $A2$ is violated (i.e. residuals and explanatory variables are correlated).

1. *Autocorrelation with a lag dependent variable*. It usually arises in models with time series.

Example

$$y_t = \alpha + \beta x_t + \gamma y_{t-1} + u_t, \quad u_t = \rho u_{t-1} + v_t \tag{50}$$

2. *Measurement error in explanatory variables*. This corresponds to the situation where explanatory variables are observed with some error.

Example

$$x_i = z_i + \varepsilon_i \tag{51}$$

where ε_i represents the measurement error which follows a distribution of zero mean and constant variance σ_ε^2 (e.g., ε_i follows a normal distribution $N(0, \sigma_\varepsilon^2)$).

3. *Omitted variables*. There is an omitted explanatory variable in x_i. If this omitted variable is correlated with the remaining variables in x_i, the OLS estimator will be biased.

Example

Suppose we have the right model

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \tag{52}$$

Suppose that $Cov(x_{1i}, x_{2i}) > 0$ and x_{2i} is unobserved. Therefore, the model we estimate is

$$y_i = \alpha + \beta_1 x_{1i} + \varepsilon_i \tag{53}$$

where the new residual becomes $\varepsilon_i = \beta_2 x_{2i} + u_i$. Consequently, this residual is correlated with the regressor of the estimated model ($E(x_{1i}\varepsilon_i) \neq 0$).

4. *Simultaneity and reverse causality.*

Example

Consider the following *structural* model with per capita consumption y_t, per capita income x_t, and per capita investment z_t:

$$y_t = \alpha + \beta x_t + u_t \tag{54}$$
$$x_t = y_t + z_t \tag{55}$$

We assume that $E(u_t|z_t) = 0$. Solving the structural model gives the following *reduced-form* equations

$$x_t = \frac{\alpha}{1-\beta} + \frac{1}{1-\beta} z_t + \frac{1}{1-\beta} u_t \tag{56}$$
$$y_t = \frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta} z_t + \frac{1}{1-\beta} u_t \tag{57}$$

Equation (56) implies that x_t and u_t are correlated. As a result, the OLS applied on the model given in (54) will be inconsistent because $E(x_t u_t) \neq 0$.

Instrumental Variables Estimator

- Suppose we have the model

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \tag{58}$$

- We can applied OLS if $E(x_{i1} u_i) = E(x_{2i} u_i) = 0$.
- However, if, for example, $E(x_{2i} u_i) \neq 0$ (i.e. x_{2i} is an *endogenous regressor*), the OLS estimator is inconsistent and biased.
- Assume there exists a variable z_{2i} (called as *instrument*) such that $E(z_{2i} u_i) = 0$ (*exclusion restriction*) and that z_{2i} is correlated with x_{2i}.
- We use the **Instrumental Variables (IV) estimator**.

- The IV estimator is based on the following moment condition:

$$E(y_i - x_i'\beta)z_i = 0 \tag{59}$$

where $x' = (x_1', x_2')$ and $z' = (x_1', z_2')'$. This leads to the following minimization program

$$\min_{\beta} \left[\frac{1}{N}\sum_{i=1}^{N}(y_i - x_i'\beta)z_i\right]' W_N \left[\frac{1}{N}\sum_{i=1}^{N}(y_i - x_i'\beta)z_i\right] \tag{60}$$

- As $dim(z_2) = dim(x_2)$, the solution is

$$\beta_{IV} = \left(\sum_{i=1}^{N} z_i x_i'\right)' \sum_{i=1}^{N} z_i y_i \tag{61}$$

$$= (Z'X)^{-1}Z'y \tag{62}$$

- The IV estimator can be also implemented in a two-step procedure (**Two-Stage Least Squares (2SLS)** estimator):
  1. Regress x_{2i} on x_{1i} and z_{2i} (including a constant) by OLS. The model of this step is
  $$x_{2i} = a + b_1 x_{1i} + b_2 z_{2i} + \varepsilon_i. \tag{63}$$
  This step gives the fitted values
  $$\hat{x}_{2i} = \hat{a} + \hat{b}_1 x_{1i} + \hat{b}_2 z_{2i} \tag{64}$$
  2. Regress y_i on x_{1i} and $\hat{x}_{2i}$ (including a constant) by OLS
  $$y_i = \alpha + x_{1,i}\beta_1 + \hat{x}_{2,i}\beta_2 + u_i \tag{65}$$
  which gives the 2SLS estimators ($\hat{\alpha}_{2SLS}$, $\hat{\beta}_{1,2SLS}$, and $\hat{\beta}_{2,2SLS}$) which are the same than the direct IV estimators.
- The standard errors in the 2nd step above should be corrected in order to have a correct inference (t- or z- statistics). This is automatically done by the Stata command `ivregress 2sls`.

- When the number of instrumental variables is higher than the number of endogenous regressors (i.e. when $dim(z) > dim(x)$), we have the **Generalized Instrumental Variables (GIV) estimator**, which can be also obtained by the **2SLS** estimator based on the two-step procedure.
- Using the optimal weighting matrix, $W_N = \left(\frac{1}{N} Z'Z\right)^{-1}$. The expression of the GIV estimator (solution of equation (60)) is

$$\beta_{GIV} = (X'P_Z X)^{-1} X'P_Z y \tag{66}$$

where $P_Z = Z(Z'Z)^{-1}Z'$ (orthogonal projection on Z space).

Tests

- The IV/GIV estimator is based on quantity $\frac{1}{N}\sum_{i=1}^{N}\hat{u}_i z_i$.
- In case of exact identification ($dim(x) = dim(z)$), this quantity is 0.
- The **overidentifying restrictions test** or **Sargan test**: the null hypothesis H_0: $E(z_i u_i) = 0$.
- If the population moments were true, $\frac{1}{N}\sum_{i=1}^{N}\hat{u}_i z_i$ is expected to be close to 0, leading to the following statistic:

$$\xi = \left(\sum_{i=1}^{N}\hat{u}_i z_i\right)' \left(\hat{\sigma}^2 \sum_{i=1}^{N}\hat{u}_i z_i'\right)^{-1} \left(\sum_{i=1}^{N}\hat{u}_i z_i\right) \sim \chi^2(q), \tag{67}$$

  where $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\hat{u}_i^2$, $\hat{u}_i = y_i - x_i'\hat{\beta}_{GIV}$, and $q = dim(z) = dim(x)$.
- This test may be performed as follows:
  1. Run 2SLS to obtain the IV residuals $\hat{u}_i$.
  2. Regress $\hat{u}_i$ on the full set of exogenous regressors, i.e. x_{1i} and z_{2i}, in order to obtain R^2.
  3. The test statistic corresponds to NR^2.

- IV estimator can be biased if instruments exhibit weak correlation with endogenous regressors: **weak instruments**.
- Consider the case of a single regressor and a constant and an instrument, the IV estimator is

$$\hat{\beta}_{IV} = \frac{(1/N)\sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y})}{(1/N)\sum_{i=1}^{N}(z_i - \bar{z})(x_i - \bar{x})} \rightarrow \frac{Cov(z, y)}{Cov(z, x)} \tag{68}$$

- If the correlation between x and z is close to 0, the IV estimator is inconsistent and its asymptotic distribution differs from a normal distribution.
- Test for **weak instruments**: this test corresponds to the null hypothesis H_0: $b_2 = 0$. The test statistic may be a t-statistic or a F statistic obtained from the first stage regression (i.e. regression of x_2 on x_1 and z_2).
- Use `reg3..., 2sls` and `ivregress 2sls` for IV estimation.
- Use post-estimation commands: `estat overid` for Sargan test, `estat endogenous` to test for endogenous regressors, and `estat firststage` to test for weak instruments.

Generalized Method of Moments (GMM)

- The **GMM** approach employs the *moment* conditions imposed by the underlying model. These conditions can be linear or nonlinear in parameters.

Example

Consider the assumption about a zero correlation between the instrument z_{2i} and the residual u_i, $E(z_{2i}u_i) = 0$. This is a linear condition which corresponds to the sample moment conditions

$$\frac{1}{N}\sum_{i=1}^{N} z_{2i}u_i = 0 \tag{69}$$

By using $u_i = y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i}$, these moments becomes

$$\frac{1}{N}\sum_{i=1}^{N} z_{2i}\left(y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i}\right) = 0. \tag{70}$$

- IV, GIV, or 2SLS are particular cases of GMM estimator.
- The GMM estimator does not have an analytical form. It is obtained by numerical solution of a minimization program based on moment conditions.
- As in the GIV (or 2SLS) case, we can test for overidentifying restrictions.

Example

Consider the estimation of the model on return to education on a sample of 663 obs.:

$$\ln wage = \alpha + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 sibs + u_i \quad (71)$$

where ln *wage*, *educ*, *exper*, and *sibs* represent respectively the log of the monthly wage, years of education, years of experience, and number of siblings.
We can use variable *brthord* (birth order, = 1 for a first-born child), *feduc* (father's education), and *meduc* (mother's education) as instruments for *educ*. The first-step of 2SLS gives the following results for the *educ* equation:

Table: 2SLS results: The first step

Variable	Coefficient	Std.Err.
exper	-0.034	0.074
$exper^2$	-0.007	0.003
sibs	-0.102	0.039
brthord	-0.019	0.059
feduc	0.151	0.027
meduc	0.109	0.031
Intercept	12.685	0.555
N	663	
R^2	0.341	

Example

(*Example (continued and end)*) Estimations of the 2SLS 2nd step (or GIV) are reported below. For comparison purpose, the OLS estimator is also presented.

Table: Estimation results: OLS and 2SLS

	OLS		2SLS	
Variable	Coefficient	Std.Err.	Coefficient	Std.Err.
educ	0.074**	0.007	0.159**	0.024
exper	0.015	0.014	0.006	0.017
$exper^2$	0.000	0.001	0.002*	0.001
sibs	-0.013*	0.006	0.005	0.008
Intercept	5.614**	0.131	4.335**	0.386
R^2	0.136		0.003	

Notes: ** significant at the 1% level, * at the 5 % level.

As there are 3 instruments for 1 endogenous variable (*educ*), the Sargan test for overidentifying restrictions can be performed here. The statistic has a $\chi^2(2)$ (for 3 instruments - 1 variable). The statistic equals to 1.274, lower than the 5% critical value of $\chi^2(2)$, which does not allow to reject the null hypothesis of overidentifying restrictions. We can conclude that the specification with these instrumental variables is not rejected.

Chapter 4

Maximum Likelihood

Maximum Likelihood Estimation

- The **Maximum Likelihood (ML)** method searches values of parameters that give the highest probability (the highest likelihood) to observe the data sample.
- The ML supposes the distribution of an observed phenomenon is known even if the parameters underlying this distribution are unknown. This distribution can be summarized in the distribution of the residual

Assumption

The residual u_i is normally and independently distributed with zero mean and variance σ^2. *(A5)*

Remark

Two events A and B are said independent if $E[g_1(A)g_2(B)] = E[g_1(A)]E[g_2(B)]$ where g_1 and g_2 are arbitrary functions.

Moreover, let $P(A)$, $P(B)$, and $P(A, B)$ denote respectively the probability (or density) of A, the probability of B, and the joint probability of A and B. If A and B are independent, $P(A, B) = P(A)P(B)$.

- Let us define θ as the set of parameters to be estimated. Therefore, the ML estimator $\hat{\theta}_{ML}$ is the result of the maximization of the likelihood (or log-likelihood) function.
- The likelihood represents the probability to observe the sample, i.e. all values (x_i, y_i), $i = 1, 2, ..., N$: $\quad L(\theta) = P(y_1, y_2, ..., y_N)$.
- As observations are independent, this likelihood becomes

$$L(\theta) = P(y_1) \times ... \times P(y_N) = \prod_{i=1}^{N} P(y_i) \tag{72}$$

  where $P(y_i)$ is the probability to observe y_i.
- Usually, we maximize $\ln L(\theta)$ with respect to θ:

$$\ln L(\theta) = \ln P(y_1) + ... + \ln P(y_N) = \sum_{i=1}^{N} \ln P(y_i) \tag{73}$$

Example

Consider the model

$$y_i = \alpha + \beta x_i + u_i \tag{74}$$

under the same assumptions A1-A4 as the OLS estimator. Moreover, we assume the residual u_i has a normal distribution: $P(u_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{u_i^2}{\sigma^2}\right)$. This is by definition the density of y_i given x_i and the parameter set $\theta \equiv (\alpha, \beta, \sigma^2)'$.

Using $u_i = y_i - \alpha - \beta x_i$ gives the density of y_i: $P(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y_i-\alpha-\beta x_i)^2}{\sigma^2}\right)$.
The probability (or likelihood) to observe the sample – i.e. all values (x_i, y_i), $i = 1, 2, ..., N$ – is $L \equiv P(y_1, y_2, ..., y_N)$. As observations are independent, this likelihood becomes

$$L(\theta) = \prod_{i=1}^{N} P(y_i) \Rightarrow \ln L = \sum_{i=1}^{N} \ln P(y_i) \tag{75}$$

The ML estimator $\hat{\theta}_{ML}$ results from maximization of $\ln L$ with respect to θ.

- The ML estimator, $\hat{\theta}_{ML}$, is derived from the maximization of the log-likelihood function ($\ln L$) with respect to θ. It does not often have an analytical form.
- The ML estimator is consistent and asymptotically efficient (i.e. it has the smallest variance among all consistent estimators)
- In the linear case, as in the example above ($y_i = \alpha + \beta x_i + u_i$), the ML estimator coincides with the OLS estimator.
- The ML estimator is useful in the nonlinear case (Chapter 2), where the OLS estimator does not work.

Specification Tests based on Maximum Likelihood

There are 3 main likelihood-based tests: Wald test, likelihood ratio test, and Lagrange multiplier test.
Assume we are interested in testing J linear restrictions on the parameter set θ (recall that $\theta \equiv (\alpha, \beta_1, ..., \beta_K)$). They are summarized in the null hypothesis H_0. Under H_0, each of three tests has a Chi-squared distribution of q degrees of freedom $\chi^2(q)$. The principle of these tests is sketched as follows:

- **Wald test**. We estimate θ by ML and check whether the restrictions in H_0 are fulfilled. The idea is close to that of the t and F tests.
- **Likelihood ratio test**. We estimate the model twice - once without the restrictions imposed (giving $\hat{\theta}_{ML}$) and once with these restrictions (giving $\hat{\theta}_{0,ML}$) and check whether the difference in log-likelihood values, $\ln L(\hat{\theta}_{ML}) - \ln L(\hat{\theta}_{0,ML})$, is significantly different from zero.
- **Lagrange multiplier test**. We estimate the model with the restrictions from the null hypothesis (giving $\hat{\theta}_{0,ML}$) and check whether $\hat{\theta}_{0,ML}$ satisfies the general model (i.e. the model without these restrictions).

Chapter 5

Limited Dependent Variable Models

What is a limited dependent variable ?

- A limited dependent variable (LDV) is broadly defined as a dependent variable whose range of values is substantively restricted.
- A binary variable takes on only two values, 0 and 1. Models related to this kind of dependent variable are *probit* and *logit*.

Example

We are interested in the fact that a family possesses a car or not. The dependent variable is defined as follows

$$\begin{aligned} y_i &= 1 \quad \text{if family } i \text{ owns a car} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

- When the dependent variable represents multiple discrete outcomes (taking a small number of discrete values), we have *multiresponse* (or *multinomial choice*) models.

Example

Concerning credit ratings for companies, firms are ranged between AAA (highest rating), AA, A, BBB, BB, B, CCC, CC, C, D (lowest rating). These ratings can be grouped into 6 categories such as

$$
\begin{aligned}
y_i &= 1 \quad \text{if firm } i \text{ has a lowest score, e.g. between D and CCC} \\
&= 2 \quad \text{if firm } i \text{ has a score B} \\
&= 3 \quad \text{if firm } i \text{ has a score BB} \\
&= 4 \quad \text{if firm } i \text{ has a score BBB} \\
&= 5 \quad \text{if firm } i \text{ has a score A} \\
&= 6 \quad \text{if firm } i \text{ has a highest score, e.g. between AA and AAA.}
\end{aligned}
$$

- The dependent variable can be also constrained (*censored models*) or missing from a certain threshold (*truncated models*).

Binary Models

- Consider the following *latent* model

$$y_i^* = \alpha + \beta x_i + u_i \tag{76}$$

where y_i^* is a continuous but unobserved (or *latent*) variable. For instance, x_i is univariate, but the model can be easily extended to the multivariate case.

- y_i^* may represent the difference between the utility levels of individual i when facing a choice for example between two brands A and B, or other quantities like her income, her well-being, etc.

- Assume now we observe the decision of individual i. For example, we say individual i chooses brand A (coded as $y_i = 1$) when the difference in her utilities is positive ($y_i^* > 0$) and brand B ($y_i = 0$) otherwise:

$$\begin{aligned} y_i &= 1 \quad \text{if } y_i^* > 0 \\ &= 0 \quad \text{if } y_i^* \leq 0 \end{aligned}$$

- Taking x_i as given, the model shows that

$$P(y_i = 1) = P(y_i^* > 0) = P(u_i > -\alpha - \beta x_i) \tag{77}$$
$$P(y_i = 0) = 1 - P(y_i = 1) = P(y_i^* \leq 0) = P(u_i \leq -\alpha - \beta x_i) \tag{78}$$

- Interpretation of $E(y|x)$ as a probability:

$$E(y|x) = 1 \times P(y = 1) + 0 \times P(y = 0) = P(y = 1). \tag{79}$$

Binary Models

Probit Model

- If u_i has a normal distribution, we have the **probit model** with

$$p \equiv P(y_i = 1) = P(u_i \leq \alpha + \beta x_i) = \Phi(\alpha + \beta x_i) \quad (80)$$

$$1 - p \equiv P(y_i = 0) = 1 - P(y_i = 1) = 1 - \Phi(\alpha + \beta x_i) \quad (81)$$

  where Φ is the cumulative normal distribution.

- Estimation of this model is usually done by Maximum Likelihood where the log likelihood to be maximized is

$$\ln L(\theta) = \ln \prod_{i=1}^{N} P(y_i = 1)^{y_i} \left[1 - P(y_i = 1)\right]^{1-y_i} \quad (82)$$

$$= \sum_{i=1}^{N} \left[y_i \ln P(y_i = 1) + (1 - y_i) \ln(1 - P(y_i = 1))\right] \quad (83)$$

- Use Stata command `probit`.

Example

We estimate the probit model for the choice between two tomato ketchup: Heinz ($y_i = 1$) and Hunts ($y_i = 0$). The initial data sample includes 300 households with 2798 observations. Estimation is done on the subsample of 2498 observations, by excluding the last observation of each household.

Table: Estimation results: Probit regression

Variable	Coefficient	Std. Err.
displheinz	0.271*	0.129
featheinz	0.188	0.157
featdisplheinz	0.255	0.248
displhunts	-0.376*	0.151
feathunts	-0.573**	0.197
featdisplhunts	-1.094**	0.275
lnprice	-3.274**	0.217
Intercept	1.846**	0.076
Log-likelihood	-598.528	

Notes: ** significant at the 1% level, * at the 5 % level.
Number of observations: 2498.

Binary Models

Logit Model

- If u_i has a logistic distribution, we have the **logit model** with

$$p \equiv P(y_i = 1) = P(u_i \leq \alpha + \beta x_i) = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} \quad (84)$$

$$1 - p \equiv P(y_i = 0) = 1 - P(y_i = 1) = \frac{1}{1 + \exp(\alpha + \beta x_i)} \quad (85)$$

- Again, estimation of this model is usually done by Maximum Likelihood.
- Use Stata command `logit`.

Example

We estimate the logit model for the choice between two tomato ketchup: Heinz ($y_i = 1$) and Hunts ($y_i = 0$). The initial data sample includes 300 households with 2798 observations. Estimation is done on the subsample of 2498 observations, by excluding the last observation of each household.

Table: Estimation results: Logit regression

Variable	Coefficient	Std. Err.
displheinz	0.526*	0.254
featheinz	0.474	0.320
featdisplheinz	0.473	0.489
displhunts	-0.651*	0.254
feathunts	-1.033**	0.361
featdisplhunts	-1.981**	0.479
lnprice	-5.987**	0.401
Intercept	3.290**	0.151
Log-likelihood	-601.238	

Notes: ** significant at the 1% level, * at the 5 % level.
Number of observations: 2498.

Binary Models

Marginal Effects

- Parameters in the logit and probit models, $\theta = (\alpha, \beta)'$, can be interpreted as the effects of explanatory variables on the latent variable.
- However, we are more interested in the *marginal effects* of these variables on individual choice probabilities.
- The marginal effect of x_i on $p \equiv P(y_i = 1) = E(y|x)$ if x_i is *continuous* is given by

$$\frac{\partial E(y|x)}{\partial x_i} = \phi(\alpha + \beta x_i)\beta \quad \text{for the probit model} \tag{86}$$

$$\frac{\partial E(y|x)}{\partial x_i} = p(1-p)\beta \quad \text{for the logit model} \tag{87}$$

where ϕ is the standard normal density.

- The sign of the marginal effect of x_i on $P(y_i = 1)$ corresponds to the sign of its coefficient, β.
- If x_i is a *dummy variable*, the marginal effect should be computed for both probit and logit models as follows

$$\frac{\partial E(y|x)}{\partial x_i} = P(y_i = 1 \text{ when } x_i = 1) - P(y_i = 1 \text{ when } x_i = 0) \tag{88}$$

- The marginal effect of x_i on $P(y_i = 0)$ is given by the opposite of the marginal effect of x_i on $P(y_i = 1)$:

$$\frac{\partial P(y_i = 0)}{\partial x_i} = -\frac{\partial P(y_i = 1)}{\partial x_i}. \tag{89}$$

- Use Stata command `margins` to calculate marginal effects.

Binary Models

Goodness-of-Fit and Tests

- Contrary to the linear model, there is no single measure for the goodness-of-fit.
- Let us define $\ln L$ and $\ln L_0$ the log-likelihood of the model of interest and that of the model with only the intercept, respectively. A goodness-of-fit is given by

$$pseudo\text{-}R^2 = 1 - \frac{1}{1 + 2(\ln L - \ln L_0)/N} \tag{90}$$

- Another measure is the McFadden R^2:

$$R^2_{McFadden} = 1 - \frac{\ln L}{\ln L_0}. \tag{91}$$

- Three ML-based tests (Wald, likelihood ratio, Lagrange multiplier) can be employed for testing hypotheses in probit and logit models.

Multiresponse Models

- In many applications, the number of alternatives to be chosen is larger than 2. For example, choice between M different products.
- If there is a logical ordering of the alternatives, in particular when the underlying latent variable that drives the choices between the alternatives, we have **ordered response models**.
- **Unordered response models** arise when there is no logical ordering.
- Estimation of these models is generally done by ML.

Multiresponse Models

Ordered Response Models

- Let us consider the choice between M alternatives, numbered from 1 to M. Assume there is a *logical ordering* in these alternatives (e.g., no car, one car, more than one car).
- The model is described by

$$\begin{aligned} y_i^* &= \beta x_i + u_i \\ y_i &= 1 \quad \text{if} \quad y_i^* \leq \gamma_1 \\ &= 2 \quad \text{if} \quad \gamma_1 < y_i^* \leq \gamma_2 \\ &= 3 \quad \text{if} \quad \gamma_2 < y_i^* \leq \gamma_3 \\ &\ldots \\ &= M \quad \text{if} \quad \gamma_{M-1} < y_i^* \end{aligned}$$

- We observe in this model that by *normalization* the intercept is set to zero (i.e. $\alpha = 0$).

- If we really want the presence of the intercept α in the model, we have to set, by *normalization*, $\gamma_1 = 0$. The model becomes:

$$
\begin{aligned}
y_i^* &= \alpha + \beta x_i + u_i \\
y_i &= 1 \quad \text{if} \quad y_i^* \leq 0 \\
&= 2 \quad \text{if} \quad 0 < y_i^* \leq \gamma_1 \\
&= 3 \quad \text{if} \quad \gamma_1 < y_i^* \leq \gamma_2 \\
&\ldots \\
&= M \quad \text{if} \quad \gamma_{M-2} < y_i^*
\end{aligned}
$$

- For example, in the example about firm's credit ratings above, a particular bond falls in the category 3 if its unobserved (or latent) creditworthiness falls within a certain range that is too low to be classified as BBB (2) or too high to be classified as B (4).
- Assuming that u_i is standard normal results in the **ordered probit model**.
- Assuming u_i has a logistic distribution gives the **ordered logit model**.
- Use Stata commands `oprobit` and `ologit`.

Example

We use a ordered probit model to study a confidential data sample on three risk profiles for 2000 clients of an investment firm. Category 1 is associated with individuals who do not take much risk as they, for example, only have a saving account. Category 3 corresponds with those who apparently are willing to take high financial risk, like those who often trade in financial derivatives. There is clearly a logical ordering in the classification of individual risk profiles.

Table: Estimation results: Ordered probit model

Variable	Coefficient	Std. Err.
TYPE2FUNDS	0.105**	0.008
TRANSTYPE1	-0.007	0.010
TRANSTYPE3	0.008**	0.002
wealth	0.173	0.110
γ_1	-0.420**	0.035
γ_2	1.305**	0.044
Log-likelihood	-1826.693	

Notes: ** significant at the 1% level, * at the 5 % level.
Number of observations: 2000.

Multiresponse Models

Multinomial Models

- There is no natural ordering in the alternatives. For example, choice of different modes of transportation (bus, train, car, bycycle, walking.
- Consider the choice over M alternatives without any logical ordering between them. The **multinomial logit model** is frequently used in this situation:

$$P(y_i = 1) = \frac{1}{1 + \exp(\alpha_2 + \beta_2 x_i) + ... + \exp(\alpha_M + \beta_M x_i)} \quad (92)$$

$$P(y_i = j) = \frac{\exp(\alpha_j + \beta_j x_i)}{1 + \exp(\alpha_2 + \beta_2 x_i) + ... + \exp(\alpha_M + \beta_M x_i)} \quad (93)$$

  where $j = 2, ..., M$.
- We remark that one category is chosen as the benchmark (here $j = 1$) and that the regressors x_i are the same for all categories.
- The model is estimated by ML. The model can be easily extended in the multivariate case. **Multinomial probit model** exists but difficult to be handled.
- Use Stata command `mlogit`.

Example

Consider the previous example on individual risk profiles. By ignoring the logical ordering in the classification, we can use the multinomial logit model with 3 categories (category 2 as the base outcome).

Table: Estimation results: Multinomial logit model

Variable	Coefficient	Std. Err.
Category 1		
TYPE2FUNDS	-0.178**	0.029
TRANSTYPE1	0.006	0.018
TRANSTYPE3	-0.037	0.026
wealth	0.108	0.391
Intercept	-0.471**	0.072
Category 3		
TYPE2FUNDS	0.161**	0.018
TRANSTYPE1	-0.010	0.029
TRANSTYPE3	0.056**	0.014
wealth	0.442	0.312
Intercept	-1.858**	0.094
Log-likelihood	-1806.83	

- In the case where the explanatory variables are different for each category (x_{ij}), we should use the **conditional logit model**:

$$P(y_i = j) = \frac{\exp(\alpha_j + \beta x_{ji})}{\exp(\alpha_1 + \beta x_{1i}) + ... + \exp(\alpha_M + \beta x_{Mi})} \tag{94}$$

  where $j = 1, ..., M$
- By identification, one intercept should be zero, for example $\alpha_1 = 0$.
- The model is estimated by ML.
- Stata command `clogit` (usual conditional logit) and `asclogit` (McFadden conditional logit) help estimate the conditional logit model. Remark: the data structure under McFadden conditional logit and under conditional logit is very particular, i.e. it is different from that under the multinomial logit, ordered probit or ordered logit models.

Example

We apply a conditional logit model to estimate the effects of marketing variables on the choice about 4 brands of saltine crackers (Private, Sunshine, Keebler, Nabisco). The initial data set contains 3292 purchases made by 136 households. By excluding the last purchases of each households, the sample used in estimation has 3156 obs.

Table: Estimation results: Conditional logit model

Variable	Coefficient	Std. Err.
Marketing variables		
displ	0.049	0.068
feat	0.412**	0.151
featdispl	0.580**	0.119
price	-3.172**	0.216
Intercepts		
Keebler	-1.968**	0.075
Private	-1.814**	0.104
Sunshine	-2.465**	0.082
Log-likelihood	-3215.830	

Notes: ** significant at the 1% level, * at the 5 % level.

Number of observations: 3156.

Censored and Truncated Models

- Standard linear models assume that we observe all values of the dependent variable (i.e. there is no missing observations).
- However, in practice we only observe the dependent variable from a certain threshold, named **truncated variable**. Observations above (or below) (including also explanatory variables) cannot be observed (or missing). For example, we can only observe positive expenditures on tobacco and alcohol. The threshold corresponding in this case is zero. People who do not want to buy tobacco and alcohol or have zero expenditures on these products cannot be observed. In this case, data are truncated from below (observations below zero are cut out from the sample).

- In some cases, the dependent variable is constrained, denoted as **censored variable**. For example, when studying the donation behavior of individuals to charity, we does observe the donations of individuals who give nothing (donation = 0) or a positive amount.
- The dependent variable may be censored below (or left censored) or above (or right censored). The case of charity donation corresponds to left censoring at 0.
- Values of explanatory variables are still observable in the censored model while they are unobserved (or missing) in the truncated model.

Censored and Truncated Models

Standard Censored or Tobit Models

- The standard censored model, also known as **Tobit model** or **Tobit I model**, is described by

$$y_i^* = \alpha + \beta x_i + u_i \tag{95}$$

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases} \tag{96}$$

- Assume that u_i is standard normal and independent of x_i, i.e. $u_i \sim N(0,1)$.

- We have

$$\begin{aligned} P(y_i = 0) &= P(y_i^* \leq 0) = P(u_i \leq -\alpha - \beta x_i) \\ &= \Phi(-\alpha - \beta x_i) = 1 - \Phi(\alpha + \beta x_i) \end{aligned}$$

and

$$P(y_i = y_i^* | y_i^* > 0) = \begin{cases} \frac{P(y_i = y_i^*)}{P(y_i > 0)} & \text{if } y_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

- The model is estimated by ML.
- Use Stata command `tobit` to estimate censored models.

Example

Consider the donation behavior on a sample of 4268 individuals who received a mailing from a charitable institution. We use the first 4000 individuals to estimate the censored model as the donation censored below (or left-censored) at 0.

Table: Estimation results: Tobit model

Variable	Coefficient	Std. Err.
resplastmail	2.476	1.359
weekslastresp	-0.119**	0.018
mailsperyear	4.725**	0.785
propresponse	32.987**	2.676
averagegift	0.025**	0.005
Intercept	-29.469**	2.888
Log-likelihood	-8793.157	

Notes: ** significant at the 1% level, * at the 5 % level.

Number of observations: 4000, left-censored: 2373, uncensored: 1627.

Censored and Truncated Models

Truncated Models

- In some situations, observations are completely missing if $y_i^* \leq 0$ (e.g. we only observe tobacco consumption and other explanatory variables of tobacco consumers; non-tobacco consumers are not observed!):

$$y_i^* = \alpha + \beta x_i + u_i \tag{97}$$

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ \text{not observed} & \text{if } y_i^* \leq 0 \end{cases} \tag{98}$$

- The probability to observe y_i is

$$P(y_i = y_i^* | y_i^* > 0) = \begin{cases} \frac{P(y_i = y_i^*)}{P(y_i > 0)} & \text{if } y_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

- Use Stata command `truncreg` to estimate truncated models.

Censored and Truncated Models

Sample Selection Model

- The **sample selection model** is also referred as the **Tobit II model**:

$$w_i^* = a + bx_i + v_i \quad (\text{ selection equation}) \tag{99}$$

$$y_i^* = \alpha + \beta x_i + u_i \quad (\sim \text{ Tobit equation}) \tag{100}$$

$$y_i = \begin{cases} y_i^* & \text{if } w_i^* > 0 \\ \text{not observed} & \text{if } w_i^* \leq 0 \end{cases} \tag{101}$$

- The unavailability of observations associated with $w_i^* \leq 0$ constitutes the sample selection problem, inducing in a bias in estimation if it is not correctly controlled for.
- This model can be appropriately estimated by ML.

- Another popular method is the **Heckman's two-step procedure** (sometimes called the **Heckit model**).
  1. Run ML on the following probit model (or *the selection equation*) to estimate a and b:
  $$w_i = \begin{cases} 1 & \text{if } w_i^* > 0 \\ 0 & \text{if } w_i^* \leq 0 \end{cases} \tag{102}$$
  2. Run the OLS on the following linear model with heteroscedasticity (with only the sub-sample associated with $w^* > 0$) to estimate α and β:
  $$y_i = \alpha + \beta x_i + \rho\lambda(\hat{a} + \hat{b}x_i) + \varepsilon_i \tag{103}$$
  where $\lambda(\hat{a} + \hat{b}x_i)$ is the **inverse Mills ratio** which is computed from the first-step estimation
  $$\lambda(\hat{a} + \hat{b}x_i) = \frac{\phi(\hat{a} + \hat{b}x_i)}{\Phi(\hat{a} + \hat{b}x_i)} \tag{104}$$
- The inclusion of $\lambda(.)$ would eliminate the problem of *selection bias*.
- Stata command `heckman` performs the estimation of this model.

Example

Consider again the example on donations to charity but now with the sample selection model. The Heckman's two-step procedure is implemented to estimate the model. Again we use the first 4000 individuals.

Table: Estimation results: Tobit II or sample selection model

	Selection equation		Tobit equation	
Variable	Coefficient	Std. Err.	Coefficient	Std. Err.
resplastmail	0.139*	0.059	-2.804	4.913
weekslastresp	-0.004**	0.001	0.325*	0.150
mailsperyear	0.162**	0.034	-3.124	4.167
propresponse	1.778**	0.116946	-85.469*	40.341
averagegift	0.001	0.001	-0.011	0.020
Intercept	-1.295**	0.121	129.472*	57.147
Inverse Mills ratio			-78.902*	36.326

Notes: ** significant at the 1% level, * at the 5 % level.

Number of observations: 4000, left-censored: 2373, uncensored: 1627.

Chapter 6

Panel Data Models

What are panel data ?

- It is the situation when we have data comprising both time series and cross-sectional elements. In other words, we have repeated observations over the same units (individuals, households, firms, countries, etc.) collected over a number of periods.
- Such a dataset is known as panel data or longitudinal data.
- Panel data is modelled as follows

$$y_{it} = \alpha + \beta x_{it} + u_{it}, \quad i = 1, 2, ..., N, \ t = 1, 2, ..., T, \tag{105}$$

- In this chapter, we only deal with *balanced panel data* (i.e. we have the same number of individuals per time period or equivalently the same number of time periods per individual). The discussion remains valid when considering *unbalanced panel data* where the number of observations per individual is not the same (or equivalently, the number of individuals per time period varies).
- Stata easily manages unbalanced panel data.

- For instance, x_{it} is univariate. It is easy to extend the analysis to the multivariate case.
- A usual panel data model assumes the existence of *individual effects*:

$$u_{it} = \mu_i + v_{it}, \tag{106}$$

- v_{it} is the standard residual term.
- μ_i is the individual effect, also referred to as *unobserved effect* or *unobserved heterogeneity* specific to individual i.
- μ_i encapsulates all factors, specific to individual i and time-invariant, that are not included in the regressors x_{it}.
- When μ_i is considered as fixed, we have the *fixed effects model*. When μ_i is considered as random, we have the *random effects model*.

- Stata command `xtreg` estimates panel data models.
- It is possible to have *time effects* with the following assumption:

$$u_{it} = \lambda_t + v_{it}, \tag{107}$$

  where λ_t is *unobserved effect* or *unobserved heterogeneity* specific to time t.
- In both cases, v_{it} is the usual regression error (known as *idiosyncratic error*).
- Notations: let z denote either x, y, u, or v, we define

$$\bar{z}_i = \frac{1}{T}\sum_{t=1}^{T} z_{it}, \quad \bar{z}_t = \frac{1}{N}\sum_{i=1}^{N} z_{it}, \quad \bar{z} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} z_{it}.$$

Fixed Effects

Individual Fixed Effects

- Consider the model

$$y_{it} = \alpha + \beta x_{it} + u_{it}, \quad i = 1, 2, ..., N, \ t = 1, 2, ..., T, \tag{108}$$

where

$$u_{it} = \mu_i + v_{it}, \tag{109}$$

- We also assume that $\sum_i^N \mu_i = 0$.
- By construction β can be estimated by OLS (called as *within estimator* or fixed effect estimator, $(\hat{\beta}_{FE})$) applied on the following demeaned model

$$y_{it} - \bar{y}_i = \beta(x_{it} - \bar{x}_i) + v_{it} - \bar{v}_i. \tag{110}$$

- α can be estimated by

$$\hat{\alpha}_{FE} = \bar{y} - \hat{\beta}_{FE}\bar{x} \tag{111}$$

Fixed Effects

Least Squares Dummy Variable Model

- The fixed effects model is equivalent to the Least Squares Dummy Variables (LSDV) model. The LSDV model with the intercept is

$$y_{it} = \alpha + \beta x_{it} + \mu_2 D_2 + \mu_3 D_3 + ... + \mu_N D_N + v_{it}. \tag{112}$$

  with the identification constraint $\sum_{i=1}^{N} \mu_i = 0$ (or $\mu_1 = 0$).
- Another writing of the LSDV model, without any constraint on μ_i, is

$$y_{it} = \beta x_{it} + \mu_1 D_1 + \mu_2 D_2 + ... + \mu_N D_N + v_{it} \tag{113}$$

  where D_i is the dummy variable that takes the value 1 for all observations on individual i in the sample and 0 otherwise.
- The LSDV model can be estimated by OLS.

Fixed Effects

Time Fixed Effects

- Consider the model

$$y_{it} = \alpha + \beta x_{it} + u_{it}, \quad i = 1, 2, ..., N, \ t = 1, 2, ..., T, \tag{114}$$

where

$$u_{it} = \lambda_t + v_{it}, \tag{115}$$

- We also assume by identification that $\sum_i^N \lambda_t = 0$.
- By construction β can be estimated by OLS (*between estimator*, $\hat{\beta}_{BE}$) applied on the following demeaned model (referred to as *between transformation*)

$$y_{it} - \bar{y}_t = \beta(x_{it} - \bar{x}_t) + v_{it} - \bar{v}_t \tag{116}$$

where $\bar{y}_t = \frac{1}{N}\sum_{i=1}^N y_{it}$, $\bar{x}_t = \frac{1}{N}\sum_{i=1}^N x_{it}$, $\bar{v}_t = \frac{1}{N}\sum_{i=1}^N v_{it}$.

- α can be estimated by

$$\hat{\alpha}_{BE} = \bar{y} - \hat{\beta}_{BE}\bar{x}. \tag{117}$$

Fixed Effects

Twoway Fixed Effects

- We can have a twoway fixed effects model with both individual and time fixed effects

$$y_{it} = \alpha + \beta x_{it} + u_{it}, \quad i = 1, 2, ..., N, \ t = 1, 2, ..., T, \tag{118}$$

where

$$u_{it} = \mu_i + \lambda_t + v_{it}, \tag{119}$$

- Similarly, we assume $\sum_i^N \mu_i = 0$ and $\sum_t^T \lambda_t = 0$.
- Estimation of β can be obtained by OLS (*within estimator*, $\hat{\beta}_{FE}$) applied to the following transformed model

$$y_{it} - \bar{y}_i - \bar{y}_t + \bar{y} = \beta(x_{it} - \bar{x}_i - \bar{x}_t + \bar{x}) + (v_{it} - \bar{v}_i - \bar{v}_t + \bar{v}) \tag{120}$$

- The intercept is estimated by

$$\hat{\alpha}_{FE} = \bar{y} - \hat{\beta}_{FE}\bar{x}. \tag{121}$$

Example

We use the panel fixed effects model to explain firm's debt ratio which is assumed to depend upon the past value of this ratio and firm characteristics, known at time $t-1$ and related to the costs and benefits of operating with various leverage ratios.

$$mdr_{it} = \alpha + x'_{i,t-1}\beta + \rho\ mdr_{i,t-1} + \mu_i + v_{it} \tag{122}$$

Table: Estimation results: Firm fixed effects

Variable	Coefficient	Std. Err.
mdr_{-1}	0.535**	0.008
lagebit_ta	-0.050**	0.008
lagmb	0.002*	0.001
lagdep_ta	-0.124*	0.058
laglnta	0.038**	0.002
lagfa_ta	0.059**	0.013
lagrd_ta	-0.066*	0.027
lagindmedian	0.167**	0.019
lagrated	0.021**	0.005
Intercept	-0.601**	0.038

Notes: ** significant at the 1% level, * at the 5 % level.

Number of observations: 19573.

Random Effects

Individual Random Effects Model

- Consider the model

$$y_{it} = \alpha + \beta x_{it} + u_{it}, \quad i = 1, 2, ..., N, \ t = 1, 2, ..., T, \tag{123}$$

where

$$u_{it} = \mu_i + v_{it}, \tag{124}$$

- We also assume that μ_i has zero mean and variance σ^2_μ, is independent of v_{it} and independent of the explanatory variables x_{it}. Notation: $\mu_i \sim IID(0, \sigma^2_\mu)$ (*Independent and Identically Distributed*).
- Another assumption is $v_{it} \sim IID(0, \sigma^2_v)$.

- Estimation of this model is usually performed by GLS, which is the OLS estimator applied to the quasi-demeaned model

$$y_{it} - \psi \bar{y}_i = \alpha(1-\psi) + \beta(x_{it} - \psi \bar{x}_i) + \underbrace{(v_{it} - \psi \bar{v}_i)}_{\text{new iid error}} \tag{125}$$

where

$$\psi = 1 - \frac{\sigma_v}{\sqrt{T\sigma_\mu^2 + \sigma_v^2}}.$$

- Usually, the RE estimator (or GLS estimator, denoted as $\hat{\beta}_{RE}$) is obtain from a two-step procedure:
  1. Use (123) to calculate ψ. Note that $\hat{\sigma}_v^2$ is obtained using the FE regression while $\hat{\sigma}_\mu^2$ is based on the between regression $\bar{y}_i = \beta \bar{x}_i + \mu_i + \bar{v}_i$ where $V(\mu_i + \bar{v}_i) = \sigma_\mu^2 + \frac{\sigma_v^2}{T}$.
  2. Run OLS on (125) in order to obtain the GLS estimator.
- Twoway random effects model (i.e. model with both random individual effects and random time effects) exists, but is very slow to be estimated.

Example

Results of the panel random effects model for firm's debt ratio are reported as follows.

Table: Estimation results: Firm random effects

Variable	Coefficient	Std. Err.
mdr_{-1}	0.788**	0.006
lagebit_ta	-0.036**	0.006
lagmb	0.001	0.001
lagdep_ta	-0.301**	0.040
laglnta	0.002*	0.001
lagfa_ta	0.032**	0.007
lagrd_ta	-0.162**	0.016
lagindmedian	0.063**	0.012
lagrated	0.011**	0.004
Intercept	0.042**	0.015

Notes: ** significant at the 1% level, * at the 5 % level.

Number of observations: 19573.

Tests

Existence of Fixed Effects

- The null hypothesis H_0: $\mu_2 = \mu_3 = ... = \mu_N = 0$ against the alternative hypothesis H_1: at least one $\mu_i \neq 0$.
- Under the null, the model can be estimated by OLS (which give $\hat{\beta}_{OLS}$). Under the alternative we have the within estimator or the LSDV estimator ($\hat{\beta}_{FE}$).
- The test corresponds to a F test, the statistic of which is

$$F = \frac{(S_0 - S)/(N-1)}{S/(NT - N - K - 1)} \tag{126}$$

where S_0 is the residual sums of squares of the null model ($S_0 = \sum_i \sum_t (y_{it} - \hat{\alpha}_{OLS} - \hat{\beta}_{OLS} x_{it})^2$) and S is the residual sums of squares of the alternative model ($S_0 = \sum_i \sum_t (y_{it} - \hat{\alpha}_{FE} - \hat{\beta}_{FE} x_{it} - \hat{\mu}_2 D_{2i} - ... - \hat{\mu}_N D_{Ni})^2$).

- The test follows under the null hypothesis a Fisher distribution, $F_{(N-1, NT-N-K-1)}$.

Tests

Existence of Random Effects

- The null hypothesis H_0: $\sigma_\mu^2 = 0$ against the alternative hypothesis H_1: $\sigma_\mu^2 \neq 0$.
- We use the Breusch-Pagan test which corresponds to a Lagrange Multiplier test.
- The test statistic, LM, follows under the null a Chi-squared distribution with 1 degree of freedom, $\chi^2(1)$.

Tests

Fixed Effects or Random Effects

- We use the Hausman test to compare the null hypothesis $H_0 : E(\mu_i|x_{it}) = 0$ against the alternative $H_1 : E(\mu_i|x_{it}) \neq 0$.
- In fact, we compare the random effects (or GLS) estimator $\hat{\beta}_{RE}$ (the null) to the fixed effects (or within) estimator $\hat{\beta}_{FE}$ (the alternative).
- Indeed, under H_0 the RE estimator is consistent and efficient, the FE estimator is consistent; under H_1 the RE estimator is inconsistent, the FE estimator is still consistent.
- The Hausman test statistic compares $\hat{\beta}_{RE}$ and $\hat{\beta}_{FE}$. It follows a Chi-squared distribution with K degrees of freedom, $\chi^2(K)$, where K is the number of elements in β.
- If α is included in the test, then there should have $K + 1$ elements.

Example

- In the example on firm's debt ratio, the F-test statistic for the existence of fixed effect is 2.15, much higher than the critical value at the 5% level of a $F(3776, 15787)$, leading to the rejection of the null hypothesis. We conclude that fixed effects exist.
- The Breusch and Pagan' Lagrangian multiplier test for random effects gives a statistic equal to 1.26, which is lower than the 5% critical value of a $\chi^2(1)$. We cannot reject the null hypothesis of non-existence of random effects.
- The Hausman statistic comparing the random effects estimator (GLS) to the fixed effects estimator (within) reports a value 2658.53, much higher than the 5% critical value of a $\chi^2(9)$. We can conclude that the GLS estimator is rejected in favor of the within estimator.

Further Developments

- When lag dependent variable is included in right hand side of the regression, alternative estimation methods should be employed. Consider the following dynamic model

$$y_{it} = \alpha + \beta x_{it} + \rho y_{i,t-1} + \underbrace{\mu_i + v_{it}}_{u_{it}} \tag{127}$$

  Clearly, the error term u_{it} is correlated with $y_{i,t-1}$ ($E(y_{i,t-1}u_{it}) \neq 0$) which induces the inconsistency of the GLS and within estimators.
- Alternative methods: IV and GMM.
- There also exist panel data models with limited dependent variable.

Chapter 7

Causal Inference

Concepts

- *Observable/Actual Outcome* vs *Potential/Counterfactual Outcome*
- Let D_i denote treatment status for unit i. $D_i = 1$ if i receives treatment (or treated), $D_i = 0$ if i does not (untreated/control).
- Observable/actual outcome: Y_i
- Each unit has 2 potential outcomes: Y_i^1 = potential outcome when treatment occurred, Y_i^0 = potential outcome when treatment did not occur.
- Switching equation:

$$Y_i = Y_i^0 + (Y_i^1 - Y_i^0)D_i \tag{128}$$

  only one outcome is observed!
- The causal effect

$$\rho_i = Y_i^1 - Y_i^0 \tag{129}$$

  is unknown.

Concepts

- *Average Treatment Effect* (ATE)

$$\begin{aligned} ATE &= E[\rho_i] \\ &= E[Y_i^1 - Y_i^0] \\ &= E[Y_i^1] - E[Y_i^0] \end{aligned}$$

  can be estimated.

- *Average Treatment Effect for the Treated* (ATT)

$$\begin{aligned} ATT &= E[\rho_i \mid D_i = 1] \\ &= E[Y_i^1 - Y_i^0 \mid D_i = 1] \\ &= E[Y_i^1 \mid D_i = 1] - E[Y_i^0 \mid D_i = 1] \end{aligned}$$

- *Average Treatment Effect for the Untreated* (ATU)

$$\begin{aligned} ATU &= E[\rho_i \mid D_i = 0] \\ &= E[Y_i^1 - Y_i^0 \mid D_i = 0] \\ &= E[Y_i^1 \mid D_i = 0] - E[Y_i^0 \mid D_i = 0] \end{aligned}$$

Concepts

- Let π denote the share of units receiving treatment (size of the treated group) and $1-\pi$ the share of the size of the control group.
- Note that

$$\text{Simple Difference in Outcomes (SDO)} = E[Y_i \mid D=1] - E[Y_i \mid D=0] \tag{130}$$

- It is shown that

$$\begin{aligned} E[Y^1 \mid D=1] - E[Y^0 \mid D=0] &= ATE \\ &\quad + E[Y_i^0 \mid D=1] - E[Y_i^0 \mid D=0] \\ &\quad + (1-\pi)\,(ATT - ATU) \end{aligned}$$

- Then,

$$SDO = ATE + \text{Selection bias} + \text{Heterogeneous treatment effect bias} \tag{131}$$

- Or, equivalently (if homogeneous treatment effect bias, $ATT = ATU$),

$$E[Y^1 \mid D=1] - E[Y^0 \mid D=0] = ATT + \text{Selection bias}. \tag{132}$$

Concepts

- If *constant treatment effect*, $\rho_i = \rho \quad \forall i$, $ATE = ATT = ATU$, and

$$SDO = ATE + \text{Selection bias} \tag{133}$$

- The *Conditional Expectation Function* (CEF) decomposition:

$$Y_i = E[Y_i \mid X_i] + \varepsilon_i \tag{134}$$

- *Conditional Independence Assumption* (CIA): Treatment is independent of potential outcomes:

$$(Y_i^1, Y_i^0) \perp D_i \mid X_i \tag{135}$$

- CIA = "*as good as random assignment*".

Concepts

- *Stable Unit Treatment Value Assumption* (SUTVA): each unit receives the same (sized-dose) treatment, there is no spillover to other units' potential outcomes.
- The CIA is the most important assumption. It eliminates selection bias, leading to a *causal* CEF: it justifies the causal interpretation of regression estimates.
- If treatment assignment is *randomized*, selection bias and heterogeneous treatment effect bias = 0. Together with the independence assumption, $SDO = ATE = ATT$:

$$\begin{aligned} E[Y_i \mid D=1] - E[Y_i \mid D=0] &= E[Y_i^1 \mid D=1] - E[Y_i^0 \mid D=0] \\ &= E[Y_i^1 \mid D=1] - E[Y_i^0 \mid D=1] \\ &= E[Y_i^1 - Y_i^0 \mid D=1] \\ &= E[Y_i^1 - Y_i^0] \end{aligned}$$

Regression Analysis of Experiments

- The model for *constant treatment effect* is

$$\begin{aligned} Y_i &= Y_i^0 + (Y_i^1 - Y_i^0)D_i && (136)\\ &= E[Y_i^0] + (Y_i^1 - Y_i^0)D_i + Y_i^0 - E[Y_i^0] && (137)\\ &= \alpha + \rho D_i + \eta_i && (138) \end{aligned}$$

- We have then

$$\begin{aligned} E[Y_i \mid D_i = 1] &= \alpha + \rho + E[\eta_i \mid D_i = 1] && (139)\\ E[Y_i \mid D_i = 0] &= \alpha + E[\eta_i \mid D_i = 0] && (140) \end{aligned}$$

- Therefore,

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = \underbrace{\rho}_{\text{treat. effect}} + \underbrace{E[\eta_i \mid D_i = 1] - E[\eta_i \mid D_i = 0]}_{\text{selection bias}} \quad (141)$$

- Note that

$$E[\eta_i \mid D_i = 1] - E[\eta_i \mid D_i = 0] = E[Y_i^0 \mid D_i = 1] - E[Y_i^0 \mid D_i = 0] \quad (142)$$

If D_i is randomly assigned, the selection bias dissapears.

Matching

- Two main assumptions:
  - CIA: $(Y_i^1, Y_i^0) \perp D_i \mid X_i$.
  - Common support: $0 < Pr(D_i = 1 \mid X_i) < 1$.
- When treatment assignment is conditional on some observable variables (X) (known as *selection on observables*), X can be thought of covariates satisfying the CIA, hence leading to identify the causal treatment effect.
- Thus,

$$\begin{aligned} E[Y^1 - Y^0 \mid X] &= E[Y^1 - Y^0 \mid X, D = 1] && (143) \\ &= E[Y^1 \mid X, D = 1] - E[Y^0 \mid X, D = 0] && (144) \\ &= E[Y \mid X, D = 1] - E[Y \mid X, D = 0] && (145) \end{aligned}$$

- And the estimator for the *Average Treatment Effect* (ATE):

$$\hat{\rho}_{ATE} = \int (E[Y \mid X, D = 1] - E[Y \mid X, D = 0]) \, \mathrm{d}Pr(X) \qquad (146)$$

Matching

- An exact matching estimator for ATT (summing over the treatment group):

$$\hat{\rho}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} \left(Y_i - Y_{j(i)} \right) \tag{147}$$

  where $Y_{j(i)}$ is the jth unit (in the control group) matched with the ith unit based on the jth being closest to the ith unit for some covariates X.
- When there is M jth units matched with ith unit,

$$\hat{\rho}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} \left(Y_i - \frac{1}{M} \sum_{m=1}^{M} Y_{j,m(i)} \right) \tag{148}$$

- Estimator for the ATE (matching over both treatment and control groups):

$$\hat{\rho}_{ATE} = \frac{1}{N} \sum_{i=1}^{N} (2D_i - 1) \left[Y_i - \frac{1}{M} \sum_{m=1}^{M} Y_{j,m(i)} \right] \tag{149}$$

Propensity Score Matching

- Under the CIA,

$$(Y_i^1, Y_i^0) \perp D \mid p(X), \tag{150}$$

where *propensity score* $p(X) \equiv Pr(D = 1 \mid X) = F(\beta_0 + X'\beta_1)$ (using logit).

- We have

$$\rho_{ATE} = E[Y^1 - Y^0] = E\left[Y \frac{D - p(X)}{p(X)(1 - p(X))}\right] \tag{151}$$

$$\rho_{ATT} = E[Y^1 - Y^0 \mid D = 1] = \frac{1}{Pr(D = 1)} E\left[Y \frac{D - p(X)}{1 - p(X)}\right] \tag{152}$$

- The corresponding estimators are

$$\hat{\rho}_{ATE} = \frac{1}{N} \sum_{i=1}^{N} Y_i \frac{D_i - \hat{p}(X_i)}{\hat{p}(X_i)(1 - \hat{p}(X_i))} \tag{153}$$

$$\hat{\rho}_{ATT} = \frac{1}{N_T} \sum_{i=1}^{N} Y_i \frac{D_i - \hat{p}(X_i)}{1 - \hat{p}(X_i)} \tag{154}$$

and the associated variances calculated by bootstrap.

Regression Discontinuity

- Assignment variable X, or *running variable*, is an observable confounder since it causes both Y (outcome) and D (treatment status).
- $X = c_0$ is the cutoff (threshold) where the treatment and the control subjects overlap in the limit.
- The probability of treatment assignment (untreated/treated) changes discontinuously at c_0.
- Comparing the outcomes for subjects above and below c_0 gives the *Local Average Treatment Effect* (LATE).
- *Continuity Assumption*: potential outcomes are continuous at cutoff c_0.
- We need the *continuity assumption* because of absence of overlap (i.e. *absence of common support*). It also means that the cutoff is exogenous (i.e. it is not endogenous to other interventions).

Regression Discontinuity

Example

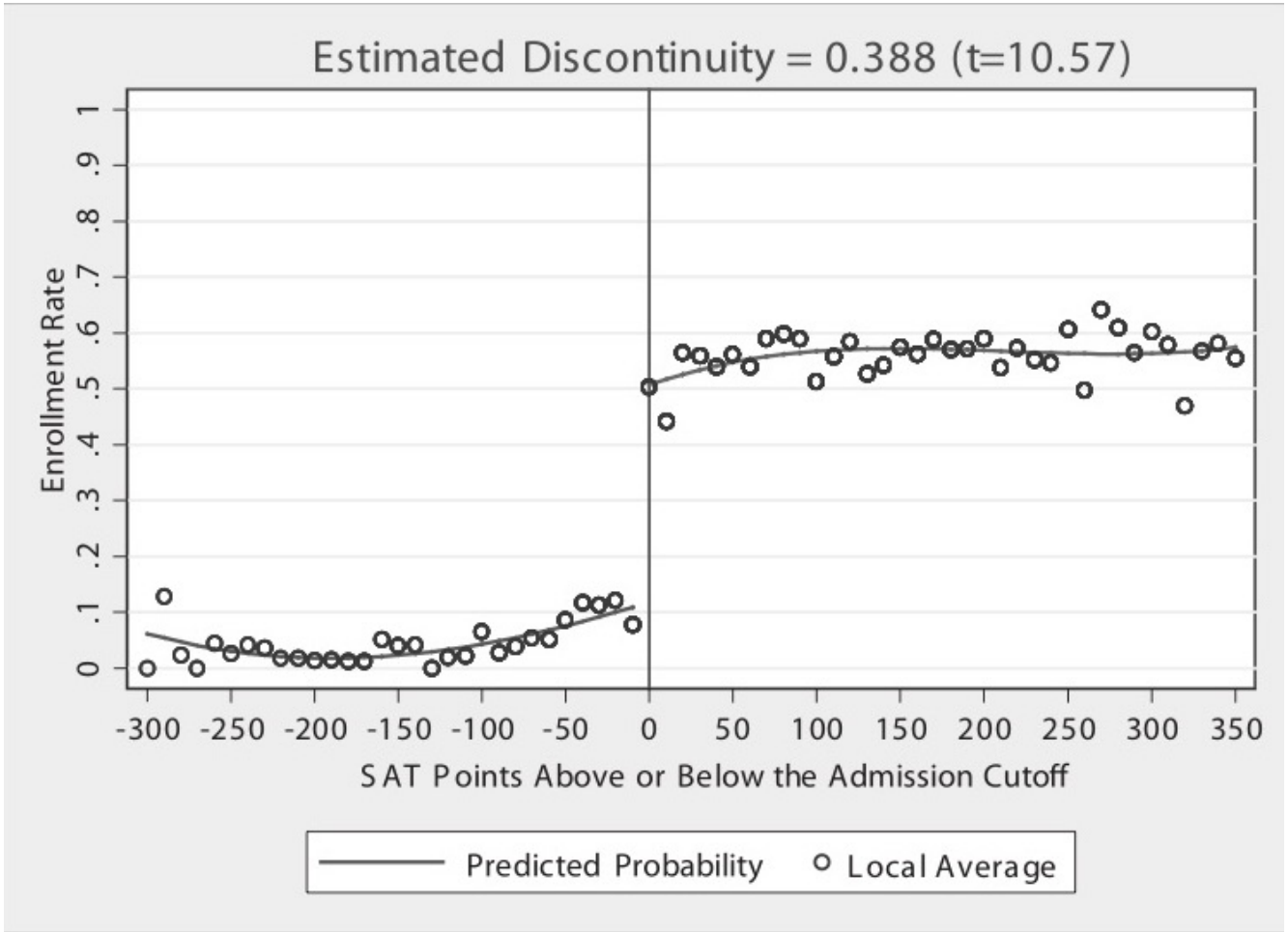


Figure: Attending the US state flagship university as a function of re-centered standardized test scores. Source: Cunningham (2021).

Regression Discontinuity

Sharp vs Fuzzy RD Design

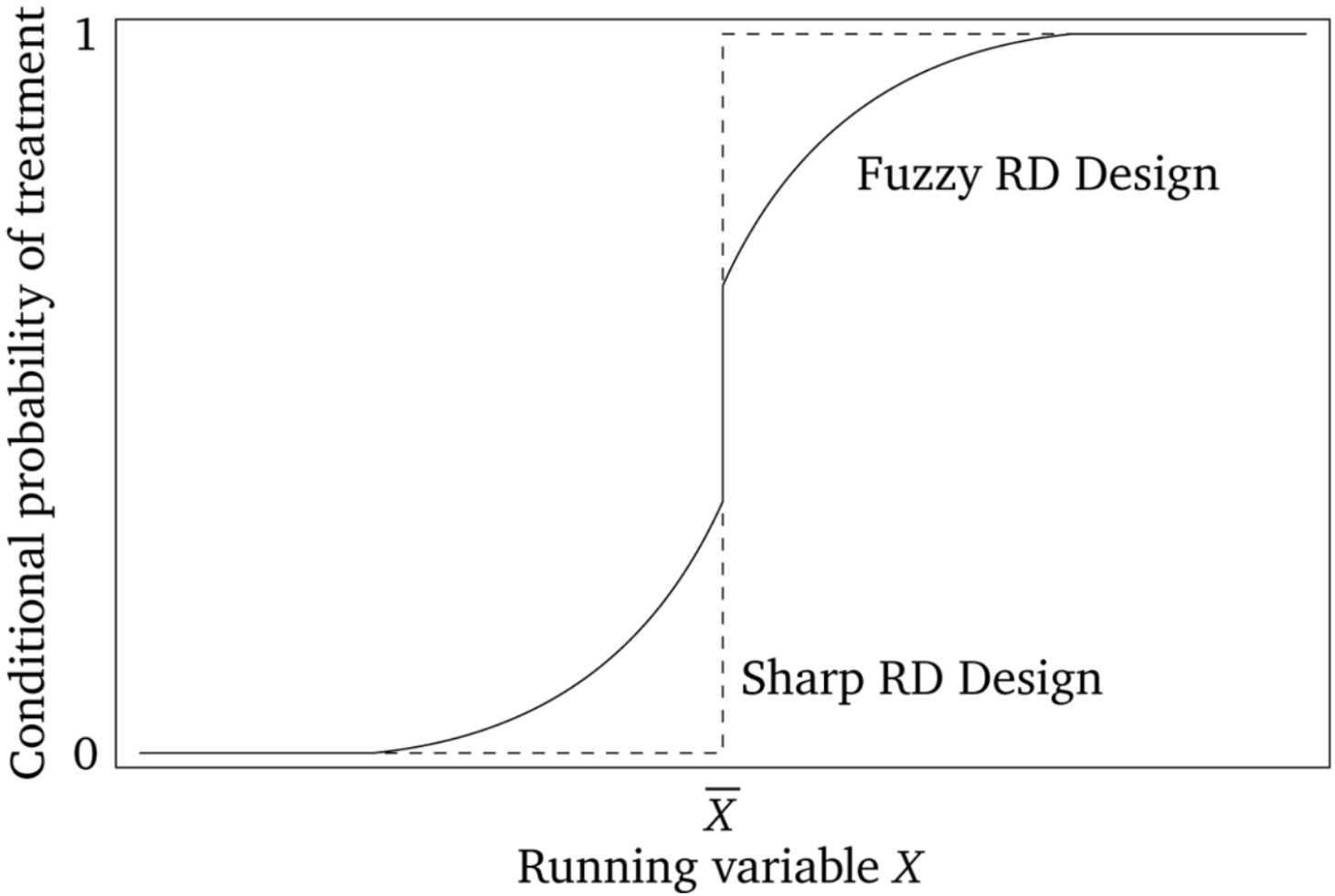


Figure: Sharp and Fuzzy RD. Source: Cunningham (2021).

Regression Discontinuity

- Formally, in a *sharp RD*, the treatment assignment status (D_i) is defined *deterministically* at cutoff/threshold c_0 of running variable X_i:

$$D_i = \begin{cases} 1 & \text{if } X_i \geq c_0 \\ 0 & \text{if } X_i < c_0 \end{cases} \tag{155}$$

- From the switching equation, $Y_i = Y_i^0 + (Y_i^1 - Y_i^0)D_i$, we get the sharp RD regression

$$Y_i = \alpha + \beta X_i + \rho D_i + \varepsilon_i \tag{156}$$

where ρ is the *constant treatment effect*.

- The continuity assumption states that $E[Y_i^1 \mid X_i = c_0]$ and $E[Y_i^0 \mid X_i = c_0]$ are continuous (smooth) functions across c_0.
- ρ, the average causal effect as the running variable approaches the cutoff, is estimated by

$$\rho_{SharpRD} = \lim_{X_i \to c_{0+}} E[Y_i^1 \mid X_i = c_0] - \lim_{X_i \to c_{0-}} E[Y_i^0 \mid X_i = c_0] \tag{157}$$

$$= \lim_{X_i \to c_{0+}} E[Y_i \mid X_i = c_0] - \lim_{X_i \to c_{0-}} E[Y_i \mid X_i = c_0] \tag{158}$$

which corresponds to the LATE at the cutoff c_0 (i.e. $\rho_{LATE} = E[Y_i^1 - Y_i^0 \mid X_i = c_0]$).

Regression Discontinuity

- Another version of the RD regression

$$Y_i = \alpha_{new} + \beta \tilde{X}_i + \rho D_i + \varepsilon_i \tag{159}$$

  where $\alpha_{new} = \alpha + \beta c_0$ and $\tilde{X}_i \equiv X_i - c_0$.
- Instead of linear function of X_i, a nonlinear function $f(X_i)$ can be used. For example,

$$Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + ... + \beta_p X_i^p + \rho D_i + \varepsilon_i \tag{160}$$

- Note that $f(X_i)$ can be nonparametric (thus a nonparametric estimator has to be used) and X_i can be also centered at c_0.

Regression Discontinuity

- For the *fuzzy RD*, treatment assignment is *stochastic*.
- Estimation of the causal treatment effect is

$$\rho_{FuzzyRD} = \frac{\lim_{X_i \to c_{0+}} E[Y_i \mid X_i = c_0] - \lim_{X_i \to c_{0-}} E[Y_i \mid X_i = c_0]}{\lim_{X_i \to c_{0+}} E[D_i \mid X_i = c_0] - \lim_{X_i \to c_{0-}} E[D_i \mid X_i = c_0]} \tag{161}$$

- There is a two-steps procedure:
  1. Estimation of the probability of treatment assignment D_i and interactions between D_i and $\tilde{X}_i$ (i.e. DX_i):

$$\begin{aligned} D_i &= \gamma_{00} + \gamma_{01}\tilde{X}_i + ... + \gamma_{0p}\tilde{X}_i^p + \eta_0 Z_i + \eta_{i0}, && (162)\\ DX_i &= \gamma_{01} + \gamma_{11}\tilde{X}_i + ... + \gamma_{1p}\tilde{X}_i^p + \eta_1 Z_i + \eta_{i1} && (163)\\ &... && (164)\\ DX_i^p &= ... && (165) \end{aligned}$$

     where Z_i is an excluded instrument (possibly including interaction terms $\tilde{X}_i Z_i$, $\tilde{X}_i^2 Z_i$, ..., $\tilde{X}_i^p Z_i$).
  2. Estimation of the treatment effect (using the fitted values from the 1st step):

$$Y_i = \alpha + \beta_1 \tilde{X}_i + ... + \beta_p \tilde{X}_i^p + \rho \hat{D}_i + \pi_1 \widehat{DX}_i + ... + \pi_p \widehat{DX}_i^p + \varepsilon_i \tag{166}$$

Instrumental Variables

- Consider the case of *constant/homogeneous treatment effect*, the switching equation and the causal regression model are respectively

$$Y_i = Y_i^0 + (Y_i^1 - Y_i^0)D_i \tag{167}$$
$$= \alpha + \rho D_i + \eta_i \tag{168}$$

- Consider the univariate IV problem with Y_i as earning, S_i as years of schooling (treatment assignment status), A_i as ability, Z_i as an excluded instrument for S_i. The regression model for *constant treatment effect* is

$$Y_i = \alpha + \rho S_i + \eta_i \tag{169}$$

  where η_i includes unobserved factor (ability A_i) which is correlated with S_i (for example, $\eta_i = \gamma A_i + \varepsilon_i$). Thus, $E[S_i \eta_i] \neq 0$.
- OLS estimator for this model is biased

$$\hat{\rho}_{OLS} = \frac{Cov(Y, S)}{V(S)} \tag{170}$$
$$= \rho + \underbrace{\gamma \frac{Cov(A, S}{V(S)}}_{\text{Omitted Variable Bias}} \tag{171}$$

Instrumental Variables

- The IV schema is

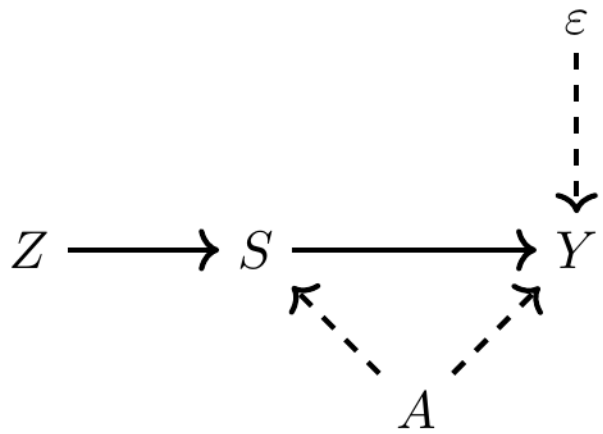


Figure: IV directed graph. Source: Cunningham (2021).

- The IV estimator:

$$\hat{\rho}_{IV} = \frac{Cov(Z,Y)}{Cov(Z,S)} \tag{172}$$

$$= \frac{\rho Cov(Z,S) + \gamma Cov(Z,A) + Cov(Z,\varepsilon)}{Cov(Z,S)} \tag{173}$$

$$= \rho + \frac{\gamma Cov(Z,A)}{Cov(Z,S)} + \frac{Cov(Z,\varepsilon)}{Cov(Z,S)} \tag{174}$$

Instrumental Variables

- If $Cov(A, Z) = 0$ and $Cov(Z, \varepsilon) = 0$ (Z satisfies the *exclusion condition*), then $\text{plim}\hat{\rho}_{IV} = \rho$.
- The IV can be rewritten as

$$\hat{\rho}_{IV} = \frac{Cov(Z, Y)}{Cov(Z, S)} \tag{175}$$

$$= \frac{Cov(Z, Y)/V(Z)}{Cov(Z, S)/V(Z)} \tag{176}$$

$$= \frac{\hat{\beta} Cov(Z, Y)}{\hat{\beta}^2 V(Z)} \quad \text{using } \hat{\beta} = Cov(Z, S)/V(Z) \tag{177}$$

$$= \frac{Cov(\hat{\beta} Z, Y)}{V(\hat{\beta} Z)} \tag{178}$$

- Note that $\hat{S} = \hat{\beta} Z$ is the fitted values of the 1st-stage regression of S_i on Z_i.
- Hence, the IV estimator is the OLS estimator of the following 2nd stage regression (2SLS estimator):

$$\hat{\rho}_{IV} = \frac{Cov(\hat{S}, Y)}{V(\hat{S})}. \tag{179}$$

Instrumental Variables

- For *heterogenous treatment effect*, the switching equation and the causal regression equation are

$$Y_i = Y_i^0 + (Y_i^1 - Y_i^0)D_i \tag{180}$$
$$= \alpha + \rho_i D_i + \eta_i \tag{181}$$

- Suppose that instrument Z_i is binary, the potential outcomes for treatment status are

$$D_i = \begin{cases} D_i^1 & \text{if } Z_i = 1 \\ D_i^0 & \text{if } Z_i = 0 \end{cases} \tag{182}$$

- The treatment status switching equation is

$$D_i = D_i^0 + (D_i^1 - D_i^0)Z_i \tag{183}$$
$$= \pi_0 + \pi_{1,i} Z_i + v_i \tag{184}$$

where π_{i1} is the heterogeneous causal effect of Z on D.

Instrumental Variables

- The IV estimator much satisfy the following assumptions
  1. SUTVA
  2. CIA: $\{Y_i(D_i^1, Z_i = 1), Y_i(D_i^0, Z_i = 0), D_i^1, D_i^0\} \perp Z_i$
  3. Exclusion restriction (outcome is independent of instrument): $Y_i(D_i, Z_i = 1) = Y_i(D_i, Z_i = 0)$ for $D_i = 0, 1$.
  4. Monotonicity: $\pi_{1,i}$ is the same sign for all i.
- Under these assumption, the IV estimator is the LATE of D on Y:

$$\rho_{IV,LATE} = \frac{\text{Effect of } Z \text{ on } Y}{\text{Effect of } Z \text{ on } D} \tag{185}$$

$$= E[Y_i^1 - Y_i^0 \mid D_i^1 - D_i^0 = 1] \tag{186}$$

- It can be obtained by 2SLS.

Instrumental Variables

- Note that
  - *Compliers*. Subpopulation whose treatment status is affected by the instrument in the correct direction: $D_i^1 = 1$ and $D_i^0 = 0$.
  - *Defiers*. Subpopulation whose treatment status is affected by the instrument in the wrong direction: $D_i^1 = 0$ and $D_i^0 = 1$.
  - *Never takers*. Subpopulation that never take the treatment: $D_i^1 = D_i^0 = 0$.
- The IV-LATE estimator is only for the compliers.
- Some advices:
  1. When there are covariates, put the same exogenous covariates in the 1st and the 2nd stage regressions.
  2. Forbidden regressions: don't (i) apply nonlinear regression to the first-stage (as D_i is binary) and (ii) use the nonlinear fitted values in the 2nd stage regression: nothing guarantees that the nonlinear fitted values are not correlated with the error term. $\Rightarrow$ Use noninear fitted values as instrument for the causal model in (181).
  3. Report details on the 1st-stage regression (F-stat), check for exogeneity, overidentifying restriction.

Differences-in-Differences

- Consider the example of an increase in minimum wage in New Jersey (from \$4.25 to \$5.05 in November 1992) while in Pennsylvania, it remains at \$4.25 (Card and Krueger, 1994). Data on fast-food employment (at restaurant i in state s and period t) are collected in February and November 1992 in NJ and PA.
- The causal regression model for the *conditional mean function*, assuming *constant treatment effect*, is

$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \varepsilon_{ist} \tag{187}$$

where $D_s t$ is a dummy for high-minimum wage states and periods (i.e. New Jersey and November 1992).

- We have

$$\begin{aligned} E[Y_{ist} \mid s = PA, t = Nov] - E[Y_{ist} \mid s = PA, t = Feb] &= \lambda_{Nov} - \lambda_{Feb} \\ E[Y_{ist} \mid s = NJ, t = Nov] - E[Y_{ist} \mid s = NJ, t = Feb] &= \lambda_{Nov} - \lambda_{Feb} + \delta \end{aligned}$$

Differences-in-Differences

- Therefore, the causal effect $E[Y^1_{ist} - Y^0_{ist} \mid s, t]$ (ATT) is

$$\begin{aligned} \hat{\delta}_{ATT} &= (E[Y_{ist} \mid s = NJ, t = Nov] - E[Y_{ist} \mid s = NJ, t = Feb]) - \\ &\quad (E[Y_{ist} \mid s = PA, t = Nov] - E[Y_{ist} \mid s = PA, t = Feb]) \end{aligned} \tag{188}$$

- Or equivalently,

$$\hat{\delta}_{ATT} = \underbrace{(E[Y_{i,NJ} \mid Post] - E[Y_{i,NJ} \mid Pre])}_{\triangle_{NJ}} - \tag{189}$$

$$\underbrace{(E[Y_{i,PA} \mid Post] - E[Y_{i,PA} \mid Pre])}_{\triangle_{PA}} \tag{190}$$

- The corresponding standard errors (if large sample with *iid* observations) are

$$SE_\delta = \sqrt{\frac{S(\triangle_{NJ})}{N_{NJ}} + \frac{S(\triangle_{PA})}{N_{PA}}}, \tag{191}$$

where $S(s)$ are estimated variance of group s.

Differences-in-Differences

- A typical DD regression equation is

$$Y_{it} = \alpha + \gamma D_i + \lambda Post_t + \delta(D_i \times Post_t) + X_{it}'\gamma + \varepsilon_{it} \tag{192}$$

or, equivalently

$$Y_{i,Post} - Y_{i,Pre} = \alpha + \delta D_i + \varepsilon_i. \tag{193}$$

- The DD regression equation for the study on minimum wage (Card and Krueger, 1994) is

$$Y_{its} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ_s \times d_t) + \varepsilon_{its} \tag{194}$$

NJ_s: dummy for treatment status (=1 if NJ, 0 if PA)
d_t: dummy for Nov. 1992 (post treatment period).

- The outcomes are then

$$\begin{aligned} PA\ Pre &= \alpha \\ PA\ Post &= \alpha + \lambda \\ NJ\ Pre &= \alpha + \gamma \\ NJ\ Post &= \alpha + \gamma + \lambda + \delta \end{aligned}$$

Differences-in-Differences

- The link between the causal model for conditional mean function and the DD regression equation is

$$
\begin{aligned}
\alpha &= E[Y_{ist} \mid s = PA, t = Feb] = \gamma_{PA} + \lambda_{Feb} \\
\gamma &= E[Y_{ist} \mid s = NJ, t = Feb] - E[Y_{ist} \mid s = PA, t = Feb] = \gamma_{NJ} - \gamma_{PA} \\
\lambda &= E[Y_{ist} \mid s = PA, t = Nov] - E[Y_{ist} \mid s = PA, t = Feb] = \lambda_{Nov} - \lambda_{Feb} \\
\delta &= (E[Y_{ist} \mid s = NJ, t = Nov] - E[Y_{ist} \mid s = NJ, t = Feb]) \\
&\quad - (E[Y_{ist} \mid s = PA, t = Nov] - E[Y_{ist} \mid s = PA, t = Feb])
\end{aligned}
$$

Differences-in-Differences

- *Parallel Trend Assumption*: If there was no treatment effect, the treatment group and the control group follow a similar trend prior to treatment (*counterfactual* trend).
- The decomposition of the DD estimate $\hat{\delta}$ is

$$\begin{aligned}\hat{\delta} &= \underbrace{\left(E[Y^1_{NJ} \mid Post] - E[Y^0_{NJ} \mid Post]\right)}_{ATT} \\ &\quad + \underbrace{\left(E[Y^0_{NJ} \mid Post] - E[Y^0_{NJ} \mid Pre]\right) - \left(E[Y^0_{PA} \mid Post] - E[E^0_{PA} \mid Pre]\right)}_{\text{Non-parallel trend bias}}.\end{aligned}$$

- With the parallel trend assumption, the estimate $\hat{\delta}$ corresponds to ATT.

Differences-in-Differences

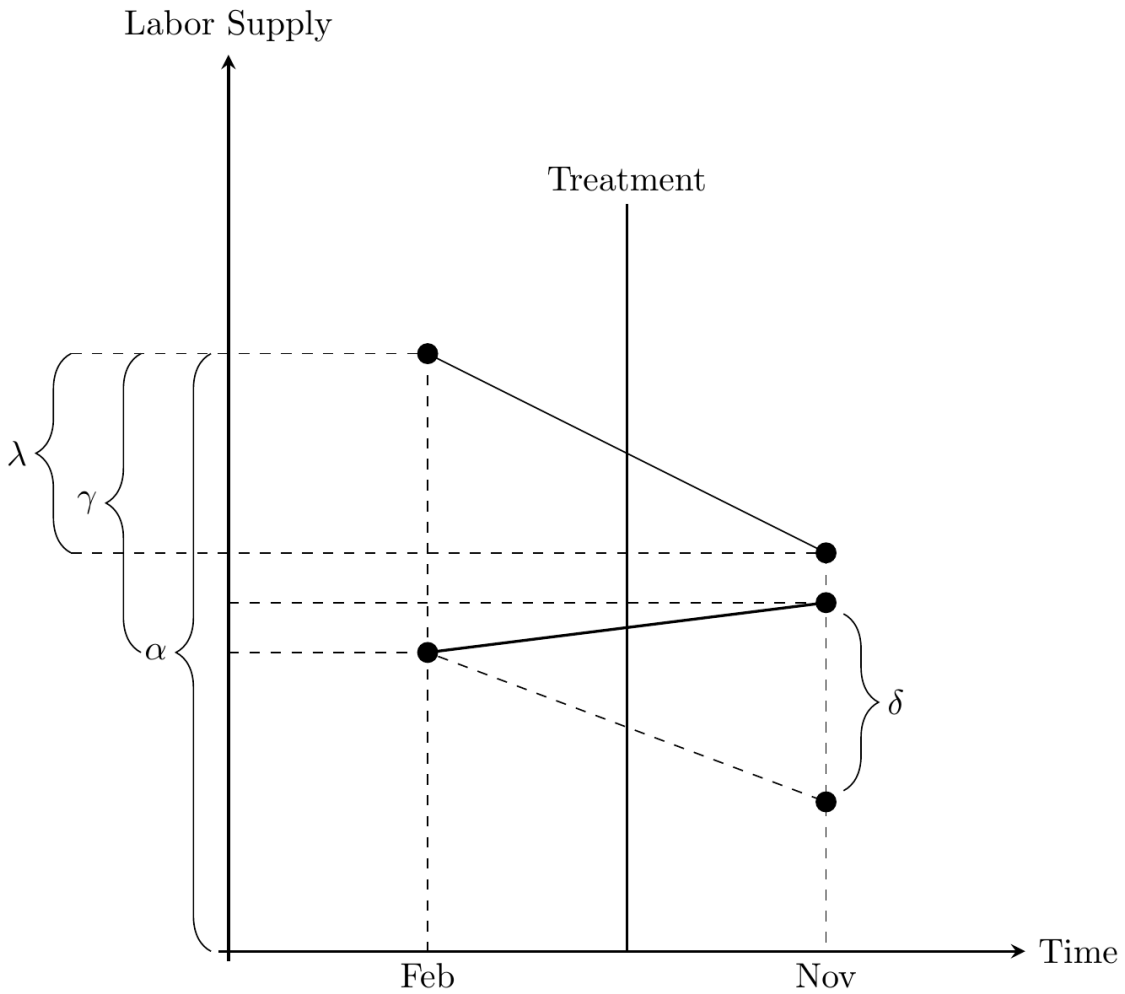


Figure: DiD regression diagram. Source: Cunningham (2021).

Differences-in-Differences

- Clustering standard errors at the group level (or level of treatment)
- Agregating the data into one pre and one post treatment period
- Checking parallel trend (pre-treatment balance between treatment and control groups)
- Placebo falsification: using alternative data (e.g. workers whose wages are not affected by minimum wage)

Differences-in-Differences

Differential timing

- Units receive treatment at different points in time.
- The DD regression equation is

$$y_{it} = \alpha_0 + \delta D_{it} + X_{it}'\gamma + \alpha_i + \lambda_t + \varepsilon_{it} \tag{195}$$

- This is a twoway fixed effect model (individual fixed effects α_i and time fixed effects λ_t).
- D_{it} is the indicator for treatment when units get treated.
- Several approaches, including Callaway and Sant'Anna (2020) solution (`did` package).

Synthetic Control

- Developed by Abadie & Gardeazabal (2003). Extension by Abadie, Diamond & Hainmueller (2010)
- Used for comparative case studies
- Particularly useful for policy evaluation
- Treatment affects single unit (e.g., region, country)
- Multiple control units available
- Traditional methods may not work well
- Need for systematic comparison approach
- Advantages: (i) Data-driven approach, (ii) Transparency in control selection, (iii) Reduces discretion in comparison unit choice, (iv) Allows for time-varying effects.

Synthetic Control

- One treated unit
- Multiple potential control units ("donor pool")
- Pre-intervention period
- Post-intervention period
- Key elements:
  - Outcome variable of interest
  - Predictors of outcome
  - Pre-treatment characteristics
  - Weights for control units

Synthetic Control

- Basic model
  - Let $i = 1$ be the treated unit and $i = 2, ..., J + 1$ be control units
  - Time periods $t = 1, ..., T_0, ..., T$ where T_0 is treatment time
  - Observed outcome:

$$Y_{1t} = Y_{1t}^N + \alpha_{1t} D_{1t} \tag{196}$$

where:

Y_{1t}^N = potential outcome without treatment
α_{1t} = treatment effect at time t
D_{1t} = treatment indicator

- Synthetic control estimator:

$$Y_{1t}^N = \sum_{j=2}^{J+1} w_j Y_{jt} \tag{197}$$

- Constraints:

$$\begin{aligned} w_j &\geq 0 \text{ for } j = 2, ..., J + 1 \\ \sum_{j=2}^{J+1} w_j &= 1 \end{aligned}$$

Synthetic Control

- Optimization problem:

$$\min_{W}(X_1 - X_0 W)' V (X_1 - X_0 W) \tag{198}$$

where

X_1: vector of pre-treatment characteristics for treated unit
X_0: matrix of pre-treatment characteristics for control units
V: weighting matrix for predictors
W: vector of weights

- Apply optimal weights: calculate $\hat{Y}_{1t}^N = \sum_{j=2}^{J+1} w_j^* =, \; t = 1, ..., T$.
- Treatment effect estimation: $\hat{\alpha}_{1t} = Y_{1t} - \hat{Y}_{1t}^N$, for $t > T_0$.

Synthetic Control: Assumptions

- No interference (SUTVA):
  - Treatment of unit 1 does not affect others
  - No spillover effects
  - No general equilibrium effects
- No anticipation:
  - No behavioral changes before T_0
  - No announcement effects
  - Clean pre-treatment period
- Convex Hull condition:
  - No extrapolation outside donor pool
  - Linear combination sufficient
  - Comparable units available
- Stable relationships:
  - No structural breaks
  - Consistent relationships
  - Stable control units

Synthetic Control: Technical Requirements

- Pre-treatment fit:
  - Good prediction pre-T_0
  - Small MSPE where

$$MSPE = \frac{1}{T_0} \sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \right)^2 \tag{199}$$

where:

Y_{1t}: outcome for treated unit at time t
w_j^*: optimal weights
Y_{jt}: outcomes for control units
T_0: pre-treatment periods

  - Balance in predictors
- Data quality:
  - No missing values
  - Consistent measurement
  - Reliable sources
- Time dimension:
  - Sufficient pre-periods
  - Regular intervals
  - Balanced panel

Contact: pnguyenvan@parisnanterre.fr