G023: Econometrics

Jérôme Adda

j.adda@ucl.ac.uk

Office # 203

I am grateful to Andrew Chesher for giving me access to his G023 course notes on which most of these slides are based.

Syllabus

Course Description:

This course is an intermediary econometrics course. There will be 3 hours of lectures per week and a class (sometimes in the computer lab) each week. Previous knowledge of econometrics is assumed. By the end of the term, you are expected to be at ease with modern econometric techniques. The computer classes introduce you to real life problems, and will help you to understand the theoretical content of the lectures. You will also learn to use a powerful and widespread econometric software, STATA.

Understanding these techniques will be of great help for your thesis over the summer, and will help you in your future workplace.

For any contact or query, please send me an email or visit my web page at:

`http://www.ucl.ac.uk/∼uctpjea/teaching.html`.

My web page contains documents which might prove useful such as notes, previous exams and answers.

Books:

There are a several good intermediate econometric books but the main book to be used for reference is the Wooldridge (J. Wooldridge (2003) Econometric Analysis of Cross-Section and Panel Data, MIT Press). Other useful books are:

- Andrew Chesher's notes, on which most of these slides are based.
- Gujurati "Basic Econometrics", Mc Graw-Hill. (Introductory text book)
- Green "Econometric Analysis", Prentice Hall International, Inc. (Intermediate text book)

Course Content

1. Introduction
   What is econometrics? Why is it useful?

2. The linear model and Ordinary Least Squares
   Model specification.

3. Hypothesis Testing
   Goodness of fit, R^2. Hypothesis tests (t and F).

4. Approximate Inference
   Slutsky's Theorem; Limit Theorems. Approximate distribution of the OLS and GLS estimators.

5. Maximum Likelihood Methods
   Properties; Limiting distribution; Logit and Probit; Count data.

6. Likelihood based Hypothesis Testing
   Wald and Score tests.

7. Endogeneity and Instrumental Variables
   Indirect Least Squares, IV, GMM; Asymptotic properties.

Definition and Examples

Econometrics: statistical tools applied to economic problems.

Examples: using data to:

- Test economic hypotheses.
- Establish a link between two phenomenons.
- Assess the impact and effectiveness of a given policy.
- Provide an evaluation of the impact of future public policies.

Provide a qualitative but also a quantitative answer.

Example 1: Global Warming

- Measuring the extent of global warming.
  - when did it start?
  - How large is the effect?
  - has it increased more in the last 50 years?

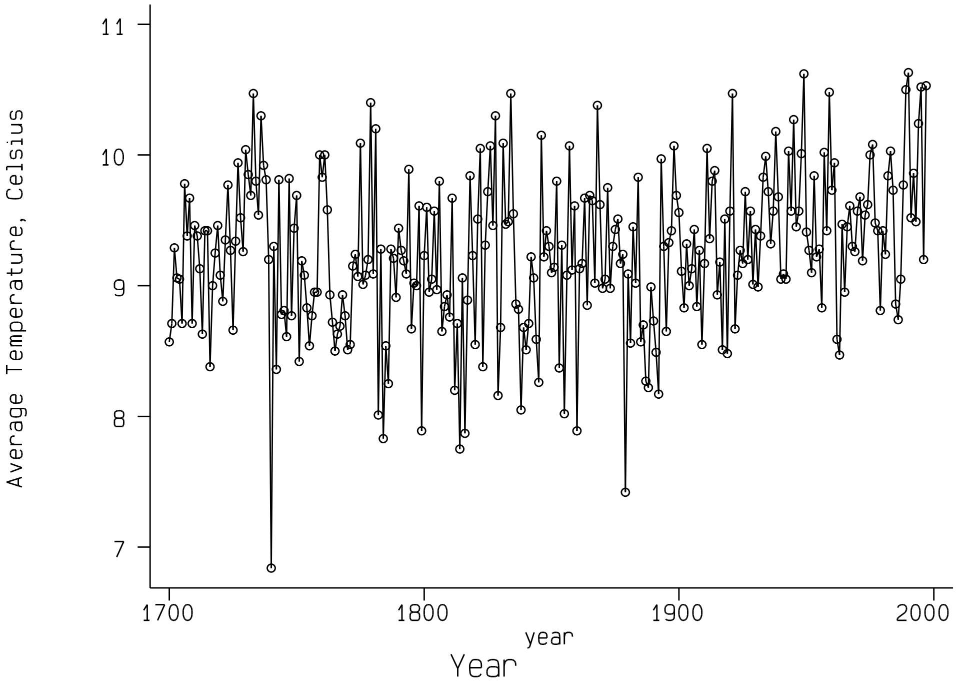
- What are the causes of global warming?
  - Does carbon dioxide *cause* global warming?
  - Are there other determinants?
  - What are their respective importance?

- Average temperature in 50 years if nothing is done?

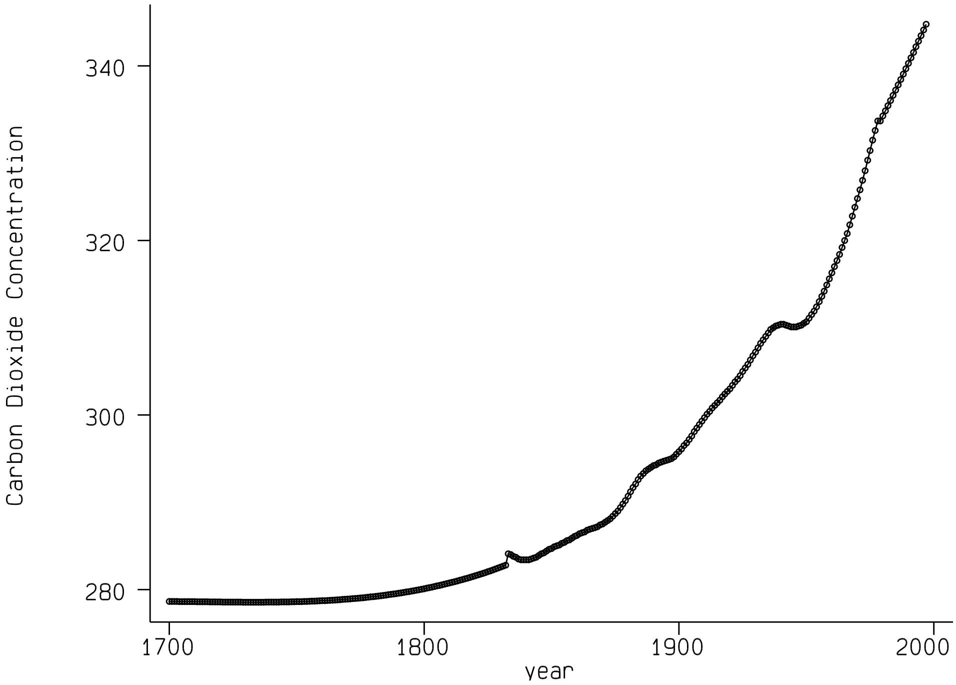
- Average temperature if carbon dioxide concentration is reduced by 10%?

Example 1: Global Warming

Average Temperature in Central England (1700-1997)



Atmospheric Concentration of Carbon Dioxide (1700-1997)



Causality

- We often observe that two variables are correlated.
  - Examples:
    - Individuals with higher education earn more.
    - Parental income is correlated with child's education.
    - Smoking is correlated with peer smoking.
    - Income and health are correlated.

- However this does NOT establish causal relationships.

Causality

- If a variable Y is causally related to X, then changing X will LEAD to a change in Y.
  - For example: Increasing VAT may cause a reduction of demand.
  - Correlation may **not** be due to causal relationship:
    - Part or the whole correlation may be induced by both variables depending on some common factor and does not imply causality.
    - For example: Individuals who smoke may be more likely to be found in similar jobs. Hence, smokers are more likely to be surrounded by smokers, which is usually taken as a sign of peer effects. The question is how much an increase in smoking by peers results in higher smoking.
    - Brighter people have more education AND earn more. The question is how much of the increased in earnings is caused by the increased education.

Linear Model

- We write the linear model as:

$$y = X\beta + \varepsilon$$

  where ε is a nx1 vector of values of the unobservable.

- X is a nxk vector of regressors (or explanatory variables).

- y is the dependent variable and is a vector.

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad X = \begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{k-1} \end{bmatrix}$$

Model Specifications

- **Linear model:**

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$\frac{\partial Y_i}{\partial X_i} = \beta_1$$

Interpretation: When X goes up by 1 unit, Y goes up by β_1 units.

- **Log-Log model** (constant elasticity model):

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + \varepsilon_i$$

$$Y_i = e^{\beta_0} X_i^{\beta_1} e^{\varepsilon_i}$$

$$\begin{aligned}\frac{\partial Y_i}{\partial X_i} &= e^{\beta_0}\beta_1 X_i^{\beta_1 - 1} e^{\varepsilon_i} \\ \frac{\partial Y_i / Y_i}{\partial X_i / X_i} &= \beta_1\end{aligned}$$

Interpretation: When X goes up by **1%**, Y goes up by β_1 **%**.

- **Log-lin model:**

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\begin{aligned}\frac{\partial Y_i}{\partial X_i} &= \beta_1 e^{\beta_0} e^{\beta_1 X_i} e^{\varepsilon_i} \\ \frac{\partial Y_i / Y_i}{\partial X_i} &= \beta_1\end{aligned}$$

Interpretation: When X goes up by 1 **unit**, Y goes up by $100\beta_1$ **%**.

Example: Global Warming

Dependent	Temperature		Log Temperature	
variable:	(Celsius)			
CO2 (ppm)	0.0094 (.0018)	-	0.00102 (.0002)	-
Log CO2	-	2.85 (.5527)	-	0.31 (0.0607)
Constant	6.47 (.5345)	-6.94 (3.13)	1.92 (.0.5879)	0.46 (0.3452)

- An increase in 1ppm in CO2 raises temperature by 0.0094 degrees Celsius. Hence, since 1700 a raise in about 60ppm leads to an increase in temperature of about 0.5 Celsius.

- A one percent increase in CO2 concentration leads to an increase of 0.0285 degrees.

- An increase of one ppm in CO2 concentration leads to an increase in temperature of 0.1%.

- A 1% increase in CO2 concentration leads to a 0.3% increase in temperature.

Assumptions of the Classical Linear Regression Model

- **Assumption 1:** $E[\varepsilon|X] = 0$
  - The expected value of the error term has mean zero *given any value of the explanatory variable.* Thus observing a high or a low value of X does not imply a high or a low value of ε.

    X and ε are uncorrelated.
  - This implies that changes in X are not associated with changes in ε in any particular direction - Hence the associated changes in Y can be attributed to the impact of X.
  - This assumption allows us to interpret the estimated coefficients as reflecting causal impacts of X on Y.
  - Note that we condition on the whole set of data for X in the sample not on just one .

Assumptions of the Classical Linear Regression Model

- **Assumption 2:** $\text{rank}(X) = k$.
- In this case, for all non-zero $k \times 1$ vectors, c, $Xc \neq 0$.
- When the rank of X is less than k, there exists a non-zero vector c such that $Xc = 0$. In words, there is a linear combination of the columns of X which is a vector of zeros. In this situation the OLS estimator cannot be calculated. β cannot be defined by using the information contained in X.
- Perhaps one could obtain other values of x and then be in a position to define β. But sometimes this is not possible, and then β is not *identifiable* given the information in X. Perhaps we could estimate functions (e.g. linear functions) of β that would be identifiable even without more x values.

OLS Estimator

- Assumption 1 state that $E[\varepsilon|X]=0$ which implies that:

$$\begin{aligned} E[y-\beta X|X] &= 0 \\ X'E[y-\beta X|X] &= E[X'(y-X\beta)|X] \\ &= E[X'y]-X'X\beta \\ &= 0 \end{aligned}$$

- and so, given that $X'X$ has full rank (Assumption 2):

$$\beta=(X'X)^{-1}E[X'y|X]$$

- Replacing $E[X'y|X]$ by $X'y$ leads to the Ordinary Least Square estimator:

$$\hat{\beta}=(X'X)^{-1}X'y$$

- Note that we can write the estimator as:

$$\begin{aligned} \hat{\beta} &= (X'X)X'(X\beta+\varepsilon) \\ &= \beta+(X'X)^{-1}X'\varepsilon \end{aligned}$$

Properties of the OLS Estimator

- Variance of the OLS estimator:

$$\begin{aligned}
Var(\hat{\beta}|X) &= E[\left(\hat{\beta}-E[\hat{\beta}|X]\right)\left(\hat{\beta}-E[\hat{\beta}|X]\right)'|X] \\
&= E[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)'|X] \\
&= E[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}|X] \\
&= (X'X)^{-1}X'E[\varepsilon\varepsilon'|X]X(X'X)^{-1} \\
&= (X'X)^{-1}X'\Sigma X(X'X)^{-1}
\end{aligned}$$

  where $\Sigma = Var[\varepsilon|X]$

- If $\Sigma = \sigma^2 I_n$ (homoskedasticity and no autocorrelation) then

$$Var(\hat{\beta}|X) = \sigma^2(X'X)^{-1}$$

- If we are able to get an estimator of σ^2:

$$\widehat{Var}(\hat{\beta}|X) = \hat{\sigma}^2(X'X)^{-1}$$

- We can re-write the variance as:

$$V\hat{a}r(\hat{\beta}|X) = \frac{\hat{\sigma}^2}{n}\left(\frac{(X'X)}{n}\right)^{-1}$$

  We can expect $(X'X)/n$ to remain fairly constant as the sample size n increases. Which means that we get more accurate OLS estimators in larger samples.

Alternative Way

- The OLS estimator is also defined as

$$\min_{\beta} \frac{\varepsilon'\varepsilon}{n} = \min_{\beta}(y - X\beta)'(y - X\beta)$$

- The first order conditions for this problem are:

$$X'(y - X\hat{\beta}) = 0$$

  This is a kx1 system of equation defining the OLS estimator.

Goodness of Fit

- We measure how well the model fits the data using the R^2.
- This is the ratio of the explained sum of squares to the total sum of squares
  - Define the Total sum of Squares as: $TSS = \sum_{i=1}^{N} (Y_i - \bar{Y})^2$
  - Define the Explained Sum of Squares as: $ESS = \sum_{i=1}^{N} [\hat{\beta}(X_i - \bar{X})]^2$
  - Define the Residual Sum of Squares as: $RSS = \sum_{i=1}^{N} \hat{\varepsilon}_i^2$
- Then we define

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

- This is a measure of how much of the variance of Y is explained by the regressor X.
- The computed R^2 following an OLS regression is always between 0 and 1.
- A low R^2 is not necessarily an indication that the model is wrong - just that the included X have low explanatory power.
- The key to whether the results are interpretable as causal impacts is whether the explanatory variable is uncorrelated with the error term.

Goodness of Fit

- The R^2 is non decreasing in the number of explanatory variables.
- To compare two different model, one would like to adjust for the number of explanatory variables: adjusted R^2:

$$\bar{R}^2 = 1 - \frac{\sum_i \hat{\varepsilon}_i^2/(N-k)}{\sum_i y_i^2/(N-1)}$$

- The adjusted and non adjusted R^2 are related:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-k}$$

- Note that to compare two different R^2 the dependent variable must be the same:

$$\begin{aligned} \ln Y_i &= \beta_0 + \beta_1 X_i + u_i \\ Y_i &= \alpha_0 + \alpha_1 X_i + v_i \end{aligned}$$

cannot be compared as the Total Sum of Squares are different.

Alternative Analogue Estimators

- Let H be a nxk matrix containing elements which are functions of the elements of X:

$$\begin{aligned} E[H'\varepsilon|X] &= 0 \\ E[H'(y - X\beta)|X] &= 0 \\ E[H'y|X] - E[H'X|X]\beta &= 0 \\ E[H'y|X] - (H'X)\beta &= 0 \end{aligned}$$

- If the matrix $H'X$ has full rank k then

$$\beta = (H'X)^{-1}E[H'y|X]$$

$$\hat{\beta}_H = (H'X)^{-1}H'y$$

$$Var(\hat{\beta}_H|X) = (H'X)^{-1}H'\Sigma H(X'H)^{-1}$$

with $\Sigma = Var(\varepsilon|X)$. If we can write $\Sigma = \sigma^2 I_n$ where I_n is a nxn identity matrix then:

$$Var(\hat{\beta}_H|X) = \sigma^2(H'X)^{-1}H'H(X'H)^{-1}$$

- Different choices of H leads to different estimators. We need a criteria that ranks estimators. Usually the estimator with the smallest variance.

Misspecification

- Suppose the true model is not linear but take the following (more general) form:

$$E[Y|X=x] = g(x,\theta)$$

so that

$$Y = g(x,\theta) + \varepsilon \qquad E[\varepsilon|X] = 0$$

Define

$$G(X,\theta) = \begin{bmatrix} g(x_1,\theta) \\ \vdots \\ g(x_n,\theta) \end{bmatrix}$$

then

$$\begin{aligned} E[\hat{\beta}|X] &= E[(X'X)^{-1}X'y|X] \\ &= (X'X)^{-1}X'G(X,\theta) \\ &\neq \beta \end{aligned}$$

- The OLS estimator is biased. The bias depends on the values of x and the parameters θ. Different researches faced with different values of x will come up with different conclusions about the value of β if they use a linear model.

- The variance of the OLS estimator is:

$$\begin{aligned} Var(\hat{\beta}|X) &= E[\left(\hat{\beta} - E[\hat{\beta}|X]\right)\left(\hat{\beta} - E[\hat{\beta}|X]\right)'|X] \\ &= E[(X'X)^{-1}X'(y - G(X,\theta))(y - G(X,\theta))'X(X'X)^{-1}|X] \\ &= E[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}|X] \\ &= (X'X)^{-1}X'\Sigma X(X'X)^{-1} \end{aligned}$$

exactly as it is when the regression function is correctly specified.

Omitted Regressors

- Suppose the true model generating y is:

$$y = Z\gamma + \varepsilon \qquad E[\varepsilon|X, Z] = 0$$

- Consider the OLS estimator $\hat{\beta} = (X'X)^{-1}X'y$ calculated using data X, where X and Z may have common columns.

$$\begin{aligned} E[\hat{\beta}|X] &= E[(X'X)^{-1}X'y|X, Z] \\ &= E[(X'X)^{-1}X'(Z\gamma + \varepsilon)|X, Z] \\ &= (X'X)^{-1}X'Z\gamma + (X'X)^{-1}X'E[\varepsilon|X, Z] \\ &= (X'X)^{-1}X'Z\gamma \end{aligned}$$

- Let $Z = [X \;\vdots\; Q]$ and $\gamma' = [\gamma'_X \;\vdots\; \gamma'_Q]$ so that the matrix X used to calculate $\hat{\beta}$ is a part of the matrix Z. In the fitted model, the variables Q have been omitted.

$$\begin{aligned} E[\hat{\beta}|X, Z] &= E[\hat{\beta}|Z] \\ &= (X'X)^{-1}X'Z\gamma \\ &= (X'X)^{-1}[X'X \;\vdots\; X'Q]\gamma \\ &= [I \;\vdots\; (X'X)^{-1}X'Q]\gamma \\ &= \gamma_X + (X'X)^{-1}X'Q\gamma_Q \end{aligned}$$

- If $X'Q = 0$ or $\gamma_Q = 0$ then $E[\hat{\beta}|X, Z] = \gamma_X$. In other words, omitting a variable from a regression bias the coefficients unless the omitted variable is uncorrelated with the other explanatory variables.

Omitted Regressors: Example

- Health and income in a sample of Swedish individuals.
- Relate Log income to a measure of overweight (body mass index).

	Log Income	
BMI low	-0.42 (.016)	-0.15 (.014)
BMI high	-0.00 (.021)	-0.12 (.018)
age		0.13 (.0012)
age square		-0.0013 (.00001)
constant	6.64 (.0053)	3.76 (.0278)

- Are obese individuals earning less than others ?
- Obesity, income and age are related:

Age	Log income	Prevalence of Obesity
<20	4.73	0.007
20-40	6.76	0.033
40-60	7.01	0.0759
60 and over	6.34	0.084

Measurement Error

- Data is often measured with error.
  - reporting errors.
  - coding errors.
- The measurement error can affect either the dependent variable or the explanatory variables. The effect is dramatically different.

Measurement Error on Dependent Variable

- Y_i is measured with error. We assume that the measurement error is **additive and not correlated with** X_i.

- We observe $\check{Y} = Y + \nu$. We regress $\check{Y}$ on X:

$$\begin{aligned}
\check{Y} &= X\beta + \varepsilon \\
Y &= X\beta + \varepsilon - \nu \\
&= X\beta + w
\end{aligned}$$

- The assumptions we have made for OLS to be unbiased and BLUE are *not* violated. **OLS estimator is unbiased**.

- The variance of the slope coefficient is:

$$\begin{aligned}
V\hat{a}r(\hat{\beta}) &= Var(w)(X'X)^{-1} \\
&= Var(\varepsilon - \nu)(X'X)^{-1} \\
&= [Var(\varepsilon) + Var(\nu)](X'X)^{-1} \\
&\geq Var(\varepsilon)(X'X)^{-1}
\end{aligned}$$

- The variance of the estimator is larger with measurement error on Y.

Measurement Error on Explanatory Variables

- X is measured with errors. We assume that the error is **additive and not correlated with X:** $E[\nu|x] = 0$.
- We observe $\check{X} = X + \nu$ instead. The regression we perform is Y on $\check{X}$. The estimator of β is expressed as:

$$\begin{aligned}
\hat{\beta} &= (\check{X}'\check{X})^{-1}\check{X}'y \\
&= (X'X + \nu'\nu + X'\nu + \nu'X)^{-1}(X+\nu)'(X\beta + \varepsilon) \\
E[\hat{\beta}|X] &= (X'X + \nu'\nu)^{-1}X'X\beta
\end{aligned}$$

- Measurement error on X leads to a **biased OLS estimate**, biased towards zero. This is also called **attenuation bias**.

Example

- True model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \qquad \text{with} \qquad \beta_0 = 1 \;\; \beta_1 = 1$$

- X_i is measured with error. We observe $\tilde{X}_i = X_i + \nu_i$.
- Regression results:

	Var(ν_i)/Var(X_i)			
	0	0.2	0.4	0.6
β_0	1	1.08	1.28	1.53
β_1	2	1.91	1.7	1.45

Estimation of linear functions of β

- Sometimes we are interested in a particular combination of the elements of β, say $c'\beta$.
  - the first element of β: $c' = [1, 0, 0, \ldots, 0]$
  - the sum of the first two elements of β: $c' = [1, 1, 0 \ldots, 0]$.
  - the expected value of Y at $x = [x_1, \ldots, x_k]$ (which might be used in predicting the value of Y at those values: $c' = [x_1, \ldots x_k]$
- An obvious estimator of $c'\beta$ is $c'\hat{\beta}$ whose variance is:

$$Var(c'\hat{\beta}|X) = \sigma^2 c'(X'X)^{-1}c$$

Minimum Variance Property of OLS

- The OLS estimator possesses an optimality property when $Var[\varepsilon|X] = \sigma^2 I_n$, namely that among the class of *linear* functions of y that are *unbiased* estimators of β the OLS estimator has the smallest variance, in the sense that, considering any other estimator,

  $$\tilde{\beta} = Q(X)y$$

  (a linear function of y, with $Q(X)$ chosen so that $\tilde{\beta}$ is unbiased),

  $$Var(c'\tilde{\beta}) \geq Var(c'\hat{\beta}) \quad \text{for all } c$$

  This is known as the *Gauss-Markov theorem*.

  OLS is the best linear unbiased (BLU) estimator.

- To show this result, let

  $$Q(X) = (X'X)^{-1} X' + R'$$

  where R may be a function of X, and note that

  $$E[\tilde{\beta}|X] = \beta + R'X\beta.$$

  This is equal to β for all β only when $R'X = 0$. This condition is required if $\tilde{\beta}$ is to be a linear unbiased estimator. Imposing that condition,

  $$Var[\tilde{\beta}|X] - Var[\hat{\beta}|X] = \sigma^2 R'R,$$

  and

  $$Var(c'\tilde{\beta}) - Var(c'\hat{\beta}) = \sigma^2 d'd = \sigma^2 \sum_{i=1}^{k} d_i^2 \geq 0$$

  where $d = Rc$.

M Estimation

- Different strategy to define an estimator.
- Estimator that ”fits the data”.
- Not obvious that this is the most desirable goal. Risk of overfitting.
- One early approach to this problem was due by the French mathematician Laplace: least absolute deviation:

  $$\tilde{\beta} = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} |Y_i - b'x_i|$$

  The estimator is quite robust to measurement error but quite difficult to compute.
- Note that OLS estimator belongs to M estimators as it can be defined as:

  $$\hat{\beta} = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - b'x_i)^2$$

Frisch-Waugh Lovell Theorem

- Suppose X is partitioned into two blocks: $X = [X_1, X_2]$, so that
$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$
where β_1 and β_2 are elements of the conformable partition of β. Let
$$M_1 = I - X_1(X1'X_1)^{-1}X_1'$$
then $\hat{\beta}_2$ can be written:
$$\begin{aligned}\hat{\beta}_2 &= ((M_1X_2)'(M_1X_2))^{-1}(M_1X_2)'M_1y \\ \hat{\beta}_2 &= (X_2'M_1X_2)^{-1}X_2'M_1y\end{aligned}$$

- Proof: writing $X'y = (X'X)\hat{\beta}$ in partitioned form:
$$\begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix} = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$$
$$\begin{aligned} X_1'y &= X_1'X_1\hat{\beta}_1 + X_1'X_2\hat{\beta}_2 \\ X_2'y &= X_2'X_1\hat{\beta}_1 + X_2'X_2\hat{\beta}_2 \end{aligned}$$
so that
$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y - (X_1'X_1)^{-1}X_1'X_2\hat{\beta}_2$$
substituting
$$X_2'y - X_2'X_1(X_1'X_1)^{-1}X_1'y = X_2'X_2\hat{\beta}_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2\hat{\beta}_2$$
which after rearrangements is
$$X_2'M_1y = (X_2'M_1X_2)\hat{\beta}_2$$

- Interpretation: the term M_1X_2 is the matrix of residuals from the OLS estimation of X_2 on X_1. The term M_1y is the residuals of the OLS regression of y on X_1. So to get the OLS estimate of β_2 we can perform OLS estimation using residuals as left and right hand side variables.

Generalised Least Squares Estimation

- The simple result $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$ is true when $Var(\varepsilon|X) = \sigma^2 I_n$ which is independent of X.

- There are many situations in which we would expect to find some dependence on X so that $Var[\varepsilon|X] \neq \sigma^2 I_n$.

- Example: in a household expenditure survey we might expect to find people with high values of time purchasing large amounts infrequently (e.g. of food, storing purchases in a freezer) and poor people purchasing small amounts frequently. If we just observed households' expenditures for a week (as in the British National Food Survey) then we would expect to see that, conditional on variables X that are correlated with the value of time, the variance of expenditure depends on X.

- When this happens we talk of the disturbances, ε, as being heteroskedastic.

- In other contexts we might expect to find correlation among the disturbances, in which case we talk of the disturbances as being serially correlated.

Generalised Least Squares Estimation

- The BLU property of the OLS estimator does not usually apply when $Var[\varepsilon|X] \neq \sigma^2 I_n$.

- Insight: suppose that Y has a much larger conditional variance at one value of x, x^*, than at other values. Realisations produced at x^* will be less informative about the location of the regression function than realisations obtained at other values of x. It seems natural to give realisations obtained at x^* less weight when estimating the regression function.

- We know how to produce a BLU estimator when $Var[\varepsilon|X] = \sigma^2 I_n$.

- Our strategy for producing a BLU estimator when this condition does not hold is to transform the original regression model so that the conditional variance of the transformed Y is proportional to an identity matrix and apply the OLS estimator in the context of that transformed model.

Generalised Least Squares Estimation

- Suppose $Var[\varepsilon|X] = \Sigma$ is positive definite.
- Then we can find a matrix P such that $P\Sigma P' = I$. Let Λ be a diagonal matrix with the (positive valued) eigenvalues of Σ on its diagonal, and let C be the matrix of associated orthonormal eigenvectors. Then $C\Sigma C' = \Lambda$ and so $\Lambda^{-1/2}C\Sigma C'\Lambda^{-1/2} = I$. The required matrix P is $\Lambda^{-1/2}C$.

  $$z = Py = PX\beta + u$$

  where $u = P\varepsilon$ and $Var[u|X] = I$
- In the context of this model the OLS estimator,

  $$\breve{\beta} = (X'P'PX)^{-1}X'P'Py,$$

  does possess the BLU property.
- Further, its conditional variance given X is $(X'P'PX)^{-1}$. Since $P\Sigma P' = I$, it follows that $\Sigma = P^{-1}P'^{-1} = (P'P)^{-1}$, so that $P'P = \Sigma^{-1}$. The estimator $\breve{\beta}$, and its conditional mean and variance can therefore be written as

  $$\begin{aligned} \breve{\beta} &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \\ E[\breve{\beta}|X] &= \beta \\ Var[\breve{\beta}|X] &= (X'\Sigma^{-1}X)^{-1} \end{aligned}$$

  The estimator is known as the *generalised least squares* (GLS) estimator.

Feasible Generalised Least Squares Estimation

- Obviously the estimator cannot be calculated unless Σ is known which is rarely the case. However sometimes it is possible to produce a well behaved estimator $\hat{\Sigma}$ in which case the *feasible GLS estimator* is:

  $$\breve{\beta} = (X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}y$$

  could be used.

- To study the properties of this estimator requires the use of asymptotic approximations and we return to this later.

Feasible GLS

- To produce the feasible GLS estimator we must impose some structure on the variance matrix of the unobservables, Σ.

- If not we would have to estimate $n(n+1)/2$ parameters using data containing just n observations: infeasible.

- One way to proceed is to impose the restriction that the diagonal elements of Σ are constant and allow nonzero off diagonal elements but only close to the main diagonal of Σ. This requires ε to have homoskedastic variation with X but allows a degree of correlation between values of ε for observations that are close together (e.g. in time if the data are in time order in the vector y).

- One could impose a parametric model on the variation of elements of Σ. You will learn more about this in the part of the course dealing with time series.

- Heteroskedasticity: a parametric approach is occasionally employed, using a model that requires $\sigma_{ii} = f(x_i)$. For example with the model $\sigma_{ii} = \gamma' x_i$ one could estimate γ, for example by calculating an OLS estimator of γ in the model with equation

  $$\hat{\varepsilon}_i^2 = \gamma' x_i + u_i$$

  where $\hat{\varepsilon}_i^2$ is the squared ith residual from an OLS estimation. Then an estimate of Σ could be produced using $\hat{\gamma}' x_i$ as the ith main diagonal element.

Feasible GLS

- Economics rarely suggests suitable parametric models for variances of unobservables.
- One may therefore not wish to pursue the gains in efficiency that GLS in principle offers.
- If the OLS estimator is used and $\Sigma \neq \sigma^2 I_n$ one must still be aware that the formula yielding standard errors, $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$ is generally *incorrect*. The correct one is:

  $$Var(\hat{\beta}) = (X'X)^{-1}X'\Sigma X(X'X)^{-1}.$$

- One popular strategy is to proceed with the OLS estimator but to use an estimate of the matrix $(X'X)^{-1}X'\Sigma X(X'X)^{-1}$ to construct standard errors.
- In models in which the off-diagonal elements of Σ are zero but heteroskedasticity is potentially present this can be done by using

  $$Var(\hat{\beta}) = (X'X)^{-1}X'\hat{\Sigma}X(X'X)^{-1}.$$

  where $\hat{\Sigma}$ is a diagonal matrix with squared OLS residuals, $\hat{\varepsilon}_i^2$, on its main diagonal.
- There exist more elaborate (and non parametric) estimators of Σ which can be used to calculate (heteroskedasticity) *robust standard errors*.

Inference: Sampling Distributions

- Suppose that y given X (equivalently ε given X) is *normally distributed*.
- The OLS estimator is a *linear* function of y and is therefore, conditional on X, normally distributed. (The same argument applies to the GLS estimator employing Σ). For the OLS estimator with $Var[\varepsilon|X] = \sigma^2 I$:

  $$\hat{\beta}|X \sim N_k[\beta, \sigma^2(X'X)^{-1}]$$

  and when $Var[\varepsilon|X] = \Sigma$, for the GLS estimator,

  $$\breve{\beta}|X \sim N_k[\beta, (X'\Sigma^{-1}X)^{-1}].$$

- Sticking with the homoskedastic case, consider a linear combination of β, $c'\beta$:

  $$c'\hat{\beta}|X \sim N[c'\beta, \sigma^2 c'(X'X)^{-1}c].$$

Inference: Confidence Intervals

- Let $Z \sim N[0,1]$ and let $z_L(\alpha)$ and $z_U(\alpha)$ be the closest pair of values such that $P[z_L(\alpha) \leq Z \leq z_U(\alpha)] = \alpha$. $z_L(\alpha)$ is the $(1-\alpha)/2$ quantile of the standard normal distribution. Choosing $\alpha = 0.95$ gives $z_U(\alpha) = 1.96$, $z_L(\alpha) = -1.96$.

- The result above concerning the distribution of $c'\hat{\beta}$ implies that

  $$P[z_L(\alpha) \leq \frac{c'\hat{\beta} - c'\beta}{\sigma\left(c'(X'X)^{-1}c\right)^{1/2}} \leq z_U(\alpha)] = \alpha$$

  which in turn implies that

  $$P[c'\hat{\beta} - z_U(\alpha)\sigma\left(c'(X'X)^{-1}c\right)^{1/2} \leq c'\beta \leq c'\hat{\beta} - z_L(\alpha)\sigma\left(c'(X'X)^{-1}c\right)^{1/2}] = \alpha.$$

  Consider the interval

  $$[c'\hat{\beta} - z_U(\alpha)\sigma\left(c'(X'X)^{-1}c\right)^{1/2}, c'\hat{\beta} - z_L(\alpha)\sigma\left(c'(X'X)^{-1}c\right)^{1/2}].$$

  This random interval covers the value $c'\beta$ with probability α. This is known as a $100\alpha\%$ confidence interval for $c'\beta$.

- Note that this interval cannot be calculated without knowledge of σ. In practice here and in the tests and interval estimators that follow one will use an estimator of σ^2.

Estimation of σ^2

- Note that

$$\sigma^2 = n^{-1} E[(y - X\beta)' (y - X\beta) | X]$$

which suggests the analogue estimator

$$\begin{aligned} \hat{\sigma}^2 &= n^{-1} \left(y - X\hat{\beta}\right)' \left(y - X\hat{\beta}\right) \\ &= n^{-1} \hat{\varepsilon}' \hat{\varepsilon} \\ &= n^{-1} y' M y \end{aligned}$$

where $\hat{\varepsilon} = y - X\hat{\beta} = My$ and $M = I - X(X'X)^{-1}X'$.

- $\hat{\sigma}^2$ is a biased estimator and the bias is in the *downward* direction:

$$E[\hat{\sigma}^2] = \frac{n-k}{n} \sigma^2 < \sigma^2$$

but note that the bias is negligible unless k the number of covariates is large relative to n the sample size.

- Intuitively, the bias arises from the fact that the OLS estimator minimises the sum of squared residuals.

- It is possible to correct the bias using the estimator $(n-k)^{-1} \hat{\varepsilon}' \hat{\varepsilon}$ but the effect is small in most economic data sets.

- Under certain conditions to be discussed shortly the estimator $\hat{\sigma}^2$ is *consistent*. This means that in large samples the inaccuracy of the estimator is small and that if in the tests described below the unknown σ^2 is replaced by $\hat{\sigma}^2$ the tests are still approximately correct.

Estimation of σ^2

- Proof of $E[\hat{\sigma}^2] = \frac{n-k}{n}\sigma^2 < \sigma^2$:
- First note that $My = M\varepsilon$ because $MX = 0$. So

$$\hat{\sigma}^2 = n^{-1}y'My = n^{-1}\varepsilon'M\varepsilon$$

.

$$\begin{aligned} E[\hat{\sigma}^2|X] &= n^{-1}E[\varepsilon'M\varepsilon|X] \\ &= n^{-1}E[\text{trace}(\varepsilon'M\varepsilon)|X] \\ &= n^{-1}E[\text{trace}(M\varepsilon\varepsilon')|X] \\ &= n^{-1}\,\text{trace}(ME[\varepsilon\varepsilon'|X]) \\ &= n^{-1}\,\text{trace}(M\Sigma) \end{aligned}$$

and when $\Sigma = \sigma^2 I_n$,

$$\begin{aligned} n^{-1}\,\text{trace}(M\Sigma) &= n^{-1}\sigma^2\,\text{trace}(M) \\ &= n^{-1}\sigma^2\,\text{trace}(I_n - X(X'X)^{-1}X') \\ &= \sigma^2\frac{n-k}{n}. \end{aligned}$$

Confidence regions

- Sometimes we need to make probability statements about the values of more than one linear combination of β. We can do this by developing *confidence regions*.

- For j linear combinations, a $100\alpha\%$ confidence region is a subset of $I\!R^j$ which covers the unknown (vector) value of the j linear combinations with probability α.

- We continue to work under the assumption that y given X (equivalently ε given X) is *normally distributed.*

- Let the j linear combinations of interest be $R\beta = r$, say, where R is $j \times k$ with rank j. The OLS estimator of r is $R\hat{\beta}$ and

  $$R\hat{\beta} \sim N[r, \sigma^2 R(X'X)^{-1}R']$$

  which implies that

  $$\left(R\hat{\beta} - r\right)' \left(R(X'X)^{-1}R'\right)^{-1} \left(R\hat{\beta} - r\right) / \sigma^2 \sim \chi^2_{(j)} \qquad (1)$$

  where $\chi^2_{(j)}$ denotes a *Chi-square* random variable with parameter (degrees of freedom) j.

Chi Square Distribution

- Let the ν x 1 element vector $Z \sim N(0, I_\nu)$.
- Then $\xi = Z'Z = \sum_{i=1}^{\nu} Z_i^2$ (positive valued) has a distribution known as a Chi-square distribution, written $Z \sim \chi^2_{(\nu)}$. The probability density function associated with the $\chi^2_{(\nu)}$ distribution is positively skewed. For small ν its mode is at zero.
- The expected value and variance of a Chi-square random variable are:

$$\begin{aligned} E[\chi^2_{(\nu)}] &= \nu \\ Var[\chi^2_{(\nu)}] &= 2\nu. \end{aligned}$$

  For large ν, the distribution is approximately normal.

- Partial proof: if $Z_i \sim N(0,1)$ then $V[Z_i] = E[Z_i^2] = 1$. Therefore $E[\Sigma_{i=1}^{v} Z_i^2] = v$.
- Generalisation: Let $A \sim N_\nu[\mu, \Sigma]$ and let P be such that $P\Sigma P' = I$, which implies that $P'P = \Sigma^{-1}$. Then $Z = P(A - \mu) \sim N_\nu[0, I]$ so that

$$\xi = Z'Z = (A-\mu)'\Sigma^{-1}(A-\mu) \sim \chi^2_{(\nu)}.$$

Confidence regions continued

- Let $q_{\chi^2(j)}(\alpha)$ denote the $\alpha-$quantile of the $\chi^2_{(j)}$ distribution. Then
$$P[\chi^2_{(j)} \leq q_{\chi^2(j)}(\alpha)] = \alpha$$
implies that
$$P[\left(R\hat{\beta} - r\right)' \left(R(X'X)^{-1}R'\right)^{-1} \left(R\hat{\beta} - r\right)/\sigma^2 \leq q_{\chi^2(j)}(\alpha)] = \alpha.$$
The region in $I\!R^j$ defined by
$$\{r : \left(R\hat{\beta} - r\right)' \left(R(X'X)^{-1}R'\right)^{-1} \left(R\hat{\beta} - r\right)/\sigma^2 \leq q_{\chi^2(j)}(\alpha)\}$$
is a $100\alpha\%$ confidence region for r, covering r with probability α. The boundary of the region is an ellipsoid centred on the point $R\hat{\beta}$.

- Setting R equal to a vector c' (note then $j = 1$) and letting $c^* = c'\beta$, produces
$$\begin{aligned}
\alpha &= P[\left(c'\hat{\beta} - c^*\right)' \left(c'(X'X)^{-1}c\right)^{-1} \left(c'\hat{\beta} - c^*\right)/\sigma^2 \leq q_{\chi^2(1)}(\alpha)] \\
&= P[\frac{\left(c'\hat{\beta} - c^*\right)^2}{\sigma^2 c'(X'X)^{-1}c} \leq q_{\chi^2(1)}(\alpha)] \\
&= P[-\left(q_{\chi^2(1)}(\alpha)\right)^{1/2} \leq \frac{\left(c'\hat{\beta} - c^*\right)}{\sigma\left(c'(X'X)^{-1}c\right)^{1/2}} \leq \left(q_{\chi^2(1)}(\alpha)\right)^{1/2}] \\
&= P[z_L(\alpha) \leq \frac{\left(c'\hat{\beta} - c^*\right)}{\sigma\left(c'(X'X)^{-1}c\right)^{1/2}} \leq z_U(\alpha)]
\end{aligned}$$
where we have used the relationship $\chi^2(1) = N(0,1)^2$.

Tests of hypotheses

- The statistics developed to construct confidence intervals can also be used to conduct tests of hypotheses.

- For example, suppose we wish to conduct a test of the null hypothesis $H_0 : R\beta - r = 0$ against the alternative $H_1 : R\beta - r \neq 0$. The statistic

  $$S = \left(R\hat{\beta} - r\right)' \left(R(X'X)^{-1}R'\right)^{-1} \left(R\hat{\beta} - r\right) / \sigma^2. \qquad (2)$$

  has a $\chi^2(j)$ distribution under the null hypothesis. Under the alternative, let

  $$R\beta - r = \delta \neq 0.$$

  Then

  $$R\hat{\beta} - r \sim N[\delta, \sigma^2 R(X'X)^{-1}R']$$

  and the statistic S will tend to be larger than we would expect to obtain from a $\chi^2(j)$ distribution. So we reject the null hypothesis for large values of S.

Tests of Hypotheses

- The *size* of a test of H_0 is the probability of rejecting H_0 when H_0 is true.
- The *power* of a test against a specific alternative H_1 is the probability of rejecting H_0 when H_1 is true.
- The following test procedure has size λ.

  > *Decision rule*: Reject H_0 if $S > q_{\chi^2(j)}(1-\lambda)$, otherwise do not reject H_0.

  Here $q_{\chi^2(j)}(1-\lambda)$ is the $(1-\lambda)$ quantile of the $\chi^2(j)$ distribution.
- Note that we do *not* talk in terms of accepting H_0 as an alternative to rejection. The reason is that a value of S that does not fall in the rejection region of the test is consonant with many values of $R\beta - r$ that are close to but not equal to 0.

Tests of Hypotheses

- To obtain a test concerning a *single* linear combination of β, $H_0 : c'\beta = c^*$, we can use the procedure above with $j = 1$, giving

  $$S = \frac{\left(c'\hat{\beta} - c^*\right)^2}{\sigma^2 c'(X'X)^{-1}c}$$

  and the following size λ test procedure.

  > *Decision rule*: Reject H_0 if $S > q_{\chi^2(1)}(1-\lambda)$, otherwise do not reject H_0.

- Alternatively we can proceed directly from the sampling distribution of $c'\hat{\beta}$. Since, when H_0 is true,

  $$\frac{\left(c'\hat{\beta} - c^*\right)}{\sigma\left(c'(X'X)^{-1}c\right)^{1/2}} \sim N(0, 1),$$

  we can obtain $z_L(\alpha)$, $z_U(\alpha)$, such that

  $$P[z_L(\alpha) < N(0, 1) < z_U(\alpha)] = \alpha = 1 - \lambda.$$

  The following test procedure has size (probability of rejecting a true null hypothesis) equal to λ.

  Decision rule: Reject H_0 if $S > z_U(\alpha)$ or $S < z_L(\alpha)$, otherwise do not reject H_0.

- Because of the relationship between the standard normal $N(0, 1)$ distribution and the $\chi^2_{(1)}$ distribution the tests are identical.

Confidence Interval: Example

- We regress log of income on age, sex and education dummies in a sample of 39000 Swedish individuals.

	Coef.	Std. Err
Age	0.1145845	0.0010962
Age square	-0.0010657	0.0000109
Male	0.060531	0.0078549
High school degree	0.5937122	0.0093677
College degree	0.7485223	0.0115236
Constant	3.563253	0.0249524
R square:	0.3435	

- 95% confidence interval for College Education:

$$[0.748 - 1.96 * 0.0115, 0.748 + 1.96 * 0.0115] = [0.726, 0.771]$$

- Test of H_0 no gender differences in income:

$$0.06/0.00785 = 7.71$$

Reject H_0.

- Test of H_0 Effect of College degree equal to High School degree:

$$(0.748 - 0.593)/0.0115 = 13.39$$

Reject H_0.

Detecting structural change

- A common application of this testing procedure in econometrics arises when attempting to detect "structural change".

- In a time series application one might imagine that up to some time T_s the vector $\beta = \beta_b$ and after T_s, $\beta = \beta_a$, that is that there are two regimes with switching occurring at time T_s. This situation can be captured by specifying the model

$$y = \begin{bmatrix} y_b \\ y_a \end{bmatrix} = \begin{bmatrix} X_b & 0 \\ 0 & X_a \end{bmatrix} \begin{bmatrix} \beta_b \\ \beta_a \end{bmatrix} + \begin{bmatrix} \varepsilon_b \\ \varepsilon_a \end{bmatrix} = X\beta + \varepsilon$$

where X_b contains data for the period before T_s and X_a contains data for the period after T_s. The null hypothesis of no structural change is expressed by $H_0 : \beta_b = \beta_a$. If all the coefficients are allowed to alter across the structural break then

$$\hat{\varepsilon}_U'\hat{\varepsilon}_U = \hat{\varepsilon}_b'\hat{\varepsilon}_b + \hat{\varepsilon}_a'\hat{\varepsilon}_a$$

where, e.g., $\hat{\varepsilon}_b'\hat{\varepsilon}_b$ is the sum of squared residuals from estimating

$$y_b = X_b\beta_b + \varepsilon_b.$$

The test statistic developed above, specialised to this problem can then be written

$$S = \frac{(\hat{\varepsilon}'\hat{\varepsilon} - (\hat{\varepsilon}_b'\hat{\varepsilon}_b + \hat{\varepsilon}_a'\hat{\varepsilon}_a))}{\sigma^2}$$

where $\hat{\varepsilon}'\hat{\varepsilon}$ is the sum of squared residuals from estimating with the constraint $\hat{\beta}_a = \hat{\beta}_b$ imposed and σ^2 is the common variance of the errors.

- When the errors are identically and independently normally distributed S has a $\chi^2_{(k)}$ distribution under H_0.

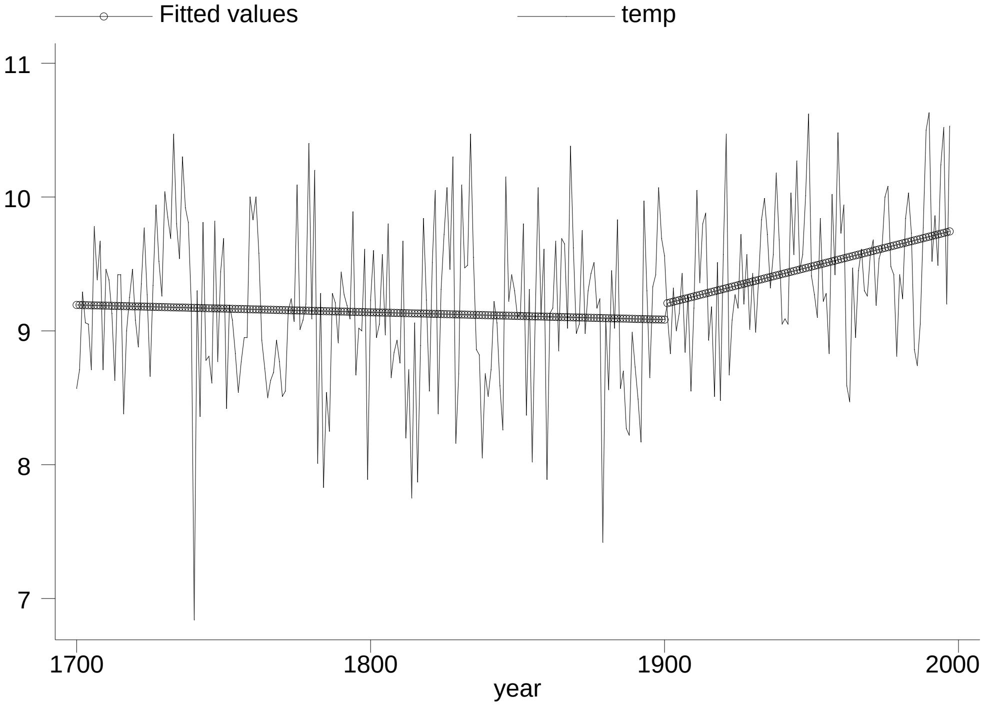
Detecting Structural Change

- In practice an estimate of σ^2 is used - for example there is the statistic
$$S^* = \frac{\left(\hat{\varepsilon}'\hat{\varepsilon} - \left(\hat{\varepsilon}_b'\hat{\varepsilon}_b + \hat{\varepsilon}_a'\hat{\varepsilon}_a\right)\right)}{\left(\hat{\varepsilon}_b'\hat{\varepsilon}_b + \hat{\varepsilon}_a'\hat{\varepsilon}_a\right)/n}$$
where n is the total number of observations in the two periods combined. S has approximately a $\chi^2_{(k)}$ distribution under H_0.

- This application of the theory of tests of linear hypotheses is given the name, "Chow test", after Gregory Chow who popularised the procedure some 30 years ago.

- The test can be modified in various ways.
  - We might wish to keep some of the elements of β constant across regimes.
  - In microeconometrics the same procedure can be employed to test for differences across groups of households, firms etc.

Example: Structural Break in Temperature



- Temperature as a function of Time:
- We test for a break in 1900:

	Coef.	Std Err	Coef.	Std Err
Time (Years)	0.0015	0.00039	-0.00054	0.00069
Time after 1900 (Years)	-	-	0.0061	0.0022

- We can test whether the slope after 1900 is different from the general one:

$$(0.0061 + 0.0054)/0.0022 = 3.03 \qquad Prob = 0.0077$$

- Or conduct a Chow test: $S = 14.62$. We come to the same conclusion. There is a break in the trend.

Estimation in non-linear regression models

- An obvious extension to the linear regression model studied so far is the non-linear regression model:

$$E[Y|X = x] = g(x, \theta)$$

equivalently, in regression function plus error form:

$$\begin{aligned} Y &= g(x, \theta) + \varepsilon \\ E[\varepsilon|X = x] &= 0. \end{aligned}$$

Consider M-estimation and in particular the non-linear least squares estimator obtained as follows.

$$\hat{\theta} = \arg\min_{\theta^*} n^{-1} \sum_{i=1}^{n} (Y_i - g(x_i; \theta^*))^2$$

- For now we just consider how a minimising value $\hat{\theta}$ can be found. Many of the statistical software packages have a routine to conduct non-linear optimisation and some have a non-linear least squares routine. Many of these routines employ a variant of Newton's method.

Numerical optimisation: Newton's method and variants

- Write the minimisation problem as:

  $$\hat{\theta} = \arg\min_{\theta^*} Q(\theta^*).$$

  Newton's method involves taking a sequence of steps, $\theta_0, \theta_1, \ldots,$ $\theta_m, \ldots \theta_M$ from a starting value, θ_0 to an approximate minimising value θ_M which we will use as our estimator $\hat{\theta}$.

- The starting value is provided by the user. One of the tricks is to use a good starting value near to the final solution. This sometimes requires some thought.

- Suppose we are at θ_m. Newton's method considers a quadratic approximation to $Q(\theta)$ which is constructed to be an accurate approximation in a neighbourhood of θ_m, and moves to the value θ_{m+1} which minimises this quadratic approximation.

- At θ_{m+1} a new quadratic approximation, accurate in a neighbourhood of θ_{m+1} is constructed and the next value in the sequence, θ_{m+2}, is chosen as the value of θ minimising this new approximation.

- Steps are taken until a convergence criterion is satisfied. Usually this involves a number of elements. For example one might continue until the following conditions is satisfied:

  $$Q_\theta(\theta_m)' Q_\theta(\theta_m) \leq \delta_1, \qquad |Q(\theta_m) - Q(\theta_{m-1})| < \delta_2.$$

  Convergence criteria vary form package to package. Some care is required in choosing these criteria. Clearly δ_1 and δ_2 above should be chosen bearing in mind the orders of magnitude of the objective function and its derivative.

Numerical optimisation: Newton's method and variants

- The quadratic approximation used at each stage is a quadratic Taylor series approximation. At $\theta = \theta_m$,

  $$Q(\theta) \simeq Q(\theta_m) + (\theta - \theta_m)' Q_\theta(\theta_m) + \frac{1}{2}(\theta - \theta_m)' Q_{\theta\theta'}(\theta_m)(\theta - \theta_m) = Q^a(\theta, \theta_m).$$

  The derivative of $Q^a(\theta, \theta_m)$ with respect to θ is

  $$Q^a_\theta(\theta, \theta_m) = Q_\theta(\theta_m) + Q_{\theta\theta'}(\theta_m)(\theta - \theta_m)$$

  and θ_{m+1} is chosen as the value of θ that solves $Q^a_\theta(\theta, \theta_m) = 0$, namely

  $$\theta_{m+1} = \theta_m - Q_{\theta\theta'}(\theta_m)^{-1} Q_\theta(\theta_m).$$

  There are a number of points to consider here.

  1. Obviously the procedure can only work when the objective function is twice differentiable with respect to θ.
  2. The procedure will stop whenever $Q_\theta(\theta_m) = 0$, which can occur at a maximum and saddlepoint as well as at a minimum. The Hessian, $Q_{\theta\theta'}(\theta_m)$, should be positive definite at a minimum of the function.
  3. When a minimum is found there is no guarantee that it is a global minimum. In problems where this possibility arises it is normal to run the optimisation from a variety of start points to guard against using an estimator that corresponds to a local minimum.
  4. If, at a point in the sequence, $Q_{\theta\theta'}(\theta_m)$ is not positive definite then the algorithm may move away from the minimum and there may be no convergence. Many minimisation (maximisation) problems we deal with involve globally convex (concave) objective functions and for these there is no problem. For other cases, Newton's method is usually modified, e.g. by taking steps

     $$\theta_{m+1} = \theta_m - A(\theta_m)^{-1} Q_\theta(\theta_m)$$

where $A(\theta_m)$ is constructed to be positive definite and in cases in which $Q_{\theta\theta'}(\theta_m)$ is in fact positive definite, to be a good approximation to $Q_{\theta\theta'}(\theta_m)$.

5. The algorithm may "overstep" the minimum to the extent that it takes an "uphill" step, i.e. so that $Q(\theta_{m+1}) > Q(\theta_m)$. This is guarded against in many implementations of Newton's method by taking steps

$$\theta_{m+1} = \theta_m - \alpha(\theta_m)A(\theta_m)^{-1}Q_\theta(\theta_m)$$

where $\alpha(\theta_m)$ is a scalar step scaling factor, chosen to ensure that $Q(\theta_{m+1}) < Q(\theta_m)$.

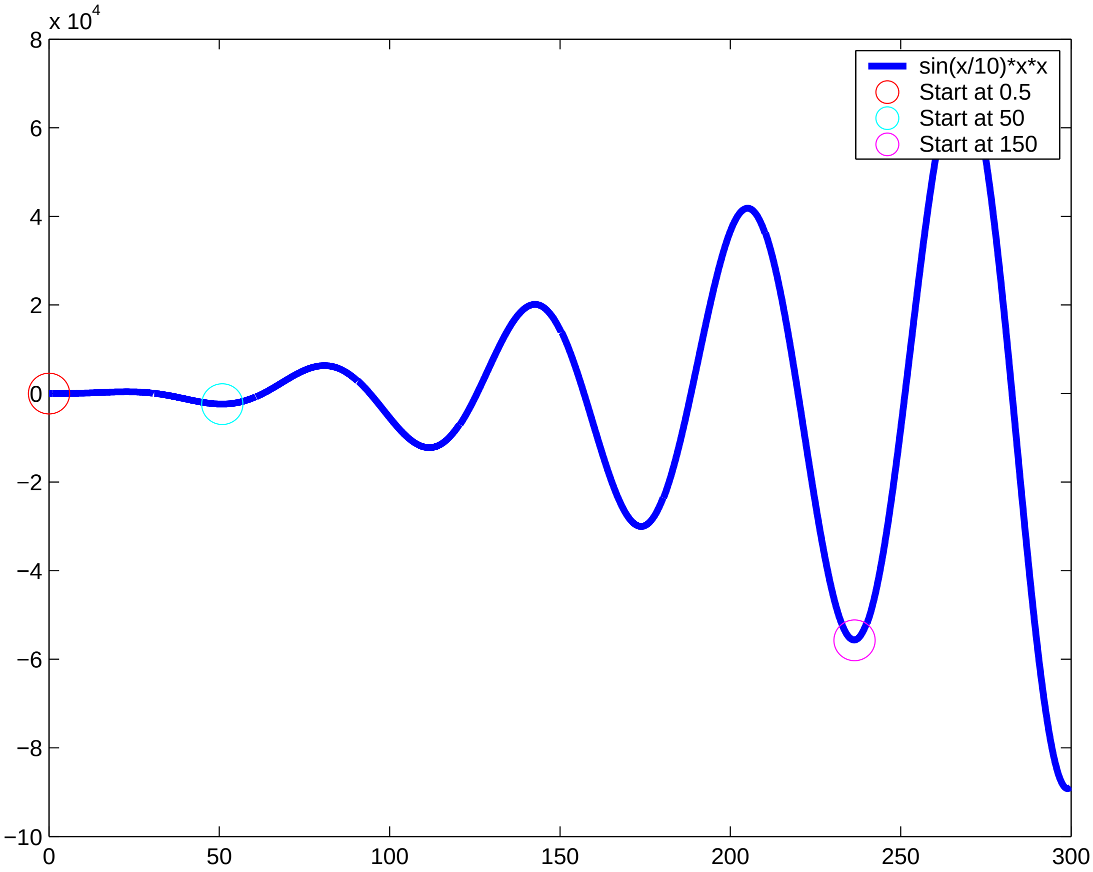
6. In practice it may be difficult to calculate exact expressions for the derivatives that appear in Newton's method. In some cases symbolic computational methods can help. In others we can use a numerical approximation, e.g.

$$Q_{\theta_i}(\theta_m) \simeq \frac{Q_\theta(\theta_m + \delta_i e_i) - Q_\theta(\theta_m)}{\delta_i}$$

where e_i is a vector with a one in position i and zeros elsewhere, and δ_i is a small perturbing value, possibly varying across the elements of θ.

Numerical Optimisation: Example

- Function $y = sin(x/10) * x^2$.
- This function has many (an infinite) number of local minimas.
- Start off the nonlinear optimisation at various points.



Approximate Inference

Approximate Inference

- The results set out in the previous notes let us make inferences about coefficients of regression functions, β, when the distribution of y given X is Gaussian (normal) and the variance of the unobservable disturbances is known.

- In practice the normal distribution at best holds approximately and we never know the value of the nuisance parameter σ. So how can we proceed?

- The most common approach and the one outlined here involves employing *approximations* to exact distributions.

- They have the disadvantage that they can be inaccurate and the magnitude of the inaccuracy can vary substantially from case to case. They have the following advantages:

  1. They are usually very much easier to derive than exact distributions,
  2. They are often valid for a wide family of distributions for y while exact distributional results are only valid for a specific distribution of y, and we rarely know which distribution to use to produce an exact distributional result.

- The most common sort of approximation employed in econometrics is what is known as a *large sample approximation*.

- The main, and increasingly popular, alternative to the use of approximations is to use the bootstrap. This is based on intensive computer replications.

Approximate Inference

- Suppose we have a statistic, S_n, computed using n realisations, for example the OLS estimator, $\hat{\beta}$, or the variance estimator $\hat{\sigma}^2$, or one of the test statistics developed earlier.

- To produce a large sample approximation to the distribution of the statistic, S_n, we regard this statistic as a member of a sequence of statistics, $S_1, \dots, S_n, \dots$, indexed by n, the number of realisations. We write this sequence as $\{S_n\}_{n=1}^{\infty}$.

- Denote the distribution function of S_n by $P[S_n \leq s] = F_{S_n}(s)$. We then consider how the distribution function $F_{S_n}(s)$ behaves as we pass through the sequence, that is as n takes larger and larger values.

- In particular we ask what properties the distribution function has as n tends to infinity. The distribution associated with the limit of the sequence of statistics is sometimes referred to as a *limiting distribution*.

- Sometimes this distribution can be used to produce an approximation to $F_{S_n}(s)$ which can be used to conduct approximate inference using S_n.

Convergence in probability

- In many cases of interest the distributions of a sequence of statistics becomes *concentrated on a single point*, say c, as we pass through the sequence, increasing n. That is, $F_{S_n}(s)$ becomes closer and closer to a step function as n is increased, a step function which is zero up to c, and at c, jumps to 1. In this case we say that S_n *converges in probability* to the constant c.

- A sequence of (possibly vector valued) statistics converges in probability to a constant (possibly vector), c, if, for all $\varepsilon > 0$,

$$\lim_{n\to\infty} P[\|S_n - c\| > \varepsilon] = 0,$$

that is, if for every $\varepsilon, \delta > 0$, there exists N (which typically depends upon ε and δ), such that for all $n > N$

$$P[\|S_n - c\| > \varepsilon] < \delta.$$

- Here the notation $|| \cdot ||$ is used to denote the Euclidean length of a vector, that is: $\|z\| = (z'z)^{1/2}$. This is the absolute value of z when z is a scalar.

- We then write $\text{plim}_{n\to\infty} S_n = c$, or, $S_n \xrightarrow{p} c$, and c is referred to as the *probability limit* of S_n.

Convergence in Probability

- When $S_n = \hat{\theta}_n$ is an *estimator* of a parameter, θ, which takes the value θ_0 and $\hat{\theta}_n \xrightarrow{p} \theta_0$, we say that $\hat{\theta}_n$ is a *consistent estimator*.
- If every member of the sequence $\{E\left[\hat{\theta}_n\right]\}_{i=1}^{\infty}$ and $\{Var\left[\hat{\theta}_n\right]\}_{i=1}^{\infty}$ exists, and

$$\begin{aligned} \lim_{n\to\infty} E\left[\hat{\theta}_n\right] &= \theta \\ \lim_{n\to\infty} Var\left[\hat{\theta}_n\right] &= 0 \end{aligned}$$

then we say that $\hat{\theta}_n$ *converges in mean square* to θ. It is quite easily shown that convergence in mean square implies convergence in probability. It is often easy to derive expected values and variances of statistics. So a quick route to proving consistency is to prove convergence in mean square.

Convergence in Probability

- Note, though, that an estimator can be consistent but *not* converge in mean square. There are commonly occurring cases in econometrics where estimators are consistent but the sequences of moments required for consideration of convergence in mean square do not exist. (For example, the two stage least squares estimator in just identified linear models, i.e. the indirect least squares estimator).

- Consistency is generally regarded as a desirable property for an estimator to possess.

- Note though that in all practical applications of econometric methods we have a finite sized sample at our disposal. The consistency property on its own does not tell us about the quality of the estimate that we calculate using such a sample. It might be better sometimes to use an inconsistent estimator that generally takes values close to the unknown θ than a consistent estimator that is very inaccurate except at a much larger sample size than we have available.

- The consistency property does tell us that with a large enough sample our estimate would likely be close to the unknown truth, but not how close, nor even how large a sample is required to get an estimate close to the unknown truth.

Convergence in distribution

- A sequence of statistics $\{S_n\}_{n=1}^{\infty}$ that converges in probability to a constant has a variance (if one exists) which becomes small as we pass to larger values of n.

- If we multiply S_n by a function of n, chosen so that the variance of the transformed statistic remains approximately constant as we pass to larger values of n, then we may obtain a sequence of statistics which converge not to a constant but to a *random variable*.

- If we can work out what the distribution of this random variable is, then we can use this distribution to approximate the distributions of the transformed statistics in the sequence.

- Consider a sequence of random variables $\{T_n\}_{n=1}^{\infty}$. Denote the distribution function of T_n by

  $$P[T_n \leq t] = F_{T_n}(t).$$

  Let T be a random variable with distribution function

  $$P[T \leq t] = F_T(t).$$

  We say that $\{T_n\}_{n=1}^{\infty}$ *converges in distribution* to T if for all $\varepsilon > 0$ there exists N (which will generally depend upon ε) such that for all $n > N$,

  $$|F_{T_n}(t) - F_T(t)| < \varepsilon$$

  at all points t at which $F_T(t)$ is continuous. Then we write

  $$T_n \xrightarrow{d} T$$

  .

- The definition applies for vector and scalar random variables. In this situation we will also talk in terms of T_n *converging in probability* to (the random variable) T.

Convergence in Distribution

- Now return to the sequence $\{S_n\}_{n=1}^{\infty}$ that converges in probability to a constant.
- Let $T_n = h(n)(S_n)$ with $h(\cdot) > 0$ chosen so that $\{T_n\}_{n=1}^{\infty}$ *converges in distribution* to a random variable T that has a non-degenerate distribution.
- A common case that will arise is that in which $h(n) = n^{\alpha}$. In this course we will only encounter the special case in which $\alpha = 1/2$, that is $h(n) = n^{1/2}$.
- We can use the limiting random variable T to make approximate probability statements as follows. Since $S_n = T_n/h(n)$,

$$\begin{aligned} P[S_n \leq s] &= P[T_n/h(n) < s] \\ &= P[T_n < s \times h(n)] \\ &\simeq P[T < s \times h(n)] \\ &= F_T(s \times h(n)) \end{aligned}$$

  which allows approximate probability statements concerning the random variable S_n.

Convergence in Distribution: Example

- Consider the mean, $\bar{X}_n$ of n independently and identically distributed random variables with common mean and variance respectively μ and σ^2.
- One of the simplest Central Limit Theorems (see below) says that, if $T_n = n^{1/2}(\bar{X}_n - \mu)/\sigma$ then $T_n \xrightarrow{d} T \sim N(0,1)$.
- We can use this result to say that $T_n \simeq N(0,1)$ where "$\simeq$" here means "is approximately distributed as". This sort of result can be used to make approximate probability statements. Since T has a standard normal distribution

$$P[-1.96 \leq T \leq 1.96] = 0.95$$

and so, approximately,

$$P[-1.96 \leq \frac{n^{1/2}(\bar{X}_n - \mu)}{\sigma} \leq 1.96] \simeq 0.95$$

leading, if σ^2 were known, to the approximate 95% *confidence interval* for μ,

$$\{\bar{X}_n - 1.96\sigma/n^{1/2}, \bar{X}_n + 1.96\sigma/n^{1/2}\},$$

approximate in the sense that

$$P[\bar{X}_n - 1.96\sigma/n^{1/2} \leq \mu \leq \bar{X}_n + 1.96\sigma/n^{1/2}] \simeq 0.95$$

Approximate Inference: Some Thoughts

- It is *very important* to realise that in making this approximation there is *no* sense in which we ever think of the sample size actually becoming large.

- The sequence $\{S_n\}_{n=1}^{\infty}$ indexed by the sample size is just a hypothetical construct in the context of which we can develop an approximation to the distribution of a statistic.

- For example we know that when y given X is normally distributed the OLS estimator is exactly normally distributed conditional on X. For non-normal y, under some conditions, as we will see, the limiting distribution of an appropriately scaled OLS estimator is normal. The quality of that normal approximation depends upon the sample size, but also upon the extent of the departure of the distribution of y given X from normality and upon the disposition of the values of the covariates. For y close to normality the normal approximation to the distribution of the OLS estimator is good even at very small sample sizes.

- The extent to which, at the value of n that we have, the deviations $|F_{T_n}(t) - F_T(t)|$ are large or small can be studied by Monte Carlo simulation or by considering higher order approximations.

Functions of statistics - Slutsky's Theorem

- Slutsky's Theorem states that if T_n is a sequence of random variables that converges in probability to a constant c, and $g(\cdot)$ is a continuous function, then $g(T_n)$ converges in probability to $g(c)$.

- T_n can be a vector or matrix of random variables in which case c is a vector or matrix of constants. Sometimes c is called the *probability limit* of T_n.

- A similar result holds for convergence to a random variable, namely that if T_n is a sequence of random variables that converges in probability to a random variable T, and $g(\cdot)$ is a continuous function, then $g(T_n)$ converges in probability to $g(T)$.

- For example, if
$$T_n' = \left[\, T_n^{1\prime} \;\vdots\; T_n^{2\prime} \, \right]$$
and
$$T_n \xrightarrow{d} T = \left[\, T^{1\prime} \;\vdots\; T^{2\prime} \, \right]'$$
then
$$T_n^1 + T_n^2 \xrightarrow{d} T^1 + T^2$$

Limit theorems

- The *Lindberg-Levy Central Limit Theorem* gives the limiting distribution of a mean of identically distributed random variables. The Theorem states that if $\{Y_i\}_{i=1}^{\infty}$ are mutually independent random (vector) variables each with expected value μ and positive definite covariance matrix Ω then if $\bar{Y}_n = n^{-1}\sum_{i=1}^{n} Y_i$,

$$n^{1/2}(\bar{Y}_n - \mu) \xrightarrow{d} Z, \qquad Z \sim N(0, \Omega).$$

- Many of the statistics we encounter in econometrics can be expressed as means of *non-identically distributed* random vectors, whose limiting distribution is the subject of the *Lindberg-Feller Central Limit Theorem*. The Theorem states that if $\{Y_i\}_{i=1}^{\infty}$ are independently distributed random variables with $E[Y_i] = \mu_i$, $Var[Y_i] = \Omega_i$ with finite third moments and

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \frac{1}{n}\sum_{i=1}^{n}\mu_i = \bar{\mu}_n$$

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mu_i = \mu \qquad \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\Omega_i = \Omega,$$

where Ω is finite and positive definite, and for each j

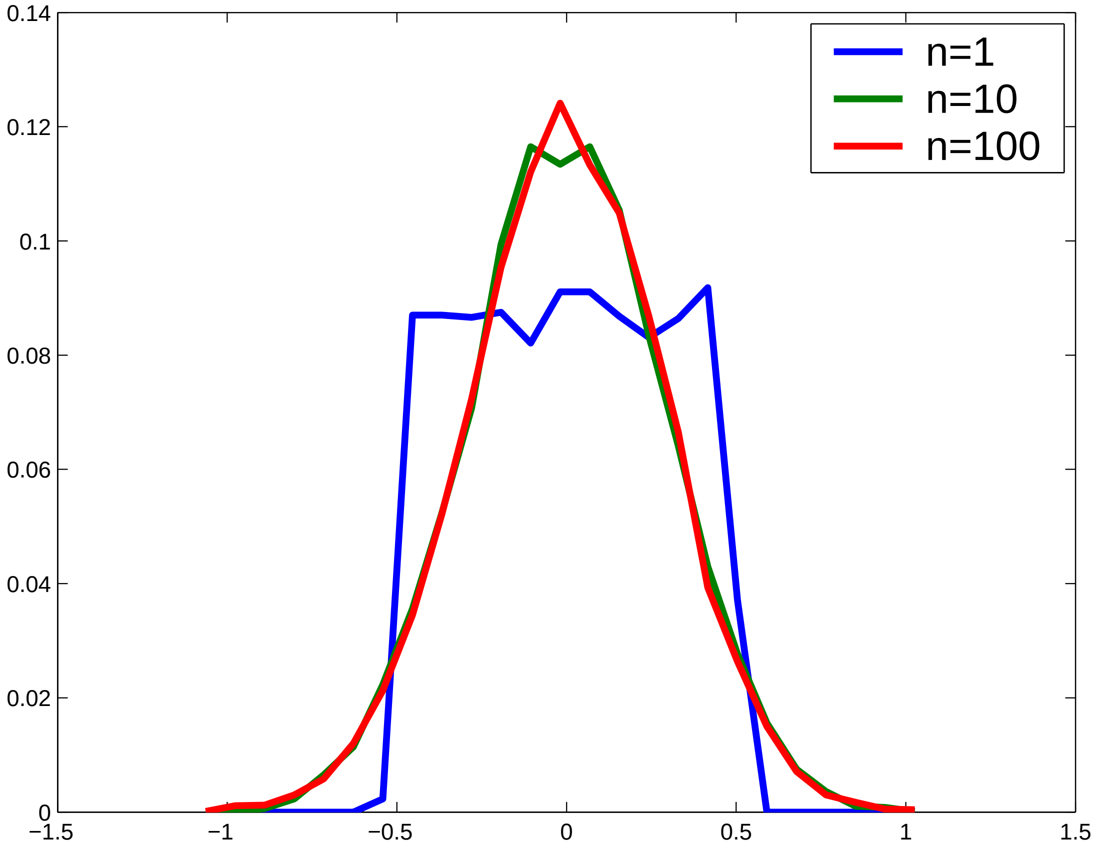
$$\lim_{n\to\infty}\left(\sum_{i=1}^{n}\Omega_i\right)^{-1}\Omega_j = 0, \tag{3}$$

then

$$n^{1/2}(\bar{Y}_n - \bar{\mu}_n) \xrightarrow{d} Z, \qquad Z \sim N(0, \Omega).$$

Limit Theorem: Example

- Start with n uniform random variables $\{Y_i\}_{i=1}^n$ over $[0, 1]$.
- Denote by $\bar{Y}_n$ the mean of Y_i based on a sample of size n.
- The graph plots the distribution of $n^{1/2}(\bar{Y}_n - 0.5)$:



Approximate Distribution Of The Ols Estimator

- Consider the OLS estimator $S_n = \hat{\beta}_n = (X_n'X_n)^{-1}X_n'y_n$ where we index by n to indicate that a sample of size n is involved. We know that when

  $$y_n = X_n\beta + \varepsilon \qquad E[\varepsilon_n|X_n] = 0 \qquad Var[\varepsilon_n|X_n] = \sigma^2 I_n$$

  then

  $$E[\hat{\beta}_n|X_n] = \beta$$

  and

  $$Var[\hat{\beta}_n|X_n] = \sigma^2(X_n'X_n)^{-1} = n^{-1}\sigma^2(n^{-1}X_n'X_n)^{-1} = n^{-1}\sigma^2(n^{-1}\sum_{i=1}^{n} x_i x_i')^{-1}.$$

- **Consistency**: If the x_i's were independently sampled from some distribution such that

  $$n^{-1}\sum_{i=1}^{n} x_i x_i' = n^{-1}X'X \xrightarrow{p} \Sigma_{xx}$$

  and if this matrix of expected squares and cross-products is non-singular then

  $$\lim_{n\to\infty} Var[\hat{\beta}_n|X_n] = 0.$$

  In this case $\hat{\beta}_n$ converges in mean square to β (recall that $E[\hat{\beta}|X] = \beta$), so $\hat{\beta}_n \xrightarrow{p} \beta$ and the OLS estimator is consistent.

OLS Estimator: Limiting distribution

- To make large sample approximate inference using the OLS estimator, consider the centred statistics

$$S_n = \hat{\beta}_n - \beta$$

and the associated scaled statistics

$$\begin{aligned} T_n &= n^{1/2} S_n \\ &= n^{1/2}(\hat{\beta}_n - \beta) \\ &= (n^{-1} X_n' X_n)^{-1} n^{-1/2} X_n' \varepsilon_n. \end{aligned}$$

Assuming $(n^{-1}X_n'X_n)^{-1} \xrightarrow{p} \Sigma_{xx}^{-1}.$, consider the term

$$n^{-1/2} X_n' \varepsilon_n = n^{-1/2} \sum_{i=1}^{n} x_i \varepsilon_i.$$

Let $R_i = x_i \varepsilon_i$ and note that

$$E[R_i] = 0, \qquad Var[R_i] = \sigma^2 x_i x_i'.$$

Under suitable conditions on the vectors x_i, the R_i's satisfy the conditions of the Lindberg-Feller Central Limit Theorem and we have

$$n^{-1/2} \sum_{i=1}^{n} R_i = n^{-1/2} X_n' \varepsilon_n \xrightarrow{d} N(0, \sigma^2 \Sigma_{xx}).$$

Finally, by Slutsky's Theorem

$$T_n = n^{1/2}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \sigma^2 \Sigma_{xx}^{-1}).$$

- We use this approximation to say that

$$n^{1/2}(\hat{\beta}_n - \beta) \simeq N(0, \sigma^2 \Sigma_{xx}^{-1}).$$

OLS Estimator: Limiting distribution

- In practice σ^2 and Σ_{xx} are unknown and we replace them by estimates, e.g. $\hat{\sigma}_n^2$ and $n^{-1}X_n'X_n$.
- If these are consistent estimates then we can use Slutsky's Theorem to obtain the limiting distributions of the resulting statistics.
- Example: testing the hypothesis $H_0 : R\beta = r$, we have already considered the statistic

$$S_n = (R\hat{\beta}_n - r)' \left(R\left(X_n'X_n\right)^{-1} R'\right)^{-1} (R\hat{\beta}_n - r)/\sigma^2$$

  where the subscript "n" is now appended to indicate the sample size under consideration. In the normal linear model, $S_n \sim \chi^2_{(j)}$.
- When y given X is non-normally distributed the limiting distribution result given above can be used, as follows. Rewrite S_n as

$$S_n = \left(n^{1/2}(R\hat{\beta}_n - r)\right)' \left(R\left(n^{-1}X_n'X_n\right)^{-1} R'\right)^{-1} \left(n^{1/2}(R\hat{\beta}_n - r)\right)/\sigma^2.$$

  Let P_n be such that

$$P_n \left(R\left(n^{-1}X_n'X_n\right)^{-1} R'\right) P_n' = I_j$$

  and consider the sequence of random variables

$$T_n = \frac{n^{1/2}}{\sigma} P_n (R\hat{\beta}_n - r).$$

  $T_n \xrightarrow{d} N(0, I_j)$ as long as $P_n \xrightarrow{p} P$ where $P'(R\Sigma_{xx}^{-1}R')P = I_j$. Application of the results on limiting distributions of functions of random variables gives

$$T_n'T_n \xrightarrow{d} \chi^2_{(j)}.$$

OLS Estimator: Limiting Distribution

- Now

  $$T_n'T_n = \frac{n}{\sigma^2}(R\hat{\beta}_n - r)' \left(R\left(n^{-1}X_n'X_n\right)^{-1}R'\right)^{-1}(R\hat{\beta}_n - r)$$

  where we have used

  $$P_n'P_n = \left(R\left(n^{-1}X_n'X_n\right)^{-1}R'\right)^{-1}.$$

  Cancelling the terms involving n:

  $$T_n'T_n = S_n \xrightarrow{d} \chi^2_{(j)}.$$

- Finally, if $\hat{\sigma}_n^2$ is a consistent estimator of σ^2 then it can replace σ^2 in the formula for S_n and the approximate $\chi^2_{(j)}$ still applies, that is:

  $$\left(n^{1/2}(R\hat{\beta}_n - r)\right)' \left(R\left(n^{-1}X_n'X_n\right)^{-1}R'\right)^{-1}\left(n^{1/2}(R\hat{\beta}_n - r)\right)/\hat{\sigma}_n^2 \xrightarrow{d} \chi^2_{(j)}.$$

- The other results we developed earlier for the normal linear model with "known" σ^2 also works as approximations when a normality restrictions does not hold and when σ^2 is replaced by a consistent estimator.

Approximate Distribution of the GLS Estimator

- Consider the following linear model:

$$\begin{aligned} y &= X\beta + \varepsilon \\ E[\varepsilon|X] &= 0 \\ Var[\varepsilon|X] &= \Omega \end{aligned}$$

- The GLS estimator $\tilde{\beta} = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}y$ is BLU, and when y given X is normally distributed:

$$\tilde{\beta} \sim N(\beta, \left(X'\Omega^{-1}X\right)^{-1}).$$

- When y given X is non-normally distributed we can proceed as above, working in the context of a transformed model in which transformed y given X has an identity covariance matrix giving, under suitable conditions $\tilde{\beta} \xrightarrow{p} \beta$ and the limiting distribution:

$$n^{1/2}(\tilde{\beta} - \beta) \xrightarrow{d} N(0, \left(n^{-1}X'\Omega^{-1}X\right)^{-1}).$$

- We noted that in practice Ω is unknown and suggested using a feasible GLS estimator, $\tilde{\beta} = \left(X'\hat{\Omega}^{-1}X\right)^{-1} X'\hat{\Omega}^{-1}y$ in which $\hat{\Omega}$ was some estimate of the conditional variance of y given X.

- Suppose $\hat{\Omega}$ is a *consistent* estimator of Ω. Then it can be shown that $\tilde{\beta}$ is a consistent estimator of β and under suitable conditions

$$n^{1/2}(\tilde{\beta} - \beta) \xrightarrow{d} N(0, \left(n^{-1}X'\Omega^{-1}X\right)^{-1}).$$

Approximate Distribution of the GLS Estimator

- When $\hat{\Omega}$ is a consistent estimator the limiting distribution of the feasible GLS estimator is the same as the limiting distribution of the estimator that employs Ω.
- The exact distributions differ in a finite sized sample to an extent that depends upon the accuracy of the estimator of Ω in that finite sized sample.
- When the elements of Ω are functions of a finite number of parameters it may be possible to produce a consistent estimator, $\hat{\Omega}$.
- Example: consider a heteroscedastic model in which Ω is diagonal with diagonal elements

  $$\omega_{ii} = f(x_i, \gamma).$$

  A first step OLS estimation produces residuals, $\hat{\varepsilon}_i$ and

  $$E[\hat{\varepsilon}_i^2 | X] = (M\Omega M)_{ii} = M_i'\Omega M_i = \omega_{ii} M_{ii}$$

  where M_i' is the ith row of M and M_{ii} is the (i,i) element of M. This simplification follows from the diagonality of Ω and the idempotency of M. We can therefore write

  $$\frac{\hat{\varepsilon}_i^2}{M_{ii}} = f(x_i, \gamma) + u_i$$

  where $E[u_i|X] = 0$, and under suitable conditions a nonlinear least squares estimation will produce a consistent estimator of γ, leading to a consistent estimator of Ω.

Approximate Distribution of M-Estimators

- It is difficult to develop exact distributions for these estimators, except under very special circumstances (e.g. for the OLS estimator with normally distributed y given X)

- Consider an M-estimator defined as

$$\hat{\theta}_n = \arg\max_{\theta} \quad U(Z_n, \theta)$$

  where θ is a vector of parameters and Z_n is a vector random variable. In the applications we will consider Z_n contains n random variables representing outcomes observed in a sample of size n. We wish to obtain the limiting distribution of $\hat{\theta}_n$.

- The first step is to show that $\hat{\theta}_n \xrightarrow{p} \theta_0$, the true value of θ. This is done by placing conditions on U and on the distribution of Z_n which ensure that:

  1. for θ in a neighbourhood of θ_0, $U(Z_n, \theta) \xrightarrow{p} U^*(\theta)$.
  2. the sequence of values (indexed by n) of θ that maximise $U(Z_n, \theta)$ converges in probability to the value of θ that maximises $U^*(\theta)$
  3. the value of θ that uniquely maximises $U^*(\theta)$ is θ_0, the unknown parameter value. (identification)

Approximate Distribution of M-Estimators

- To obtain the limiting distribution of $n^{1/2}\left(\hat{\theta}_n - \theta_0\right)$, consider situations in which the M-estimator can be defined as the unique solution to first order conditions

$$U_\theta(Z_n, \hat{\theta}_n) = 0 \qquad \text{where} \quad U_\theta(Z_n, \hat{\theta}_n) = \frac{\partial}{\partial \theta} U(Z_n, \theta)|_{\theta = \hat{\theta}_n}$$

This is certainly the case when $U(Z_n, \theta)$ is concave.

- We first consider a Taylor series expansion of $U(Z_n, \theta)$ regarded as a function of θ around $\theta = \theta_0$, as follows:

$$U_\theta(Z_n, \theta) = U_\theta(Z_n, \theta_0) + U_{\theta\theta}(Z_n, \theta_0)\,(\theta - \theta_0) + R(\theta, \theta_0, Z_n)$$

Evaluating this at $\theta = \hat{\theta}_n$ gives:

$$0 = U_\theta(Z_n, \hat{\theta}_n) = U_\theta(Z_n, \theta_0) + U_{\theta\theta}(Z_n, \theta_0)\left(\hat{\theta}_n - \theta_0\right) + R(\hat{\theta}_n, \theta_0, Z_n)$$

where

$$U_{\theta\theta}(Z_n, \theta) = \frac{\partial^2}{\partial\theta\partial\theta'} U(Z_n, \theta).$$

The remainder term, $R(\hat{\theta}_n, \theta_0, Z_n)$, involves the third derivatives of $U(Z_n, \theta)$ and in many situations converges in probability to zero as n becomes large. This allows us to write:

$$U_\theta(Z_n, \theta_0) + U_{\theta\theta}(Z_n, \theta_0)\left(\hat{\theta}_n - \theta_0\right) \simeq 0$$

and then

$$\left(\hat{\theta}_n - \theta_0\right) \simeq -U_{\theta\theta}(Z_n, \theta_0)^{-1} U_\theta(Z_n, \theta_0).$$

Equivalently:

$$n^{1/2}\left(\hat{\theta}_n - \theta_0\right) \simeq -\left(n^{-1} U_{\theta\theta}(Z_n, \theta_0)\right)^{-1} n^{-1/2} U_\theta(Z_n, \theta_0).$$

Approximate Distribution of M-Estimator

- In the situations we will encounter it is possible to find conditions under which

$$n^{-1}U_{\theta\theta}(Z_n,\theta_0) \xrightarrow{p} A(\theta_0) \qquad n^{-1/2}U_\theta(Z_n,\theta_0) \xrightarrow{d} N(0,B(\theta_0)),$$

for some matrices $A(\theta_0)$ and $B(\theta_0)$, concluding that

$$n^{1/2}\left(\hat{\theta}_n-\theta_0\right) \xrightarrow{d} N(0,A(\theta_0)^{-1}B(\theta_0)A(\theta_0)^{-1\prime}).$$

- Example: OLS estimator

$$\hat{\theta}_n = \arg\max_\theta \left\{-\sum_{i=1}^n (Y_i - x_i'\theta)^2\right\}$$

when $Y_i = x_i'\theta_0 + \varepsilon_i$ and the ε_i's are independently distributed with expected value zero and common variance σ_0^2.

$$\begin{aligned} U(Z_n,\theta) &= -\sum_{i=1}^n (Y_i - x_i'\theta)^2 \\ n^{-1/2}U_\theta(Z_n,\theta) &= 2n^{-1/2}\sum_{i=1}^n (Y_i - x_i'\theta)x_i \\ n^{-1}U_{\theta\theta}(Z_n,\theta) &= -2n^{-1}\sum_{i=1}^n x_i x_i' \end{aligned}$$

and, defining $\Sigma_{XX} \equiv \operatorname{plim}_{n\to\infty} n^{-1}\sum_{i=1}^n x_i x_i'$:

$$A(\theta_0) = -2\Sigma_{XX}$$

which does not depend upon θ_0 in this special case,

$$B(\theta_0) = 4\sigma_0^2\Sigma_{XX}$$

$$A(\theta_0)^{-1}B(\theta_0)A(\theta_0)^{-1\prime} = \sigma_0^2\Sigma_{XX}^{-1}$$

and finally the OLS estimator has the following limiting normal distribution.

$$n^{1/2}\left(\hat{\theta}_n-\theta_0\right) \xrightarrow{d} N(0,\sigma_0^2\Sigma_{XX}^{-1}).$$

Approximate distributions of functions of estimators the "delta method"

- We proceed in a more general context in which we are interested in a scalar function of a vector of parameters, $h(\theta)$, and suppose that we have a consistent estimator $\hat{\theta}$ of θ whose approximate distribution is given by

$$n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Omega)$$

where θ_0 is the data generating value of θ.

- What is the approximate distribution of $h(\hat{\theta})$?

Consider a Taylor series expansion of $h(\theta)$ around $\theta = \theta_0$ as follows

$$h(\theta) = h(\theta_0) + (\theta - \theta_0)' h_\theta(\theta_0) + \frac{1}{2}(\theta - \theta_0)' h_{\theta\theta}(\theta^*)(\theta - \theta_0)$$

where $h_\theta(\theta_0)$ is the vector of derivatives of $h(\theta)$ evaluated at $\theta = \theta_0$, $h_{\theta\theta}(\theta^*)$ is the matrix of second derivatives of $h(\theta)$ evaluated at $\theta = \theta^*$, a value between θ and θ_0. Evaluate this at $\theta = \hat{\theta}$ and rearrange to give

$$n^{1/2}\left(h(\hat{\theta}) - h(\theta_0)\right) = n^{1/2}(\hat{\theta} - \theta_0)' h_\theta(\theta_0) + \frac{1}{2} n^{1/2}(\hat{\theta} - \theta_0)' h_{\theta\theta}(\hat{\theta}^*)(\hat{\theta} - \theta_0)$$

where $\hat{\theta}^*$ lies between $\hat{\theta}$ and θ_0. Since $\hat{\theta}$ is consistent, $\hat{\theta}^*$ must converge to θ_0 and if $h_{\theta\theta}(\theta_0)$ is bounded then the second term above disappears[1] as $n \to \infty$. So, we have

$$n^{1/2}\left(h(\hat{\theta}) - h(\theta_0)\right) \xrightarrow{d} h_\theta(\theta_0)' Z$$

where

$$n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} Z \sim N(0, \Omega).$$

Using our result on linear functions of normal random variables

$$n^{1/2}\left(h(\hat{\theta}) - h(\theta_0)\right) \xrightarrow{d} N(0, h_\theta(\theta_0)' \Omega h_\theta(\theta_0)).$$

[1] $n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Omega)$ and $(\hat{\theta} - \theta_0) \xrightarrow{p} 0$.

Delta Method: Example

- Suppose we have $\theta = [\theta_1, \theta_2]'$ and that we are interested in $h(\theta) = \theta_2/\theta_1$ leading to

$$h_\theta(\theta) = \begin{bmatrix} -\theta_2/\theta_1^2 \\ 1/\theta_1 \end{bmatrix}.$$

Write the approximate variance of $n^{1/2}(\hat{\theta} - \theta_0)$ as

$$\Omega = \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{12} & \omega_{22} \end{bmatrix}.$$

Then the approximate variance of $n^{1/2}\left(h(\hat{\theta}) - h(\theta_0)\right)$ is

$$\begin{aligned} &= \begin{bmatrix} -\theta_2/\theta_1^2 \\ 1/\theta_1 \end{bmatrix}' \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{12} & \omega_{22} \end{bmatrix} \begin{bmatrix} -\theta_2/\theta_1^2 \\ 1/\theta_1 \end{bmatrix} \\ &= \left(\theta_2^2/\theta_1^4\right)\omega_{11} - 2\left(\theta_2/\theta_1^3\right)\omega_{12} + \left(1/\theta_1^2\right)\omega_{22} \\ &= \left(1/\theta_1^2\right)\left(\left(\theta_2^2/\theta_1^2\right)\omega_{11} - 2\left(\theta_2/\theta_1\right)\omega_{12} + \omega_{22}\right) \end{aligned}$$

in which θ_1 and θ_2 are here taken to indicate the data generating values. Clearly if θ_1 is very close to zero then this will be large. Note that if θ_1 were actually zero then the development above would not go through because the condition on $h_{\theta\theta}(\theta_0)$ being bounded would be violated.

The method we have used here is sometimes called the “delta method”.

Maximum Likelihood Methods

Maximum Likelihood Methods

- Some of the models used in econometrics specify the complete probability distribution of the outcomes of interest rather than just a regression function.

- Sometimes this is because of special features of the outcomes under study - for example because they are discrete or censored, or because there is serial dependence of a complex form.

- When the complete probability distribution of outcomes given covariates is specified we can develop an expression for the probability of observation of the responses we see as a function of the unknown parameters embedded in the specification.

- We can then ask what values of these parameters maximise this probability for the data we have. The resulting statistics, functions of the observed data, are called *maximum likelihood estimators*. They possess important optimality properties and have the advantage that they can be produced in a rule directed fashion.

Estimating a Probability

- Suppose $Y_1, \ldots Y_n$ are binary independently and identically distributed random variables with $P[Y_i = 1] = p$, $P[Y_i = 0] = 1-p$ for all i.

- We might use such a model for data recording the occurrence or otherwise of an event for n individuals, for example being in work or not, buying a good or service or not, etc.

- Let $y_1, \ldots, y_n$ indicate the data values obtained and note that in this model

$$\begin{aligned} P[Y_1 = y_1 \cap \cdots \cap Y_n = y_n, p] &= \prod_{i=1}^{n} p^{y_i}(1-p)^{(1-y_i)} \\ &= p^{\sum_{i=1}^n y_i}(1-p)^{\sum_{i=1}^n (1-y_i)} \\ &= L(p; y). \end{aligned}$$

  With any set of data $L(p; y)$ can be calculated for any value of p between 0 and 1. The result is the probability of observing the data to hand for each chosen value of p.

- One strategy for estimating p is to use that value that maximises this probability. The resulting estimator is called the *maximum likelihood estimator* (MLE) and the maximand, $L(p; y)$, is called the *likelihood function*.

Log Likelihood Function

- The maximum of the *log likelihood function*, $l(p;y) = \log L(p,y)$, is at the same value of p as is the maximum of the likelihood function (because the log function is monotonic).
- It is often easier to maximise the log likelihood function (LLF). For the problem considered here the LLF is

  $$l(p;y) = \left(\sum_{i=1}^{n} y_i\right) \log p + \sum_{i=1}^{n} (1-y_i) \log(1-p).$$

  Let

  $$\hat{p} = \arg\max_p L(p;y) = \arg\max_p l(p;y).$$

  On differentiating we have the following.

  $$\begin{aligned} l_p(p;y) &= \frac{1}{p}\sum_{i=1}^{n} y_i - \frac{1}{1-p}\sum_{i=1}^{n}(1-y_i) \\ l_{pp}(p;y) &= -\frac{1}{p^2}\sum_{i=1}^{n} y_i - \frac{1}{(1-p)^2}\sum_{i=1}^{n}(1-y_i). \end{aligned}$$

  Note that $l_{pp}(p;y)$ is always negative for admissable p so the optimisation problem has a unique solution corresponding to a maximum. The solution to $l_p(\hat{p};y) = 0$ is

  $$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

  just the mean of the observed values of the binary indicators, equivalently the proportion of 1's observed in the data.

Likelihood Functions and Estimation in General

- Let Y_i, $i = 1, \ldots, n$ be continuously distributed random variables with joint probability density function $f(y_1, \ldots, y_n, \theta)$.
- The probability that Y falls in infinitesimal intervals of width $dy_1, \ldots dy_n$ centred on values $y_1, \ldots, y_n$ is

  $$A = f(y_1, \ldots, y_n, \theta) dy_1 dy_2 \ldots dy_n$$

  Here only the joint density function depends upon θ and the value of θ that maximises $f(y_1, \ldots, y_n, \theta)$ also maximises A.
- In this case the likelihood function is defined to be the joint *density* function of the Y_i's.
- When the Y_i's are discrete random variables the likelihood function is the joint probability mass function of the Y_i's, and in cases in which there are discrete and continuous elements the likelihood function is a combination of probability density elements and probability mass elements.
- In all cases the likelihood function is a function of the observed data values that is equal to, or proportional to, the probability of observing these particular values, where the constant of proportionality does not depend upon the parameters which are to be estimated.

Likelihood Functions and Estimation in General

- When Y_i, $i = 1, \ldots, n$ are *independently* distributed the joint density (mass) function is the *product* of the marginal density (mass) functions of each Y_i, the likelihood function is

  $$L(y;\theta) = \prod_{i=1}^{n} f_i(y_i;\theta),$$

  and the log likelihood function is the *sum*:

  $$l(y;\theta) = \sum_{i=1}^{n} \log f_i(y_i;\theta).$$

  There is a subscript i on f to allow for the possibility that each Y_i has a distinct probability distribution.

- This situation arises when modelling conditional distributions of Y given some covariates x. In particular, $f_i(y_i;\theta) = f_i(y_i|x_i;\theta)$, and often $f_i(y_i|x_i;\theta) = f(y_i|x_i;\theta)$.

- In time series and panel data problems there is often dependence among the Y_i's. For any list of random variables $Y = \{Y_1, \ldots, Y_n\}$ define the $i-1$ element list $Y_{i-} = \{Y_1, \ldots, Y_{i-1}\}$. The joint density (mass) function of Y can be written as

  $$f(y) = \prod_{i=2}^{n} f_{y_i|y_{i-}}(y_i|y_{i-}) f_{y_1}(y_1),$$

Invariance

- Note that (parameter free) monotonic transformations of the Y_i's (for example, a change of units of measurement, or use of logs rather than the original y data) usually leads to a change in the value of the maximised likelihood function when we work with continuous distributions.

- If we transform from y to z where $y = h(z)$ and the joint density function of y is $f_y(y;\theta)$ then the joint density function of z is
$$f_z(z;\theta) = \left|\frac{\partial h(z)}{\partial z}\right| f_y(h(z);\theta).$$

- For any given set of values, y^*, the value of θ that maximises the likelihood function $f_y(y^*,\theta)$ also maximises the likelihood function $f_z(z^*;\theta)$ where $y^* = h(z^*)$, so the maximum likelihood estimator is invariant with respect to such changes in the way the data are presented.

- However the maximised likelihood functions will differ by a factor equal to $\left|\frac{\partial h(z)}{\partial z}\right|_{z=z^*}$.

- The reason for this is that we omit the infinitesimals $dy_1, \ldots dy_n$ from the likelihood function for continuous variates and these change when we move from y to z because they are denominated in the units in which y or z are measured.

Maximum Likelihood: Properties

- Maximum likelihood estimators possess another important *invariance property.* Suppose two researchers choose different ways in which to parameterise the same model. One uses θ, and the other uses $\lambda = h(\theta)$ where this function is one-to-one. Then faced with the same data and producing estimators $\hat{\theta}$ and $\hat{\lambda}$, it will always be the case that $\hat{\lambda} = h(\hat{\theta})$.

- There are a number of important consequences of this:

  - For instance, if we are interested in the ratio of two parameters, the MLE of the ratio will be the ratio of the ML estimators.

  - Sometimes a re-parameterisation can improve the numerical properties of the likelihood function. Newton's method and its variants may in practice work better if parameters are rescaled.

Maximum Likelihood: Improving Numerical Properties

- An example of this often arises when, in index models, elements of x involve squares, cubes, etc., of some covariate, say x_1. Then maximisation of the likelihood function may be easier if instead of x_1^2, x_1^3, etc., you use $x_1^2/10$, $x_1^3/100$, etc., with consequent rescaling of the coefficients on these covariates. You can always recover the MLEs you would have obtained without the rescaling by rescaling the estimates.

- There are some cases in which a re-parameterisation can produce a globally concave likelihood function where in the original parameterisation there was not global concavity.

- An example of this arises in the "Tobit" model.
  - This is a model in which each Y_i is $N(x_i'\beta, \sigma^2)$ with negative realisations replaced by zeros. The model is sometimes used to model expenditures and hours worked, which are necessarily non-negative.
  - In this model the likelihood as parameterised here is not globally concave, but re-parameterising to $\lambda = \beta/\sigma$, and $\gamma = 1/\sigma$, produces a globally concave likelihood function.
  - The invariance property tells us that having maximised the "easy" likelihood function and obtained estimates $\hat{\lambda}$ and $\hat{\gamma}$, we can recover the maximum likelihood estimates we might have had difficulty finding in the original parameterisation by calculating $\hat{\beta} = \hat{\lambda}/\hat{\gamma}$ and $\hat{\sigma} = 1/\hat{\gamma}$.

Properties Of Maximum Likelihood Estimators

- First we just sketch the main results:
  - Let $l(\theta; Y)$ be the log likelihood function now regarded as a random variable, a function of a set of (possibly vector) random variables $Y = \{Y_1, \ldots, Y_n\}$.
  - Let $l_\theta(\theta; Y)$ be the gradient of this function, itself a vector of random variables (scalar if θ is scalar) and let $l_{\theta\theta}(\theta; Y)$ be the matrix of second derivatives of this function (also a scalar if θ is a scalar).
  - Let
    $$\hat{\theta} = \arg\max_{\theta} \quad l(\theta; Y).$$
    In order to make inferences about θ using $\hat{\theta}$ we need to determine the distribution of $\hat{\theta}$. We consider developing a large sample approximation. The limiting distribution for a quite wide class of maximum likelihood problems is as follows:
    $$n^{1/2}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V_0)$$
    where
    $$V_0 = -\operatorname*{plim}_{n\to\infty}(n^{-1} l_{\theta\theta}(\theta_0; Y))^{-1}$$
    and θ_0 is the unknown parameter value. To get an approximate distribution that can be used in practice we use $(n^{-1} l_{\theta\theta}(\hat{\theta}; Y))^{-1}$ or some other consistent estimator of V_0 in place of V_0.

Properties Of Maximum Likelihood Estimators

- We apply the method for dealing with M-estimators.
- Suppose $\hat{\theta}$ is uniquely determined as the solution to the first order condition

$$l_\theta(\hat{\theta}; Y) = 0$$

  and that $\hat{\theta}$ is a consistent estimator of the unknown value of the parameter, θ_0. Weak conditions required for consistency are quite complicated and will not be given here.

- Taking a Taylor series expansion around $\theta = \theta_0$ and then evaluating this at $\theta = \hat{\theta}$ gives

$$0 \simeq l_\theta(\theta_0; Y) + l_{\theta\theta'}(\theta_0; Y)(\hat{\theta} - \theta_0)$$

  and rearranging and scaling by powers of the sample size n

$$n^{1/2}(\hat{\theta} - \theta_0) \simeq -\left(n^{-1} l_{\theta\theta'}(\theta; Y)\right)^{-1} n^{-1/2} l_\theta(\theta; Y).$$

  As in our general treatment of M-estimators if we can show that

$$n^{-1} l_{\theta\theta'}(\theta_0; Y) \xrightarrow{p} A(\theta_0)$$

  and

$$n^{-1/2} l_\theta(\theta_0; Y) \xrightarrow{d} N(0, B(\theta_0))$$

  then

$$n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, A(\theta_0)^{-1} B(\theta_0) A(\theta_0)^{-1\prime}).$$

Maximum Likelihood: Limiting Distribution

- What is the limiting distribution of $n^{-1/2}l_\theta(\theta_0; Y)$?
- First note that in problems for which the Y_i's are independently distributed, $n^{-1/2}l_\theta(\theta_0; Y)$ is a scaled mean of random variables and we may be able to find conditions under which a central limit theorem applies, indicating a limiting *normal* distribution.
- We must now find the mean and variance of this distribution.

  Since $L(\theta; Y)$ is a joint probability density function (we just consider the continuous distribution case here),

  $$\int L(\theta; y)dy = 1$$

  where multiple integration is over the support of Y. If this support *does not depend upon* θ, then

  $$\frac{\partial}{\partial\theta}\int L(\theta; y)dy = \int L_\theta(\theta; y)dy = 0.$$

  But, because $l(\theta; y) = \log L(\theta; y)$, and $l_\theta(\theta; y) = L_\theta(\theta; y)/L(\theta; y)$, we have

  $$\int L_\theta(\theta; y)dy = \int l_\theta(\theta; y)L(\theta; y)dy = E\left[l_\theta(\theta; Y)\right]$$

  and so $E\left[l_\theta(\theta; Y)\right] = 0$.
- This holds for any value of θ, in particular for θ_0 above. If the variance of $l_\theta(\theta_0; Y)$ converges to zero as n becomes large then $l_\theta(\theta_0; Y)$ will converge in probability to zero and the mean of the limiting distribution of $n^{-1/2}l_\theta(\theta_0; Y)$ will be zero.

Maximum Likelihood: Limiting Distribution

- We turn now to the variance of the limiting distribution. We have just shown that

$$\int l_\theta(\theta;y)L(\theta;y)dy = 0.$$

Differentiating again

$$\begin{aligned}
\frac{\partial}{\partial\theta'}\int l_\theta(\theta;y)L(\theta;y)dy &= \int\left(l_{\theta\theta'}(\theta;y)L(\theta;y)+l_\theta(\theta;y)L_{\theta'}(\theta;y)\right)dy \\
&= \int\left(l_{\theta\theta'}(\theta;y)+l_\theta(\theta;y)l_\theta(\theta;y)'\right)L(\theta;y)dy \\
&= E\left[l_{\theta\theta'}(\theta;Y)+l_\theta(\theta;Y)l_\theta(\theta;Y)'\right] \\
&= 0.
\end{aligned}$$

Separating the two terms in the penultimate line,

$$E\left[l_\theta(\theta;Y)l_\theta(\theta;Y)'\right] = -E\left[l_{\theta\theta'}(\theta;Y)\right] \qquad (4)$$

and note that, since $E\left[l_\theta(\theta;Y)\right]=0$,

$$Var[l_\theta(\theta;Y)] = E\left[l_\theta(\theta;Y)l_\theta(\theta;Y)'\right]$$

and so

$$\begin{aligned}
Var[l_\theta(\theta;Y)] &= -E\left[l_{\theta\theta'}(\theta;Y)\right] \\
\Rightarrow Var[n^{-1/2}l_\theta(\theta;Y)] &= -E\left[n^{-1}l_{\theta\theta'}(\theta;Y)\right]
\end{aligned}$$

giving

$$B(\theta_0) = -\operatorname*{plim}_{n\to\infty} n^{-1}l_{\theta\theta'}(\theta_0;Y).$$

The matrix

$$I(\theta) = -E\left[l_{\theta\theta}(\theta;Y)\right]$$

plays a central role in likelihood theory - it is called the *Information Matrix*.

Finally, because $B(\theta_0) = -A(\theta_0)$

$$A(\theta)^{-1}B(\theta)A(\theta)^{-1\prime} = -\left(\operatorname*{plim}_{n\to\infty} n^{-1}l_{\theta\theta'}(\theta;Y)\right)^{-1}.$$

- Of course a number of conditions are required to hold for the results above to hold. These include the boundedness of third order derivatives of the log likelihood function, independence or at most weak dependence of the Y_i's, existence of moments of derivatives of the log likelihood, or at least of probability limits of suitably scaled versions of them, and lack of dependence of the support of the Y_i's on θ.

- The result in equation (4) above leads, under suitable conditions concerning convergence, to

  $$\operatorname*{plim}_{n\to\infty}\left(n^{-1}l_\theta(\theta;Y)l_\theta(\theta;Y)'\right) = -\operatorname*{plim}_{n\to\infty}\left(n^{-1}l_{\theta\theta'}(\theta;Y)\right).$$

  This gives an alternative way of "estimating " V_0, namely

  $$\hat{V}_0^o = \left\{n^{-1}l_\theta(\hat{\theta};Y)l_\theta(\hat{\theta};Y)'\right\}^{-1}$$

  which compared with

  $$\tilde{V}_0^o = \left\{-n^{-1}l_{\theta\theta'}(\hat{\theta};Y)\right\}^{-1}$$

  has the advantage that only first derivatives of the log likelihood function need to be calculated. Sometimes $\hat{V}_0^o$ is referred to as the "outer product of gradient" (OPG) estimator. Both these estimators use the "observed" values of functions of derivatives of the LLF and. It may be possible to derive explicit expressions for the expected values of these functions. Then one can estimate V_0 by

  $$\begin{aligned}\hat{V}_0^e &= \left\{E[n^{-1}l_\theta(\theta;Y)l_\theta(\theta;Y)']|_{\theta=\hat{\theta}}\right\}^{-1}\\ &= \left\{-E[n^{-1}l_{\theta\theta'}(\theta;Y)]|_{\theta=\hat{\theta}}\right\}^{-1}.\end{aligned}$$

  These two sorts of estimators are sometimes referred to as "observed information" ($\hat{V}_0^o$, $\tilde{V}_0^o$) and "expected information" ($\hat{V}_0^e$) estimators.

- Maximum likelihood estimators possess optimality property, namely that, among the class of consistent and asymptotically normally distributed estimators, the variance matrix of their limiting distribution is the smallest that can be achieved in the sense that other estimators in the class have limiting distributions with variance matrices exceeding the MLE's by a positive semidefinite matrix.

Estimating a Conditional Probability

- Suppose $Y_1, \dots Y_n$ are binary independently and identically distributed random variables with

  $$\begin{aligned} P[Y_i &= 1|X = x_i] = p(x_i, \theta) \\ P[Y_i &= 0|X = x_i] = 1 - p(x_i, \theta). \end{aligned}$$

  This is an obvious extension of the model in the previous section.

- The likelihood function for this problem is

  $$\begin{aligned} P[Y_1 = y_1 \cap \cdots \cap Y_n = y_n|x] &= \prod_{i=1}^{n} p(x_i, \theta)^{y_i}(1 - p(x_i, \theta))^{(1-y_i)} \\ &= L(\theta; y). \end{aligned}$$

  where y denotes the complete set of values of y_i and dependence on x is suppressed in the notation. The log likelihood function is

  $$l(\theta; y) = \sum_{i=1}^{n} y_i \log p(x_i, \theta) + \sum_{i=1}^{n} (1 - y_i) \log(1 - p(x_i, \theta))$$

  and the maximum likelihood estimator of θ is

  $$\hat{\theta} = \arg\max_{\theta} \quad l(\theta; y).$$

  So far this is an obvious generalisation of the simple problem met in the last section.

Estimating a Conditional Probability

- To implement the model we choose a form for the function $p(x,\theta)$, which must of course lie between zero and one.
  - One common choice is

    $$p(x,\theta) = \frac{\exp(x'\theta)}{1+\exp(x'\theta)}$$

    which produces what is commonly called a *logit model*.
  - Another common choice is

    $$\begin{aligned} p(x,\theta) &= \Phi(x'\theta) = \int_{-\infty}^{x'\theta} \phi(w)dw \\ \phi(w) &= (2\pi)^{-1/2}\exp(-w^2/2) \end{aligned}$$

    in which Φ is the standard normal distribution function. This produces what is known as a *probit model*.
- Both models are widely used. Note that in both cases a single index model is specified, the probability functions are monotonic increasing, probabilities arbitrarily close to zero or one are obtained when $x'\theta$ is sufficiently large or small, and there is a symmetry in both of the models in the sense that $p(-x,\theta) = 1 - p(x,\theta)$.
- Any or all of these properties might be inappropriate in a particular application but there is rarely discussion of this in the applied econometrics literature.

More on Logit and Probit

- Both models can also be written as a linear model involving a latent variable.
- We define a **latent variable** Y_i^*, which is unobserved, but determined by the following model:

  $$Y_i^* = X_i\theta + \varepsilon_i$$

  We observe the variable Y_i which is linked to Y_i^* as:

  $$\begin{cases} Y_i = 0 & \text{if } Y_i^* < 0 \\ Y_i = 1 & \text{if } Y_i^* \geq 0 \end{cases}$$

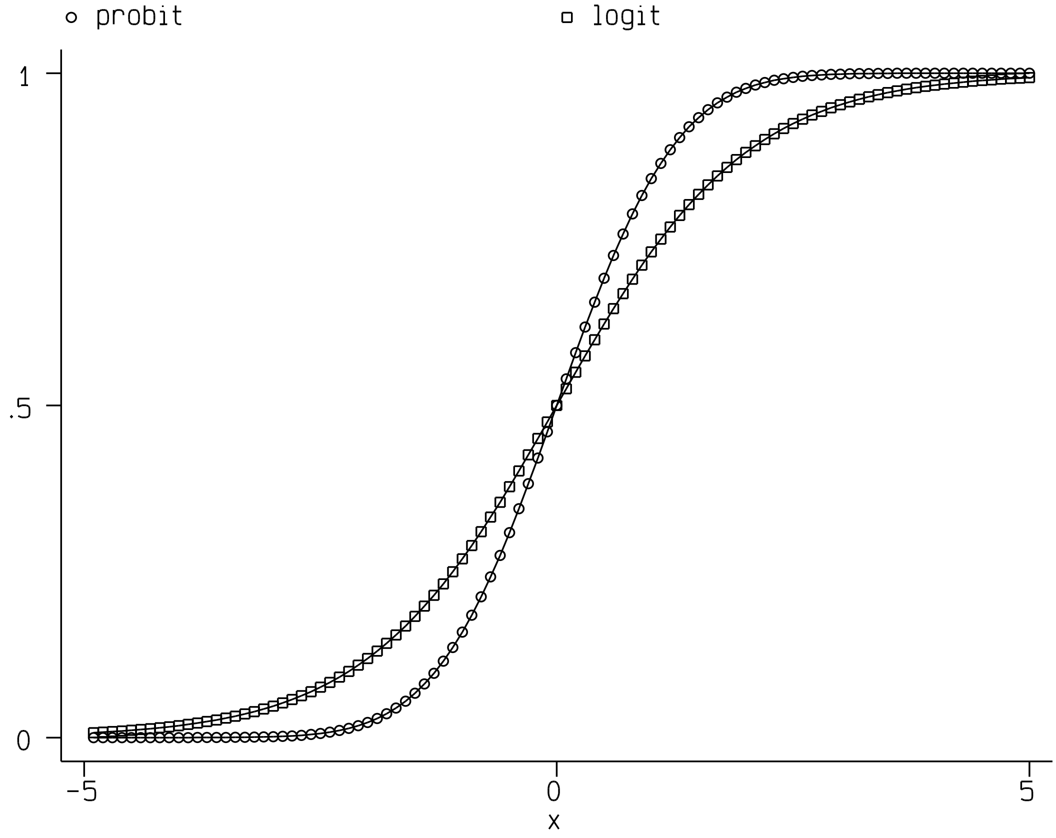
- The probability of observing $Y_i = 1$ is:

  $$\begin{aligned} p_i = P(Y_i = 1) &= P(Y_i^* \geq 0) \\ &= P(X_i\theta + \varepsilon_i \geq 0) \\ &= P(\varepsilon_i \geq -X_i\theta) \\ &= 1 - F_\varepsilon(-X_i\theta) \end{aligned}$$

  where F_ε is the cumulative distribution function of the random variable ε.

- If ε_i is distributed normally, the model is the probit model.
- If ε_i follows an extreme value distribution, the model is the logit model.

Shape of Logit and Probit Models



Odds-Ratio

- Define the ratio $p_i/(1-p_i)$ as the **odds-ratio**. This is the ratio of the probability of outcome 1 over the probability of outcome 0. If this ratio is equal to 1, then both outcomes have equal probability ($p_i = 0.5$). If this ratio is equal to 2, say, then outcome 1 is twice as likely than outcome 0 ($p_i = 2/3$).
- In the logit model, the log odds-ratio is linear in the parameters:
$$\ln \frac{p_i}{1-p_i} = X_i\theta$$
- In the logit model, θ is the marginal effect of X on the log odds-ratio. A unit increase in X leads to an increase of θ % in the odds-ratio.

Marginal Effects

- Logit model:

$$\begin{aligned}\frac{\partial p_i}{\partial X} &= \frac{\theta \exp(X_i\theta)(1+\exp(X_i\theta)) - \theta \exp(X_i\theta)^2}{(1+\exp(X_i\theta))^2} \\ &= \frac{\theta \exp(X_i\theta)}{(1+\exp(X_i\theta))^2} \\ &= \theta p_i(1-p_i)\end{aligned}$$

A one unit increase in X leads to an increase in the probability of choosing option 1 of $\theta p_i(1-p_i)$.

- Probit model:

$$\frac{\partial p_i}{\partial X_i} = \theta\phi(X_i\theta)$$

A one unit increase in X leads to an increase in the probability of choosing option 1 of $\theta\phi(X_i\theta)$.

Maximum Likelihood in Single Index Models

- We can cover both cases by considering general single index models, so for the moment rewrite $p(x, \theta)$ as $g(w)$ where $w = x'\theta$.
- The log-likelihood is then:

$$l(\theta, y) = \sum_{i=1}^{n} y_i \log(g(w_i)) + \sum_{i=1}^{n} (1 - y_i) \log(1 - g(w_i))$$

- The first derivative of the log likelihood function is:

$$\begin{aligned} l_\theta(\theta; y) &= \sum_{i=1}^{n} \frac{g_w(x_i'\theta)x_i}{g(x_i'\theta)} y_i - \frac{g_w(x_i'\theta)x_i}{1 - g(x_i'\theta)} (1 - y_i) \\ &= \sum_{i=1}^{n} (y_i - g(x_i'\theta)) \frac{g_w(x_i'\theta)}{g(x_i'\theta)\,(1 - g(x_i'\theta))} x_i \end{aligned}$$

  Here $g_w(w)$ is the derivative of $g(w)$ with respect to w.

- The expression for the second derivative is rather messy. Here we just note that its expected value given x is quite simple, namely

$$E[l_{\theta\theta}(\theta; y)|x] = -\sum_{i=1}^{n} \frac{g_w(x_i'\theta)^2}{g(x_i'\theta)\,(1 - g(x_i'\theta))} x_i x_i',$$

  the negative of which is the Information Matrix for general single index binary data models.

Asymptotic Properties of the Logit Model

- For the logit model there is major simplification

$$\begin{aligned} g(w) &= \frac{\exp(w)}{1+\exp(w)} \\ g_w(w) &= \frac{\exp(w)}{(1+\exp(w))^2} \\ \Rightarrow \frac{g_w(w)}{g(w)\,(1-g(w))} &= 1. \end{aligned}$$

Therefore in the logit model the MLE satisfies

$$\sum_{i=1}^{n}\left(y_i - \frac{\exp(x_i'\hat{\theta})}{1+\exp(x_i'\hat{\theta})}\right)x_i = 0,$$

the Information Matrix is

$$I(\theta) = \sum_{i=1}^{n}\frac{\exp(x_i'\theta)}{(1+\exp(x_i'\theta))^2}x_i x_i',$$

the MLE has the limiting distribution

$$\begin{aligned} &n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, V_0) \\ V_0 &= \left(\operatorname*{plim}_{n\to\infty} n^{-1}\sum_{i=1}^{n}\frac{\exp(x_i'\theta)}{(1+\exp(x_i'\theta))^2}x_i x_i'\right)^{-1}, \end{aligned}$$

and we can conduct approximate inference using the following approximation

$$n^{1/2}(\hat{\theta}_n - \theta_0) \simeq N(0, V_0)$$

using the estimator

$$\hat{V}_0 = \left(n^{-1}\sum_{i=1}^{n}\frac{\exp(x_i'\hat{\theta})}{\left(1+\exp(x_i'\hat{\theta})\right)^2}x_i x_i'\right)^{-1}$$

when producing approximate hypothesis tests and confidence intervals.

Asymptotic Properties of the Probit Model

- In the probit model

$$\begin{aligned} g(w) &= \Phi(w) \\ g_w(w) &= \phi(w) \\ \Rightarrow \frac{g_w(w)}{g(w)\left(1-g(w)\right)} &= \frac{\phi(w)}{\Phi(w)(1-\Phi(w))}. \end{aligned}$$

Therefore in the probit model the MLE satisfies

$$\sum_{i=1}^{n}\left(y_i-\Phi(x_i'\hat{\theta})\right)\frac{\phi(x_i'\hat{\theta})}{\Phi(x_i'\hat{\theta})(1-\Phi(x_i'\hat{\theta}))}x_i=0,$$

the Information Matrix is

$$I(\theta)=\sum_{i=1}^{n}\frac{\phi(x_i'\theta)^2}{\Phi(x_i'\theta)(1-\Phi(x_i'\theta))}x_ix_i',$$

the MLE has the limiting distribution

$$\begin{aligned} & n^{1/2}(\hat{\theta}_n-\theta_0)\xrightarrow{d}N(0,V_0) \\ V_0 &= \left(\operatorname*{plim}_{n\to\infty} n^{-1}\sum_{i=1}^{n}\frac{\phi(x_i'\theta)^2}{\Phi(x_i'\theta)(1-\Phi(x_i'\theta))}x_ix_i'\right)^{-1}, \end{aligned}$$

and we can conduct approximate inference using the following approximation

$$n^{1/2}(\hat{\theta}_n-\theta_0)\simeq N(0,V_0)$$

using the estimator

$$\hat{V}_0=\left(n^{-1}\sum_{i=1}^{n}\frac{\phi(x_i'\hat{\theta})^2}{\Phi(x_i'\hat{\theta})(1-\Phi(x_i'\hat{\theta}))}x_ix_i'\right)^{-1}$$

when producing approximate tests and confidence intervals.

Example: Logit and Probit

- We have data from households in Kuala Lumpur (Malaysia) describing household characteristics and their concern about the environment. The question is

  ”Are you concerned about the environment? Yes / No”.

  We also observe their age, sex (coded as 1 men, 0 women), income and quality of the neighborhood measured as air quality. The latter is coded with a dummy variable *smell*, equal to 1 if there is a bad smell in the neighborhood. The model is:

  $$Concern_i = \beta_0 + \beta_1 age_i + \beta_2 sex_i + \beta_3 \log income_i + \beta_4 smell_i + u_i$$

- We estimate this model with three specifications, linear probability model (LPM), logit and probit:

Probability of being concerned by Environment

Variable	LPM		Logit		Probit	
	Est.	t-stat	Est.	t-stat	Est.	t-stat
age	.0074536	3.9	.0321385	3.77	.0198273	3.84
sex	.0149649	0.3	.06458	0.31	.0395197	0.31
log income	.1120876	3.7	.480128	3.63	.2994516	3.69
smell	.1302265	2.5	.5564473	2.48	.3492112	2.52
constant	-.683376	-2.6	-5.072543	-4.37	-3.157095	-4.46
			Some Marginal Effects			
Age	.0074536		.0077372		.0082191	
log income	.1120876		.110528		.1185926	
smell	.1302265		.1338664		.1429596	

Multinomial Logit

- The logit model was dealing with two qualitative outcomes. This can be generalized to multiple outcomes:
  - choice of transportation: car, bus, train...
  - choice of dwelling: house, apartment, social housing.
- The multinomial logit: Denote the outcomes as $j = 1, \dots, J$ and p_j the probability of outcome j.

$$p_j = \frac{\exp(X\theta^j)}{\sum_{k=1}^{J} \exp(X\theta^k)}$$

  where θ^j is a vector of parameter associated with outcome j.

Identification

- If we multiply all the coefficients by a factor λ this does not change the probabilities p_j, as the factor cancel out. This means that there is under identification. We have to normalize the coefficients of one outcome, say, J to zero. All the results are interpreted as deviations from the baseline choice.

- We write the probability of choosing outcome $j = 1, \ldots, J-1$ as:

$$p_j = \frac{\exp(X\theta^j)}{1 + \sum_{k=1}^{J-1} \exp(X\theta^k)}$$

- We can express the logs odds-ratio as:

$$\ln \frac{p_j}{p_J} = X\theta^j$$

- The odds-ratio of choice j versus J is only expressed as a function of the parameters of choice j, but not of those other choices: Independence of Irrelevant Alternatives (IIA).

Independence of Irrelevant Alternatives

An anecdote which illustrates a violation of this property has been attributed to Sidney Morgenbesser:

After finishing dinner, Sidney Morgenbesser decides to order dessert. The waitress tells him he has two choices: apple pie and blueberry pie. Sidney orders the apple pie.

After a few minutes the waitress returns and says that they also have cherry pie at which point Morgenbesser says ”In that case I'll have the blueberry pie.”

Independence of Irrelevant Alternatives

- Consider travelling choices, by car or with a red bus. Assume for simplicity that the choice probabilities are equal:

$$P(car) = P(\text{red bus}) = 0.5 \quad \Longrightarrow \quad \frac{P(car)}{P(\text{red bus})} = 1$$

- Suppose we introduce a blue bus, (almost) identical to the red bus. The probability that individuals will choose the blue bus is therefore the same as for the red bus and the odd ratio is:

$$P(\text{blue bus}) = P(\text{red bus}) \quad \Longrightarrow \quad \frac{P(\text{blue bus})}{P(\text{red bus})} = 1$$

- However, the IIA implies that odds ratios are the same whether of not another alternative exists. The only probabilities for which the three odds ratios are equal to one are:

$$P(car) = P(\text{blue bus}) = P(\text{red bus}) = 1/3$$

However, the prediction we ought to obtain is:

$$P(\text{red bus}) = P(\text{blue bus}) = 1/4 \quad P(car) = 0.5$$

Marginal Effects: Multinomial Logit

- θ^j can be interpreted as the marginal effect of X on the log odds-ratio of choice j to the baseline choice.

- The marginal effect of X on the probability of choosing outcome j can be expressed as:

$$\frac{\partial p_j}{\partial X} = p_j[\theta^j - \sum_{k=1}^{J} p_k \theta^k]$$

  Hence, the marginal effect on choice j involves not only the coefficients relative to j but also the coefficients relative to the other choices.

- Note that we can have $\theta^j < 0$ and $\partial p_j / \partial X > 0$ or vice versa.

  Due to the non linearity of the model, the sign of the coefficients does not indicate the direction nor the magnitude of the effect of a variable on the probability of choosing a given outcome. One has to compute the marginal effects.

Example

- We analyze here the choice of dwelling: house, apartment or low cost flat, the latter being the baseline choice. We include as explanatory variables the age, sex and log income of the head of household:

Variable	Estimate	Std. Err.	Marginal Effect
	Choice of House		
age	.0118092	.0103547	-0.002
sex	-.3057774	.2493981	-0.007
log income	1.382504	.1794587	0.18
constant	-10.17516	1.498192	
	Choice of Apartment		
age	.0682479	.0151806	0.005
sex	-.89881	.399947	-0.05
log income	1.618621	.2857743	0.05
constant	-15.90391	2.483205	

Ordered Models

- In the multinomial logit, the choices were not ordered. For instance, we cannot rank cars, busses or train in a meaningful way. In some instances, we have a natural ordering of the outcomes even if we cannot express them as a continuous variable:
  - Yes / Somehow / No.
  - Low / Medium / High
- We can analyze these answers with ordered models.

Ordered Probit

- We code the answers by arbitrary assigning values:

$$Y_i = 0 \text{ if No}, \quad Y_i = 1 \text{ if Somehow}, \quad Y_i = 2 \text{ if Yes}$$

- We define a latent variable Y_i^* which is linked to the explanatory variables:

$$Y_i^* = X_i'\theta + \varepsilon_i$$

$$\begin{array}{ll} Y_i = 0 & \text{if } Y_i^* < 0 \\ Y_i = 1 & \text{if } Y_i^* \in [0, \mu[\\ Y_i = 2 & \text{if } Y_i^* \geq \mu \end{array}$$

  μ is a threshold and an auxiliary parameter which is estimated along with θ.

- We assume that ε_i is distributed normally.

- The probability of each outcome is derived from the normal cdf:

$$\begin{array}{l} P(Y_i = 0) = \Phi(-X_i'\theta) \\ P(Y_i = 1) = \Phi(\mu - X_i'\theta) - \Phi(-X_i'\theta) \\ P(Y_i = 2) = 1 - \Phi(\mu - X_i'\theta) \end{array}$$

Ordered Probit

- Marginal Effects:

$$\begin{aligned}\frac{\partial P(Y_i=0)}{\partial X_i} &= -\theta\phi(-X_i'\theta)\\ \frac{\partial P(Y_i=1)}{\partial X_i} &= \theta\left(\phi(X_i'\theta)-\phi(\mu-X_i'\theta)\right)\\ \frac{\partial P(Y_i=2)}{\partial X_i} &= \theta\phi(\mu-X_i'\theta)\end{aligned}$$

- Note that if $\theta > 0$, $\partial P(Y_i=0)/\partial X_i < 0$ and $\partial P(Y_i=2)/\partial X_i > 0$:
  - If X_i has a positive effect on the latent variable, then by increasing X_i, fewer individuals will stay in category 0.
  - Similarly, more individuals will be in category 2.
  - In the intermediate category, the fraction of individual will either increase or decrease, depending on the relative size of the inflow from category 0 and the outflow to category 2.

Ordered Probit: Example

- We want to investigate the determinants of health.
- Individuals are asked to report their health status in three categories: poor, fair or good.
- We estimate an ordered probit and calculate the marginal effects at the mean of the sample.

Variable	Coeff	sd. err.	Marginal Effects			Sample
			Poor	Fair	Good	Mean
Age 18-30	-1.09**	.031	-.051**	-.196**	.248**	.25
Age 30-50	-.523**	.031	-.031**	-.109**	.141**	.32
Age 50-70	-.217**	.026	-.013**	-.046**	.060**	.24
Male	-.130**	.018	-.008**	-.028**	.037**	.48
Income low third	.428**	.027	.038**	.098**	-.136**	.33
Income medium third	.264**	.022	.020**	.059**	-.080**	.33
Education low	.40**	.028	.031**	.091**	-.122**	.43
Education Medium	.257**	.026	.018**	.057**	-.076**	.37
Year of interview	-.028	.018	-.001	-.006	.008	1.9
Household size	-.098**	.008	-.006**	-.021**	.028**	2.5
Alcohol consumed	.043**	.041	.002**	.009**	-.012**	.04
Current smoker	.160**	.018	.011**	.035**	-.046**	.49
cut1	.3992**	.058				
cut2	1.477**	.059				

Age group	Proportion		
	Poor Health	Fair Health	Good Health
Age 18-30	.01	.08	.90
Age 30-50	.03	.13	.83
Age 50-70	.07	.28	.64
Age 70 +	.15	.37	.46

Ordered Probit: Example

- Marginal Effects differ by individual characteristics.
- Below, we compare the marginal effects from an ordered probit and a multinomial logit.

	Marginal Effects for Good Health			
Variable	Ordered	X	Ordered	Multinomial
	Probit at mean		Probit at X	Logit at X
Age 18-30	.248**	1	.375**	.403**
Age 30-50	.141**	0	.093**	.077**
Age 50-70	.060**	0	.046**	.035**
Male	.037**	1	.033**	.031**
Income low third	-.136**	1	-.080**	-.066**
Income medium third	-.080**	0	-.071**	-.067**
Education low	-.122**	1	-.077**	-.067**
Education Medium	-.076**	0	-.069**	-.064**
Year of interview	.008	1	.006	.003
Household size	.028**	2	.023**	.020**
Alcohol consumed	-.012**	0	-.010**	-.011**
Current smoker	-.046**	0	-.041**	-.038**

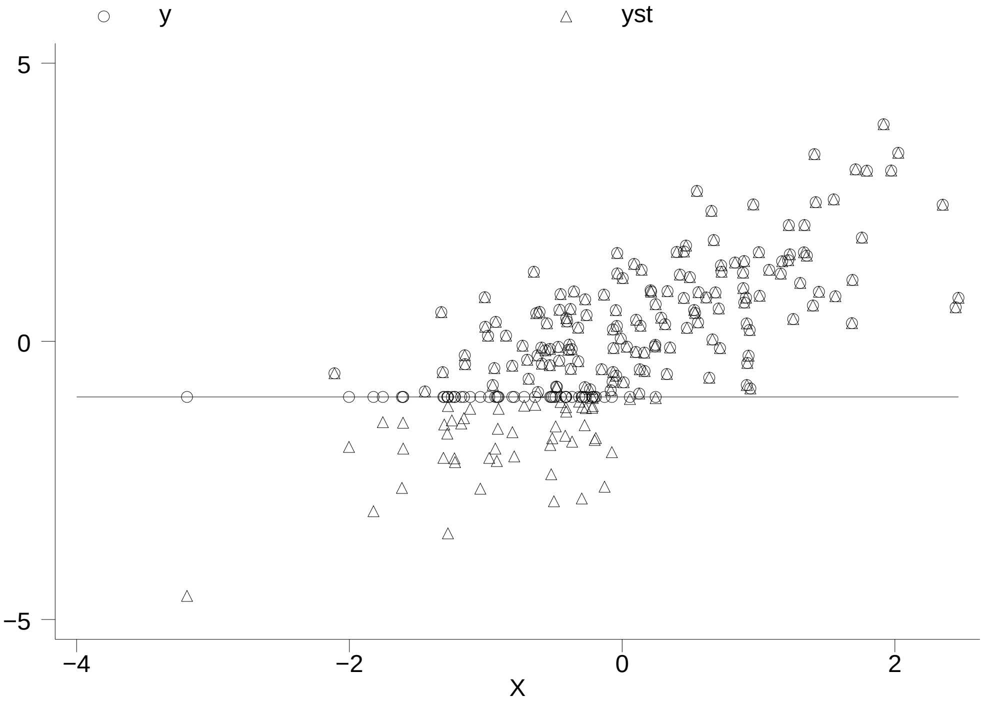
Tobit Model

- First proposed by Tobin (1958), [2]
- We define a latent (unobserved) variable Y^* such as:

$$Y^* = X\beta + \varepsilon \qquad \varepsilon \simeq \quad N(0, \sigma^2)$$

- We only observe a variable Y which is related to Y^* such as:

$$\begin{array}{lll} Y = Y^* & \text{if} & Y^* > a \\ Y = a & \text{if} & Y^* \leq a \end{array}$$



[2] Tobin, J. (1958), Estimation of Relationships for Limited Dependent Variables, *Econometrica 26, 24-36.*

Truncation Bias

- The conditional mean of Y given X takes the form:

$$E[Y|Y^* > a, X] = X\beta + \sigma \frac{\phi(\alpha)}{1-\Phi(\alpha)}$$

with $\alpha = \frac{a-X\beta}{\sigma}$. The ratio $\phi(\alpha)/(1-\Phi(\alpha))$ is called the inverse Mills ratio.

- Therefore, if you regress only the Ys which are above a on the corresponding Xs then, due to the latter term, the OLS parameters estimate of β will be biased and inconsistent.

- Proof: Note that the conditional c.d.f of $Y^*|Y^* > a$ is:

$$\begin{aligned} H(y|Y^* > a, X) &= P(Y^* \leq y | Y^* > a) = \frac{P(a < Y^* \leq y)}{P(Y^* > a)} \\ &= \frac{P(a - X\beta < \varepsilon \leq y - X\beta)}{P(\varepsilon > a - X\beta)} \\ &= \frac{\Phi(\frac{y-X\beta}{\sigma}) - \Phi(\frac{a-X\beta}{\sigma})}{1-\Phi(\frac{a-X\beta}{\sigma})} \end{aligned}$$

so that the conditional distribution is:

$$h(y|Y^* > a, X) = \frac{\partial H(y|Y^* > a, X)}{\partial y} = \frac{\phi(\frac{y-X\beta}{\sigma})}{\sigma(1-\Phi(\frac{a-X\beta}{\sigma}))}$$

$$\begin{aligned} E[Y|Y^* > a, X] &= \int_a^{+\infty} y h(y|Y^* > a, X) dy \\ &= \frac{1}{\sigma(1-\Phi(\alpha))} \int_a^{+\infty} y \phi(\frac{y - X\beta}{\sigma}) dy \\ &= \frac{1}{1-\Phi(\alpha)} \int_{(a-X\beta)/\sigma}^{+\infty} (X\beta + \sigma z) \phi(z) dz \\ &= \beta X - \frac{1}{1-\Phi(\alpha)} \sigma \int_{(a-X\beta)/\sigma}^{+\infty} \phi'(z) dz \\ &= X\beta + \sigma \frac{\phi(\alpha)}{1-\Phi(\alpha)} \end{aligned}$$

Tobit Model: Marginal Effects

- How do we interpret the coefficient β?

$$\beta = \frac{\partial Y^*}{\partial X}$$

  This is the marginal effect of X on the (latent) variable Y^*.

- Note that

$$E[Y|X] = X\beta\,(1 - \Phi(\alpha)) + \sigma\phi(\alpha)$$

  Therefore, if you treat the censored values of Y as regular dependent variable values in a linear regression model the OLS parameters estimate of β will be biased and inconsistent as well.

Likelihood for Tobit Model

- The conditional c.d.f of Y given X is:

$$\begin{aligned} G(y|X,\beta,\sigma) &= P(Y \leq y|X] \\ &= P(Y \leq y|X, Y > a)P(Y > a|X) \\ &\quad +P(Y \leq y|X, Y = a)P(P = a|X) \\ &= I(y > a)H(y|Y > a, X)(1 - \Phi(\frac{a - X\beta}{\sigma})) \\ &\quad +I(y = a)\Phi(\frac{a - X\beta}{\sigma}) \end{aligned}$$

where $I(.)$ is the indicator function: $I(true) = 1$, $I(false) = 0$.

- The corresponding conditional density is:

$$g(y|X,\beta,\sigma) = I(y > a)h(y|Y > a, X)(1 - \Phi(\frac{a - X\beta}{\sigma})) + I(y = a)\Phi(\frac{a - X\beta}{\sigma}$$

- The log-likelihood function of the Tobit model is:

$$\begin{aligned} l(\beta,\sigma) &= \sum_{i=1}^{n} \log(g(Y_i|X_i,\beta,\sigma)) \\ &= \sum_{i=1}^{n} I(y_i > a)\log(h(Y_j|Y_j > 0, X_j)) \\ &\quad +\sum_{i=1}^{n} I(y_i > a)(1 - \Phi(\frac{a - X_i\beta}{\sigma})) + \sum_{i=1}^{n} I(y_i = a)\log(\Phi(\frac{a - X\beta}{\sigma})) \\ &= \sum_{i=1}^{n} I(y_i > a)\left(-\frac{1}{2}(Y_i - X_i\beta)^2/\sigma^2 - \log(\sigma) - \log(\sqrt{2\pi})\right) \\ &\quad +\sum_{i=1}^{n} I(y_i = a)\log(\Phi(\frac{a - X_i\beta}{\sigma})) \end{aligned}$$

- This can be maximised with respect to β, σ or $\gamma = 1/\sigma$ and $\lambda = \beta/\sigma$.

Example: WTP

- The WTP is censored at zero. We can compare the two regressions:

OLS: $WTP_i = \beta_0 + \beta_1 lny + \beta_2 age_i + \beta_3 smell_i + u_i$

Tobit:
$$WTP_i^* = \beta_0 + \beta_1 lny + \beta_2 age_i + \beta_3 smell_i + u_i$$
$$WTP_i = WTP_i^* \quad \text{if} \quad WTP_i^* > 0$$
$$WTP_i = 0 \quad \text{if} \quad WTP_i^* < 0$$

	OLS		Tobit		
Variable	Estimate	t-stat	Estimate	t-stat	Marginal effect
lny	2.515	2.74	2.701	2.5	2.64
age	-.1155	-2.00	-.20651	-3.0	-0.19
sex	.4084	0.28	.14084	0.0	.137
smell	-1.427	-0.90	-1.8006	-0.9	-1.76
constant	-4.006	-0.50	-3.6817	-0.4	

Models for Count Data

- The methods developed above are useful when we want to model the occurrence or otherwise of an event. Sometimes we want to model the number of times an event occurs. In general it might be any nonnegative integer. Count data are being used increasingly in econometrics.

- An interesting application is to the modelling of the returns to R&D investment in which data on numbers of patents filed in a series of years by a sample of companies is studied and related to data on R&D investments.

- Binomial and Poisson probability models provide common starting points in the development of count data models.

- If $Z_1, \dots, Z_m$ are identically and independently distributed binary random variables with $P[Z_i = 1] = p$, $P[Z_i = 0] = 1 - p$, then the sum of the Z_i's has a Binomial distribution,

$$Y = \sum_{i=1}^{m} Z_i \sim Bi(m, p)$$

and

$$P[Y = j] = \frac{m!}{j!(m-j)!} p^j (1-p)^{m-j}, \qquad j \in \{0, 1, 2, \dots, m\}$$

Models for Count Data

- As m becomes large, $m^{1/2}(m^{-1}Y - p)$ becomes approximately normally distributed, $N(0, p(1-p))$, and as m becomes large while $mp = \lambda$ remains constant, Y comes to have a Poisson distribution,

  $$Y \sim Po(\lambda)$$

  and

  $$P[Y = j] = \frac{\lambda^j}{j!}\exp(-\lambda), \qquad j \in \{0, 1, 2, \dots\}.$$

- In each case letting p or λ be functions of covariates creates a model for the conditional distribution of a count of events given covariate values.

- The Poisson model is much more widely used, in part because there is no need to specify or estimate the parameter m.

- In the application to R&D investment one might imagine that a firm seeds a large number of research projects in a period of time, each of which has only a small probability of producing a patent. This is consonant with the Poisson probability model but note that one might be concerned about the underlying assumption of independence across projects built into the Poisson model.

Models for Count Data

- The estimation of the model proceeds by maximum likelihood. The Poisson model is used as an example. Suppose that we specify a single index model:

$$P[Y_i = y_i | x_i] = \frac{\lambda(x_i'\theta)^{y_i}}{y_i!} \exp(-\lambda(x_i'\theta)), \qquad j \in \{0, 1, 2, \dots\}.$$

- The log likelihood function is

$$l(\theta, y) = \sum_{i=1}^{n} y_i \log \lambda(x_i'\theta) - \lambda(x_i'\theta) - \log y_i!$$

with first derivative

$$\begin{aligned} l_\theta(\theta, y) &= \sum_{i=1}^{n} \left(y_i \frac{\lambda_w(x_i'\theta)}{\lambda(x_i'\theta)} - \lambda_w(x_i'\theta) \right) x_i \\ &= \sum_{i=1}^{n} (y_i - \lambda(x_i'\theta)) \frac{\lambda_w(x_i'\theta)}{\lambda(x_i'\theta)} x_i \end{aligned}$$

where $\lambda_w(w)$ is the derivative of $\lambda(w)$ with respect to w.

- The MLE satisfies

$$\sum_{i=1}^{n} \left(y_i - \lambda(x_i'\hat{\theta}) \right) \frac{\lambda_w(x_i'\hat{\theta})}{\lambda(x_i'\hat{\theta})} x_i = 0.$$

Models for Count Data

- The second derivative matrix is

$$l_{\theta\theta}(\theta, y) = \sum_{i=1}^{n} (y_i - \lambda(x_i'\theta)) \left(\frac{\lambda_{ww}(x_i'\theta)}{\lambda(x_i'\theta)} - \left(\frac{\lambda_w(x_i'\theta)}{\lambda(x_i'\theta)} \right)^2 \right) x_i x_i' - \sum_{i=1}^{n} \frac{\lambda_w(x_i'\theta)^2}{\lambda(x_i'\theta)} x_i x_i'$$

where, note, the first term has expected value zero. Therefore the Information Matrix for this conditional Poisson model is

$$I(\theta) = \sum_{i=1}^{n} \frac{\lambda_w(x_i'\theta)^2}{\lambda(x_i'\theta)} x_i x_i'.$$

The limiting distribution of the MLE is (under suitable conditions)

$$\begin{aligned} n^{1/2}(\hat{\theta} - \theta_0) &\xrightarrow{d} N(0, V_0) \\ V_0 &= \left(\operatorname*{plim}_{n\to\infty} n^{-1} \sum_{i=1}^{n} \frac{\lambda_w(x_i'\theta)^2}{\lambda(x_i'\theta)} x_i x_i' \right)^{-1} \end{aligned}$$

and we can make approximate inference about θ_0 using

$$(\hat{\theta} - \theta_0) \simeq N\left(0, n^{-1}V_0\right)$$

with V_0 estimated by

$$\hat{V}_0 = \left(n^{-1} \sum_{i=1}^{n} \frac{\lambda_w(x_i'\hat{\theta})^2}{\lambda(x_i'\hat{\theta})} x_i x_i' \right)^{-1}.$$

- In applied work a common choice is $\lambda(w) = \exp(w)$ for which

$$\frac{\lambda_w(w)}{\lambda(w)} = 1 \qquad \frac{\lambda_w(w)^2}{\lambda(w)} = \exp(w).$$

Likelihood Based Hypothesis Testing

Likelihood Based Hypothesis Testing

- We now consider test of hypotheses in econometric models in which the complete probability distribution of outcomes given conditioning variables is specified.
- There are three natural ways to develop tests of hypotheses when a likelihood function is available.

  1. Is the unrestricted ML estimator significantly far from the hypothesised value? This leads to what is known as the Wald test.
  2. If the ML estimator is restricted to satisfy the hypothesis, is the value of the maximised likelihood function significantly smaller than the value obtained when the restrictions of the hypothesis are not imposed? This leads to what is known as the likelihood ratio test.
  3. If the ML estimator is restricted to satisfy the hypothesis, are the Lagrange multipliers associated with the restrictions of the hypothesis significantly far from zero? This leads to what is known as the Lagrange multiplier or score test.

Likelihood Based Hypothesis Testing

- In the normal linear regression model all three approaches, after minor adjustments, lead to the same statistic which has an $F^{(j)}_{(n-k)}$ distribution when the null hypothesis is true and there are j restrictions.

- Outside that special case, in general the three methods lead to different statistics, but in large samples the differences tend to be small.

- All three statistics have, under certain weak conditions, $\chi^2_{(j)}$ limiting distributions when the null hypothesis is true and there are j restrictions.

- The exact distributional result in the normal linear regression model fits into this large sample theory on noting that $\operatorname{plim}_{n\to\infty}\left(jF^{(j)}_{(n-k)}\right) = \chi^2_{(j)}$.

Test of Hypothesis

- We now consider tests of a hypothesis $H_0 : \theta_2 = 0$ where the full parameter vector is partitioned into $\theta' = [\theta_1' \vdots \theta_2']$ and θ_2 contains j elements. Recall that the MLE has the approximate distribution

  $$n^{1/2}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V_0)$$

  where

  $$V_0 = -\operatorname*{plim}_{n \to \infty}(n^{-1} l_{\theta\theta}(\theta_0; Y))^{-1} = \overline{I}(\theta_0)^{-1}$$

  and $\overline{I}(\theta_0)$ is the asymptotic information matrix per observation.

Wald Test

- This test is obtained by making a direct comparison of $\hat{\theta}_2$ with the hypothesised value of θ_2, zero.
- Using the approximate distributional result given above leads to the following test statistic.

$$S_W = n\hat{\theta}_2'\widehat{W}_{22}^{-1}\hat{\theta}_2'$$

where $\widehat{W}_{22}$ is a consistent estimator of the lower right hand $j \times j$ block of V_0.

- Under the null hypothesis $S_W \xrightarrow{d} \chi^2_{(j)}$ and we reject the null hypothesis for large values of S_W.
- Using one of the formulas for the inverse of a partitioned matrix the Wald statistic can also be written as

$$S_W = n\hat{\theta}_2'\left(\widehat{\overline{I}}(\hat{\theta})_{22} - \widehat{\overline{I}}(\hat{\theta})'_{21}\widehat{\overline{I}}(\hat{\theta})^{-1}_{11}\widehat{\overline{I}}(\hat{\theta})_{12}\right)\hat{\theta}_2'$$

where the elements $\widehat{\overline{I}}(\hat{\theta})_{ij}$ are consistent estimators of the appropriate blocks of the asymptotic Information Matrix per observation evaluated at the (unrestricted) MLE.

The Score - or Lagrange Multiplier - test

- Sometimes we are in a situation where a model has been estimated with $\theta_2 = 0$, and we would like to see whether the model should be extended by adding additional parameters and perhaps associated conditioning variables or functions of ones already present.

- It is convenient to have a method of conducting a test of the hypothesis that the additional parameters are zero (in which case we might decide not to extend the model) without having to estimate the additional parameters. The *score test* provides such a method.

The Score - or Lagrange Multiplier - test

- The score test considers the gradient of the log likelihood function evaluated at the point

$$\hat{\theta}^R = [\hat{\theta}_1^{R\prime}, 0]'$$

and examines the departure from zero of that part of the gradient of the log likelihood function that is associated with θ_2.

- Here $\hat{\theta}_1^R$ is the MLE of θ_1 when θ_2 is restricted to be zero. If the unknown value of θ_2 is in fact zero then this part of the gradient should be close to zero. The score test statistic is

$$S_S = n^{-1} l_\theta(\hat{\theta}^R; Y)' \widehat{\bar{I}}(\hat{\theta}^R)^{-1} l_\theta(\hat{\theta}^R; Y)$$

and $S_S \xrightarrow{d} \chi^2_{(j)}$ under the null hypothesis. There are a variety of ways of estimating $\widehat{\bar{I}}(\theta_0)$ and hence its inverse.

- Note that the complete score (gradient) vector appears in this formula. Of course the part of that associated with θ_1 is zero because we are evaluating at the restricted MLE. That means the score statistic can also be written, using the formula for the inverse of a partitioned matrix, as the algebraically identical

$$S_S = n^{-1} l_{\theta_2}(\hat{\theta}^R; Y)' \left(\widehat{\bar{I}}(\hat{\theta}^R)_{22} - \widehat{\bar{I}}(\hat{\theta}^R)_{21}' \widehat{\bar{I}}(\hat{\theta}^R)_{11}^{-1} \widehat{\bar{I}}(\hat{\theta}^R)_{12} \right)^{-1} l_{\theta_2}(\hat{\theta}^R; Y).$$

- When the information matrix is block diagonal, which means that the MLEs of θ_1 and θ_2 are asymptotically uncorrelated, the second term in the inverse above vanishes.

Likelihood ratio tests

- The final method for constructing hypothesis tests that we will consider involves comparing the value of the maximised likelihood function at the restricted MLE ($\hat{\theta}^R$) and the unrestricted MLE (now written as $\hat{\theta}^U$).
- This likelihood ratio test statistic takes the form

$$S_L = 2\left(l(\hat{\theta}^U;Y) - l(\hat{\theta}^R;Y)\right)$$

and it can be shown that under H_0, $S_L \xrightarrow{d} \chi^2_{(j)}$.

Specification Testing

- Maximum likelihood estimation requires a complete specification of the probability distribution of the random variables whose realisations we observe.

- In practice we do not *know* this distribution though we may be able to make a good guess. If our guess is badly wrong then we may produce poor quality estimates, for example badly biased estimates, and the inferences we draw using the properties of the likelihood function may be incorrect.

- In regression models the same sorts of problems occur. If there is heteroskedasticity or serial correlation then, though we may produce reasonable point estimates of regression coefficients if we ignore these features of the data generating process, our inferences will usually be incorrect if these features are not allowed for, because we will use incorrect formulae for standard errors and so forth.

- It is important then to seek for evidence of departure from a model specification, that is to conduct *specification tests*.

- In a likelihood context the score test provides an easy way of generating specification tests.

- The score specification test does not tell us exactly how the model should be extended.

Detecting Heteroskedasticity

- We consider one example here, namely detecting heteroskedasticity in a normal linear regression model.
- In the model considered, $Y_{1,} \ldots, Y_n$ are independently distributed with Y_i given x_i being $N(x_i'\beta, \sigma^2 h(z_i'\alpha))$ where $h(0) = 1$ and $h'(0) = 1$, both achievable by suitable scaling of $h(\cdot)$.
- Let $\theta^U = [\beta, \sigma^2, \alpha]$ and let $\theta^R = [\beta, \sigma^2, 0]$. A score test of $H_0 : \alpha = 0$ will provide a specification test to detect heteroskedasticity.
- The log likelihood function when $\alpha = 0$, in which case there is homoskedasticity, is as follows.

$$l(\theta^R; y|x) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2$$

whose gradients with respect to β and σ^2 are

$$\begin{aligned} l_\beta(\theta^R; y|x) &= -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)\,x_i \\ l_{\sigma^2}(\theta^R; y|x) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 \end{aligned}$$

which lead to the restricted MLEs under homoskedasticity, as follows.

$$\begin{aligned} \hat\beta &= (X'X)^{-1} X'y \\ \hat\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i'\hat\beta\right)^2 \end{aligned}$$

Detecting Heteroskedasticity

- The log likelihood function for the unrestricted model is

$$l(\theta^U; y|x) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\log h(z_i'\alpha) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{(y_i - x_i'\beta)^2}{h(z_i'\alpha)}$$

whose gradient with respect to α is

$$l_\alpha(\theta^U; y|x) = -\frac{1}{2}\sum_{i=1}^{n}\frac{h'(z_i'\alpha)}{h(z_i'\alpha)}z_i + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{(y_i - x_i'\beta)^2\, h'(z_i'\alpha)}{h(z_i'\alpha)^2}z_i$$

which evaluated at the restricted MLE (for which $\alpha = 0$) is

$$\begin{aligned} l_\alpha(\hat{\theta}^R; y|x) &= -\frac{1}{2}\sum_{i=1}^{n} z_i + \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}\left(y_i - x_i'\hat{\beta}\right)^2 z_i \\ &= \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}\left(\hat{\varepsilon}_i^2 - \hat{\sigma}^2\right) z_i. \end{aligned}$$

- The specification test examines the correlation between the squared OLS residuals and z_i. The score test will lead to rejection when this correlation is large.

- Details of calculation of this test are given in the intermediate textbooks and the test (Breusch-Pagan-Godfrey) is built into many of the econometric software packages.

- Note that the form of the function $h(\cdot)$ does not figure in the score test. This would not be the case had we developed either a Wald test or a Likelihood Ratio test.

Information Matrix Tests

- We have seen that the results on the limiting distribution of the MLE rest at one point on the Information Matrix Equality

$$E[l_\theta(\theta_0,Y)l_\theta(\theta_0,Y)'] = -E[l_{\theta\theta'}(\theta_0,Y)]$$

  where $Y = (Y_1,\ldots,Y_n)$ are n random variables whose realisations constitute our data.

- In the case relevant to much microeconometric work the log likelihood function is a sum of independently distributed random variables, e.g. in the continuous Y case:

$$l(\theta,Y) = \sum_{i=1}^{n} \log f(Y_i,\theta),$$

  where $f(Y_i,\theta)$ is the probability density function of Y_i. Here the Information Matrix Equality derives from the result

$$E[\frac{\partial}{\partial\theta}\log f(Y,\theta)\frac{\partial}{\partial\theta'}\log f(Y,\theta) + \frac{\partial^2}{\partial\theta\partial\theta'}\log f(Y,\theta)] = 0.$$

- Given a value $\hat{\theta}$ of the MLE we can calculate a sample analogue of the left hand side of this equation:

$$IM = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}\log f(Y_i,\theta)\frac{\partial}{\partial\theta'}\log f(Y_i,\theta) + \frac{\partial^2}{\partial\theta\partial\theta'}\log f(Y_i,\theta)|_{\theta=\hat{\theta}}\right)$$

- If the likelihood function is a correct specification for the data generating process, then we expect the resulting statistic (which is a matrix of values unless θi is scalar) to be close to (a matrix of zeros).

- A general purpose statistic for detecting incorrect specification of a likelihood function is produced by considering a quadratic form in a vectorised version of all or part of $n^{1/2}IM$. This Information Matrix Test statistic was introduced by Halbert White[3] in 1982.

[3] See "Maximum Likelihood Estimation in Misspecified Models", Halbert White

Endogeneity and Instrumental Variables

Endogeneity and Simultaneity

- In many problems studied in econometrics it is *not* possible to maintain restrictions requiring that the expected value of the latent variable in an equation is zero given the values of the right hand side variables in the equation:

$$E(\varepsilon|X) \neq 0$$

- This leads to a biased OLS estimate.
- There are many cases in which the OLS identification assumption does not hold:
  - simultaneous equations.
  - explanatory variables measured with error.
  - omitted variables correlated with explanatory variables.

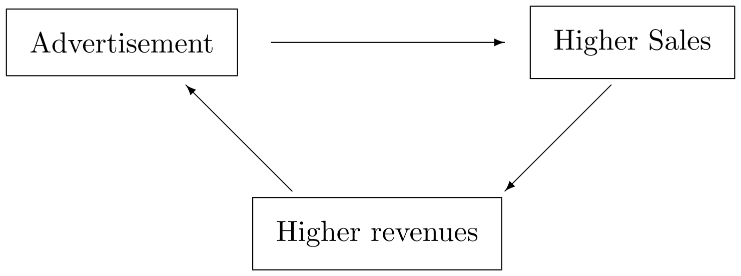
Simultaneity

- **Definition:** Simultaneity arises when the causal relationship between Y and X runs both ways. In other words, the explanatory variable X is a function of the dependent variable Y, which in turn is a function of X.

$$\begin{tikzcd}
Y \arrow[r, bend right, "\text{Indirect Effect}"'] & X \arrow[l, bend right, "\text{Direct effect}"']
\end{tikzcd}$$

- This arises in many economic examples:
  - Income and health.
  - Sales and advertizing.
  - Investment and productivity.
- What are we estimating when we run an OLS regression of Y on X? Is it the direct effect, the indirect effect or a mixture of both.

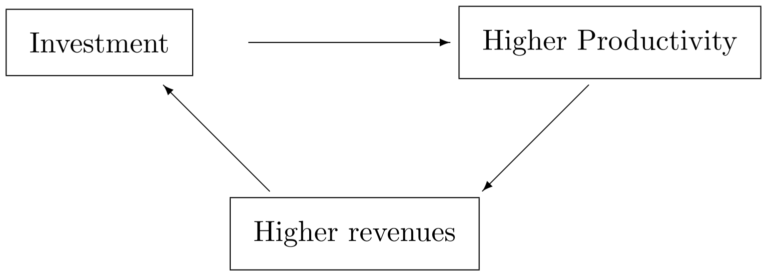
Examples



Advertisement

Higher Sales

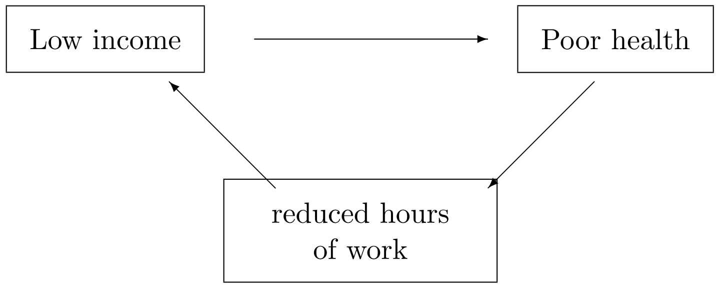
Higher revenues



Investment

Higher Productivity

Higher revenues



Low income

Poor health

reduced hours of work

Implications of Simultaneity

- $$\begin{cases} Y_i = \beta_0 + \beta_1 X_i + u_i & \text{(direct effect)} \\ \\ X_i = \alpha_0 + \alpha_1 Y_i + v_i & \text{(indirect effect)} \end{cases}$$

- Replacing the second equation in the first one, we get an equation expressing Y_i as a function of the parameters and the error terms u_i and v_i only. Substituting this into the second equation, we get X_i also as a function of the parameters and the error terms:

$$\begin{cases} Y_i = \dfrac{\beta_0 + \beta_1 \alpha_0}{1 - \alpha 1 \beta_1} + \dfrac{\beta_1 v_i + u_i}{1 - \alpha_1 \beta_1} = B_0 + \tilde{u}_i \\ \\ X_i = \dfrac{\alpha_0 + \alpha_1 \beta_0}{1 - \alpha_1 \beta_1} + \dfrac{v_i + \alpha_1 u_i}{1 - \alpha_1 \beta_1} = A_0 + \tilde{v}_i \end{cases}$$

- This is the **reduced form** of our model. In this rewritten model, Y_i is not a function of X_i and vice versa. However, Y_i and X_i are both a function of the two original error terms u_i and v_i.

- Now that we have an expression for X_i, we can compute:

$$\begin{aligned} cov(X_i, u_i) &= cov(\frac{\alpha_0 + \alpha_1 \beta_0}{1 - \alpha_1 \beta_1} + \frac{v_i + \alpha_1 u_i}{1 - \alpha_1 \beta_1}, u_i) \\ &= \frac{\alpha_1}{1 - \alpha_1 \beta_1} \; Var(u_i) \end{aligned}$$

  which, in general is different from zero. Hence, with simultaneity, our assumption 1 is violated. **An OLS regression of Y_i on X_i will lead to a <u>biased</u> estimate of β_1.** Similarly, an OLS regression of X_i on Y_i will lead to a biased estimate of α_1.

What are we estimating?

- For the model:
$$Y_i = \beta_0 + \beta_1 + X_i + u_i$$

- The OLS estimate is:
$$\begin{aligned}\hat{\beta}_1 &= \beta_1 + \frac{cov(X_i, u_i)}{Var(X_i)} \\ &= \beta_1 + \frac{\alpha_1}{1 - \alpha_1\beta_1}\frac{Var(u_i)}{Var(X_i)}\end{aligned}$$

- So
  - $E\hat{\beta}_1 \neq \beta_1$
  - $E\hat{\beta}_1 \neq \alpha_1$
  - $E\hat{\beta}_1 \neq$ an average of β_1 and α_1.

Identification

- Suppose a more general model:

$$\begin{cases} Y_i = \beta_0 + \beta_1 X_i + \beta_2 T_i + u_i \\ X_i = \alpha_0 + \alpha_1 Y_i + \alpha_2 Z_i + v_i \end{cases}$$

- We have two sorts of variables:
  - **Endogenous:** Y_i and X_i because they are determined within the system. They appear on the right and left hand side.
  - **Exogenous:** T_i and Z_i. They are determined outside of our model, and in particular are not caused by either X_i or Y_i. They appear only on the right-hand-side.

Example

- Consider a simple version of the Mincer model for returns to schooling with the following *structural equations*.

$$\begin{aligned} W &= \alpha_0 + \alpha_1 S + \alpha_2 Z + \varepsilon_1 \\ S &= \beta_0 + \beta_1 Z + \varepsilon_2 \end{aligned}$$

  Here W is the log wage, S is years of schooling, Z is some characteristic of the individual, and ε_1 and ε_2 are unobservable latent random variables.

- We might expect those who receive unusually high levels of schooling given Z to also receive unusually high wages given Z and S, a situation that would arise if ε_1 and ε_2 were affected positively by ability, a characteristic not completely captured by variation in Z.

- In this problem we might be prepared to impose the following restrictions.

$$\begin{aligned} E[\varepsilon_1|Z = z] &= 0 \\ E[\varepsilon_2|Z = z] &= 0 \end{aligned}$$

  but not

$$E[\varepsilon_1|S = s, Z = z] = 0$$

  unless ε_1 was believed to be uncorrelated with ε_2.

- Considering just the first (W) equation,

$$E[W|S = s, Z = z] = \alpha_0 + \alpha_1 s + \alpha_2 z + E[\varepsilon_1|S = s, Z = z]$$

- A variable like S, appearing in a structural form equation and correlated with the latent variable in the equation, is called an *endogenous* variable.

Reduced Form Equations

- Substitute for S in the wage equation:

$$\begin{aligned} W &= (\alpha_0 + \alpha_1\beta_0) + (\alpha_1\beta_1 + \alpha_2)\, Z + \varepsilon_1 + \alpha_1\varepsilon_2 \\ S &= \beta_0 + \beta_1 Z + \varepsilon_2 \end{aligned}$$

- Equations like this, in which each equation involves exactly one endogenous variable are called *reduced form* equations.
- The restrictions $E[\varepsilon_1|Z = z] = 0$ and $E[\varepsilon_2|Z = z] = 0$ imply that

$$\begin{aligned} E[W|Z &= z] = (\alpha_0 + \alpha_1\beta_0) + (\alpha_1\beta_1 + \alpha_2)\, z \\ E[S|Z &= z] = \beta_0 + \beta_1 z \end{aligned}$$

- Given enough (at least 2) distinct values of z and knowledge of the left hand side quantities we can solve for $(\alpha_0 + \alpha_1\beta_0)$, $(\alpha_1\beta_1 + \alpha_2)$, β_0 and β_1. So, the values of these *functions* of parameters of the structural equations *can* be identified.
- In practice we do not know the left hand side quantities but with enough data we can estimate the data generating values of $(\alpha_0 + \alpha_1\beta_0)$, $(\alpha_1\beta_1 + \alpha_2)$, β_0 and β_1, for example by OLS applied first to (W, Z) data and then to (S, Z) data.
- The values of β_0 and β_1 *are* identified but the values of α_0, α_1 and α_2 are *not*, for without further restrictions their values cannot be deduced from knowledge of $(\alpha_0 + \alpha_1\beta_0)$, $(\alpha_1\beta_1 + \alpha_2)$, β_0.

Identification using an Exclusion Restriction

- One restriction we might be prepared to add to the model is the restriction $\alpha_2 = 0$. Whether or not that is a reasonable restriction to maintain depends on the nature of the variable Z.

- If Z were a measure of some characteristic of the environment of the person at the time that schooling decisions were made (for example the parents' income, or some measure of an event that perturbed the schooling choice) then we might be prepared to maintain the restriction that, given schooling achieved (S), Z does not affect W, i.e. that $\alpha_2 = 0$.

- This restriction may be sufficient to identify the remaining parameters. If the restriction is true then the coefficients on Z become $\alpha_1\beta_1$.

- We have already seen that (the value of) the coefficient β_1 is identified. If β_1 is not itself zero (that is Z does indeed affect years of schooling) then α_1 is identified as the ratio of the coefficients on Z in the regressions of W and S on Z. With α_1 identified and β_0 already identified, identification of α_0 follows directly.

Indirect Least Squares Estimation

- Estimation could proceed under the restriction $\alpha_2 = 0$ by calculating OLS (or GLS) estimates of the "reduced form" equations:

$$\begin{aligned} W &= \pi_{01} + \pi_{11} Z + U_1 \\ S &= \pi_{02} + \pi_{12} Z + U_2 \end{aligned}$$

where

$$\begin{aligned} \pi_{01} &= \alpha_0 + \alpha_1 \beta_0 & \pi_{11} &= \alpha_1 \beta_1 \\ \pi_{02} &= \beta_0 & \pi_{12} &= \beta_1 \\ U_1 &= \varepsilon_1 + \alpha_1 \varepsilon_2 & U_2 &= \varepsilon_2 \end{aligned}$$

and

$$E[U_1|Z = z] = 0 \quad E[U_2|Z = z] = 0$$

solving the equations:

$$\begin{aligned} \hat{\pi}_{01} &= \hat{\alpha}_0 + \hat{\alpha}_1 \hat{\beta}_0 & \hat{\pi}_{11} &= \hat{\alpha}_1 \hat{\beta}_1 \\ \hat{\pi}_{02} &= \hat{\beta}_0 & \hat{\pi}_{12} &= \hat{\beta}_1 \end{aligned}$$

given values of the $\hat{\pi}$'s for values of the $\hat{\alpha}$'s and $\hat{\beta}$'s, as follows.

$$\begin{aligned} \hat{\alpha}_0 &= \hat{\pi}_{01} - \hat{\pi}_{02} \left(\hat{\pi}_{11} / \hat{\pi}_{12} \right) & \hat{\alpha}_1 &= \hat{\pi}_{11} / \hat{\pi}_{12} \\ \hat{\beta}_0 &= \hat{\pi}_{02} & \hat{\beta}_1 &= \hat{\pi}_{12} \end{aligned}$$

- Estimators obtained in this way, by solving the equations relating structural form parameters to reduced form parameters with OLS estimates replacing the reduced form parameters, are known as *Indirect Least Squares* estimators. They were first proposed by Jan Tinbergen in 1930.

Over Identification

- Suppose that there are *two* covariates, Z_1 and Z_2 whose impact on the structural equations we are prepared to restrict so that *both* affect schooling choice but *neither* affect the wage given the amount of schooling achieved:

$$\begin{aligned} W &= \alpha_0 + \alpha_1 S + \varepsilon_1 \\ S &= \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \varepsilon_2 \end{aligned}$$

- the reduced form equations are as follows

$$\begin{aligned} W &= \pi_{01} + \pi_{11} Z_1 + \pi_{21} Z_2 + U_1 \\ S &= \pi_{02} + \pi_{12} Z_1 + \pi_{22} Z_2 + U_2 \end{aligned}$$

  where

$$\begin{array}{lllllllll} \pi_{01} &=& \alpha_0 + \alpha_1 \beta_0 & \pi_{11} &=& \alpha_1 \beta_1 & \pi_{21} &=& \alpha_1 \beta_2 \\ \pi_{02} &=& \beta_0 & \pi_{12} &=& \beta_1 & \pi_{22} &=& \beta_2 \end{array}$$

  and

$$U_1 = \varepsilon_1 + \alpha_1 \varepsilon_2 \qquad U_2 = \varepsilon_2 .$$

- The values of the reduced form equations' coefficients are identified under restrictions .

- Note, there are *two* ways in which the coefficient α_1 can be identified, as follows

$$\alpha_1 = \alpha_1^{Z_1} = \frac{\pi_{11}}{\pi_{12}} \qquad \alpha_1 = \alpha_1^{Z_2} = \frac{\pi_{21}}{\pi_{22}}$$

- In this situation we say that the value of the parameter α_1 is *over identified*.

- We will usually find that $\hat{\alpha}_1^{Z_1} \neq \hat{\alpha}_1^{Z_2}$ even though these are both estimates of the value of the same structural form parameter.

- If the discrepancy was found to be very large then we might doubt whether the restrictions of the model are correct. This suggests that tests of over identifying restrictions can detect misspecification of the econometric model.

- If the discrepancy is not large then there is scope for combining the estimates to produce a single estimate that is more efficient than either taken alone.

Instrumental Variables

- Consider the linear model for an outcome Y given covariates X

$$Y = X\beta + \varepsilon$$

- Suppose that the restriction $E[\varepsilon|X = x] = 0$ cannot be maintained but that there exist m variables Z for which the restriction $E[\varepsilon|Z = z] = 0$ can be maintained. It implies:

$$E[Y - X\beta|Z = z] = 0$$

and thus that

$$E[Z'(Y - X\beta)|Z = z] = 0$$

which implies that, unconditionally

$$E[Z'(Y - X\beta)] = 0.$$

and thus

$$E[Z'Y] = E[Z'X]\beta.$$

- First suppose $m = k$, and that $E[Z'X]$ has rank k. Then β can be expressed in terms of moments of Y, X and Z as follows

$$\beta = E[Z'X]^{-1}E[Z'Y].$$

and β is (just) identifiable. This leads directly to an analogue type estimator:

$$\hat{\beta} = (Z'X)^{-1}(Z'Y)$$

In the context of the just identified returns to schooling model this is the Indirect Least Squares estimator.

Generalised Method of Moments estimation

- Suppose that $m > k$. We will not find a solution since we have $m > k$ equations in k unknowns.
- Define a family of estimators, $\hat{\beta}_W$ as

$$\hat{\beta}_W = \arg\min_{\beta} \left(Z'Y - Z'X\beta\right)' W \left(Z'Y - Z'X\beta\right)$$

  where W is a $m \times m$ full rank, positive definite symmetric matrix.
- This M-estimator is an example of what is known as the *Generalised Method of Moments* (GMM) estimator.
- Different choices of W lead to different estimators unless $m = k$.
- The choice among these is commonly made by considering their accuracy. We consider the limiting distribution of the GMM estimator for alternative choices of W and choose W to minimise the variance of the limiting distribution of $n^{1/2}(\hat{\beta}_W - \beta_0)$.
- In standard cases this means choosing W to be proportional to a consistent estimator of the inverse of the variance of the limiting distribution of $n^{1/2}\left(Z'Y - Z'X\beta\right)$.

Generalised Instrumental Variables Estimation

- Write $\hat{\beta}_W$ explicitly in terms of sample moments:

$$\hat{\beta}_W = \arg\min_{\beta} \left(\frac{Z_n' y_n - Z_n' X_n \beta}{n^{1/2}} \right)' W \left(\frac{Z_n' y_n - Z_n' X_n \beta}{n^{1/2}} \right)$$

- Consider what the (asymptotically) efficient choice of W is by examining the variance of $n^{-1/2}(Z_n' y_n - Z_n' X_n \beta)$.

- We have, since $y_n = X_n \beta + \varepsilon_n$,

$$n^{-1/2}(Z_n' y_n) - n^{-1/2}(Z_n' X_n)\beta = n^{-1/2}(Z_n' \varepsilon_n)$$

and *if* we suppose that $Var(\varepsilon_n | Z_n) = \sigma^2 I_n$,

$$Var\left(n^{-1/2}(Z_n' \varepsilon_n) | Z_n\right) = \sigma^2 (n^{-1} Z_n' Z_n).$$

This suggests choosing $W = (n^{-1} Z_n' Z_n)^{-1}$ leading to the following minimisation problem:

$$\hat{\beta}_n \quad = \quad \arg\min_{\beta} \left(Z_n' y_n - Z_n' X_n \beta\right)' \left(Z_n' Z_n\right)^{-1} \left(Z_n' y_n - Z_n' X_n \beta\right)$$

- The first order conditions for this problem, satisfied by $\hat{\beta}_n$ are:

$$2\hat{\beta}_n'(X_n' Z_n)(Z_n' Z_n)^{-1}(Z_n' X_n) - 2(X_n' Z_n)(Z_n' Z_n)^{-1}(Z_n' y_n) = 0$$

leading to the following estimator.

$$\hat{\beta} = \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} X'Z(Z'Z)^{-1}Z'y$$

This is known as the *generalised instrumental variable estimator* (GIVE).

GIVE: Asymptotic Properties

- The asymptotic properties of this estimator are obtained as follows. Substituting $y_n = X_n\beta + \varepsilon_n$ gives

$$\begin{aligned}\hat{\beta}_n &= \beta + \left(X_n'Z_n(Z_n'Z_n)^{-1}Z_n'X_n\right)^{-1} X_n'Z_n(Z_n'Z_n)^{-1}Z_n'\varepsilon_n \\ &= \beta + \left(n^{-1}X_n'Z_n(n^{-1}Z_n'Z_n)^{-1}n^{-1}Z_n'X_n\right)^{-1} n^{-1}X_n'Z_n(n^{-1}Z_n'Z_n)^{-1}n^{-1}Z_n'\varepsilon_n\end{aligned}$$

and if

$$\begin{aligned}\operatorname*{plim}_{n\to\infty}(n^{-1}Z_n'Z_n) &= \Sigma_{ZZ} \\ \operatorname*{plim}_{n\to\infty}(n^{-1}X_n'Z_n) &= \Sigma_{XZ} \\ \operatorname*{plim}_{n\to\infty}(n^{-1}Z_n'\varepsilon_n) &= 0\end{aligned}$$

with Σ_{ZZ} having full rank (m) and Σ_{XZ} having full rank (k) then

$$\operatorname*{plim}_{n\to\infty}\hat{\beta}_n = \beta$$

and we have a *consistent* estimator.

GIVE Asymptotic Properties

- To obtain the limiting distribution of $n^{1/2}(\hat{\beta}-\beta)$ note that

$$\begin{aligned} n^{1/2}(\hat{\beta}-\beta) &= \left(n^{-1}X_n'Z_n(n^{-1}Z_n'Z_n)^{-1}n^{-1}Z_n'X_n\right)^{-1} \\ &\qquad n^{-1}X_n'Z_n(n^{-1}Z_n'Z_n)^{-1}n^{-1/2}Z_n'\varepsilon_n \end{aligned}$$

- Under the conditions in the previous slide we have the limiting distribution:

  $$\operatorname{plim} n^{1/2}(\hat{\beta}-\beta) = \left(\Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}\right)^{-1}\Sigma_{XZ}\Sigma_{ZZ}^{-1}\operatorname{plim}\left(n^{-1/2}Z_n'\varepsilon_n\right)$$

  where $\Sigma_{ZX} = \Sigma_{XZ}'$, and if a Central Limit Theorem applies to $n^{-1/2}Z_n'\varepsilon_n$

  $$\operatorname{plim}\left(n^{-1/2}Z_n'\varepsilon_n\right) = N(0,\sigma^2\Sigma_{ZZ})$$

  then

  $$\operatorname{plim} n^{1/2}(\hat{\beta}-\beta) = N(0,V)$$

  where

  $$\begin{aligned} V &= \sigma^2\left(\Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{XZ}\right)^{-1}\Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}\left(\Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}\right)^{-1} \\ &= \sigma^2\left(\Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}\right)^{-1} \end{aligned}$$

  and so

  $$\operatorname{plim} n^{1/2}(\hat{\beta}-\beta) \simeq N(0,\sigma^2\left(\Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}\right)^{-1}).$$

GIVE and Two Stage OLS

- Suppose there is a model for X,

  $$X = Z\Phi + V$$

  where $E[V|Z] = 0$. The OLS estimator of Φ is

  $$\hat{\Phi}_n = \left(Z_n'Z_n\right)^{-1} Z_n'X_n$$

  and the "predicted value" of X for a given Z is

  $$\hat{X}_n = Z_n \left(Z_n'Z_n\right)^{-1} Z_n'X_n.$$

  Note that

  $$\hat{X}_n'\hat{X}_n = X_n'Z_n(Z_n'Z_n)^{-1}Z_n'X_n$$

  and

  $$\hat{X}_n'y_n = X_n'Z_n(Z_n'Z_n)^{-1}Z_n'y_n.$$

  So the Generalised Instrumental Variables Estimator can be written as

  $$\hat{\beta}_n = \left(\hat{X}_n'\hat{X}_n\right)^{-1} \hat{X}_n'y_n.$$

  that is, as the OLS estimator of the coefficients of a linear relationship between y_n and the *predicted values* of X_n got from OLS estimation of a linear relationship between X_n and the instrumental variables Z_n.

Examples: Measurement Errors

- Suppose we are measuring the impact of income, X, on consumption, Y. The true model is:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$\beta_0 = 0, \quad \beta_1 = 1$$

- Suppose we have two measures of income, both with measurement errors.
  - $\check{X}_{1i} = X_i + v_{1i}, \quad s.d.(v_{1i}) = 0.2 * \bar{Y}$
  - $\check{X}_{2i} = X_i + v_{2i}, \quad s.d.(v_{2i}) = 0.4 * \bar{Y}$

  If we use $\check{X}_2$ to instrument $\check{X}_1$, we get:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (\check{X}_{2i} - \bar{\check{X}}_2)(Y_i - \bar{Y})}{\sum_{i=1}^{N} (\check{X}_{2i} - \bar{\check{X}}_2)(\check{X}_{1i} - \bar{\check{X}}_1)}$$

- Results:

Method	Estimate of β_1
OLS regressing Y on $\check{X}_1$	0.88
OLS regressing Y on $\check{X}_2$	0.68
IV, using $\check{X}_2$ as instrument	0.99