Lecture 1
Least Squares

1

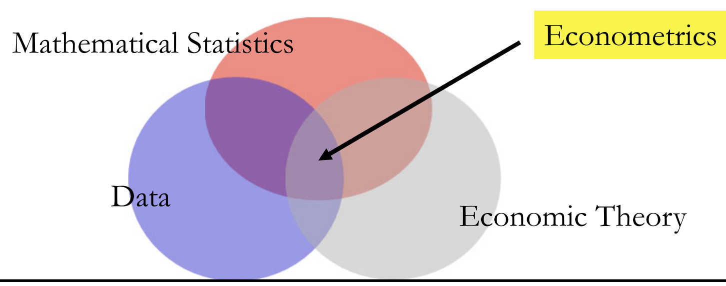
(for private use, not to be posted/shared online)

What is Econometrics?

• Ragnar Frisch, Econometrica Vol.1 No. 1 (1933) revisited

"Experience has shown that each of these three view-points, that of *statistics*, *economic theory*, and *mathematics*, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life.

It is the unification of all three aspects that is powerful. And it is this unification that constitutes econometrics."



Not to be posted/distributed online without written consent from author.

What is Econometrics?

- Economic Theory:

- The CAPM: $E[r_i - r_f] = \beta_i \, E[(r_M - r_f)]$

- Mathematical Statistics:

- Method to estimate CAPM. For example,

  Linear regression: $r_i - r_f = \alpha_i + \beta_i \, (r_M - r_f) + \varepsilon_i$

- Properties of $\mathbf{b}_i$ (the LS estimator of β_i)

- Properties of different tests of CAPM. For example, a t-test for H_0: $\alpha_i = 0$.

- Data: r_i, r_f, and r_M

- Typical problems: Missing data, Measurement errors, Survivorship bias, Auto- and Cross-correlated returns, Time-varying moments.

Data: Population and Sample

Definition: Sample

The *sample* is a (manageable) subset of elements of the population.

Example: The total returns of the stocks on the S&P 500 index.

Population (DGP) $\xrightarrow{\text{Get a sample}}$ *Sample* (size N or T)

Samples are collected to learn about the population. The process of collecting information from a sample is referred to as *sampling*.

Definition: Random Sample

A *random sample* is a sample where the probability that any individual member from the population being selected as part of the sample is exactly the same as any other individual member of the population.

4

Data: Population and Sample

Example: The total returns of the stocks on the S&P 500 index is *not* a random sample of stock returns.

In mathematical terms, given a random variable X with distribution F, a *random sample* of length N is a set of N independent, identically distributed (*i.i.d.*) random variables with distribution F.

• We will estimate population parameters using sample analogues: mean, sample mean; variance, sample variance; $\boldsymbol{\beta}$, $\mathbf{b}$; etc.

• In general, in finance and economics, we do not deal with random samples. The collected observations will have issues that make the sample not a truly a random sample.

5

Data: Samples and Types of Data

• The samples we collect are classified in three groups:

• **Time Series Data**: Collected over time on one or more variables, with a particular *frequency* of observation.

Example: We record for 10 years the monthly S&P 500 returns, or 10' IBM returns.

Usual notation: x_t, $t = 1, 2, \ldots, T$.

• **Cross-sectional Data**: Collected on one or more variables collected at a single point in time.

Example: Today we record all closing returns for the members of the S&P 500 index.

Usual notation: x_i, $i = 1, 2, \ldots, N$.

6

Not to be posted/distributed online without written consent from author.

Data: Samples and Types of Data

• **Panel Data**: Cross-sectional data collected over time.

Example: The CRSP database collects daily prices of all U.S. traded stocks since 1962.

Usual notation: $x_{i,t}$, $\quad i = 1, 2, \ldots, N$ & $t = 1, 2, \ldots, T$.

• The different types of data will present different problems; for example, autocorrelation is a common problem in time series, while cross-correlation is a common problem in cross-sections.

7

Estimation

• Two philosophies regarding models (assumptions) in statistics:

(1) Parametric statistics.

It assumes data come from a type of probability distribution and makes inferences about the parameters of the distribution. Models are parameterized before collecting the data.

Example: Maximum likelihood estimation.

(2) Non-parametric statistics.

It assumes no probability distribution –i.e., they are "distribution free." Models are not imposed *a priori,* but determined by the data.

Examples: histograms, kernel density estimation.

• In general, in parametric statistics we make more assumptions.

Least Squares Estimation

- Old method: Gauss (1795, 1801) used it in astronomy.

Carl F. Gauss (1777 – 1855, Germany)

Idea:

- We model the behavior of a dependent variable y as a function of k explanatory variables $\boldsymbol{x}$. This function depends on q unknown parameters, $\boldsymbol{\theta}$. The relation between y and $\boldsymbol{x}$ is not exact; there is an error, ε. We have T observations of Y and **X**.

- We assume that the functional form is known. The model is:

$$y_i = f(x_{1,i}, x_{2,i}, \ldots, x_{k,i}; \boldsymbol{\theta}) + \varepsilon_i, \qquad i = 1, 2, \ldots, T.$$

- We estimate $\boldsymbol{\theta}$ by minimizing a sum of squared errors:

$$\min_\theta \{S(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{T} \varepsilon_i^{\,2} = \sum_{i=1}^{T} (y_i - f(x_{1,i}, x_{2,i}, \ldots, x_{k,i}; \boldsymbol{\theta}))^2\}$$

Least Squares Estimation: OLS

- The estimator obtained is called the *Least Squares* (LS) estimator.

- LS is a general estimation method. It can be applied to almost any function $f(\boldsymbol{x}_i, \boldsymbol{\theta})$.

- The functional form, $f(\boldsymbol{x}_i, \boldsymbol{\theta})$, is dictated by theory or experience. In this lecture, we work with the **linear** case:

$$f(\boldsymbol{x}_i, \boldsymbol{\theta}) = \beta_1\, x_{1,i} + \beta_2\, x_{2,i} + \beta_3\, x_{3,i} + \ldots + \beta_k\, x_{k,i}.$$

- Now, we estimate the vector $\boldsymbol{\theta} = \{\beta_1, \beta_2, \ldots, \beta_k\}$ by minimizing

$$S(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{T} \varepsilon_i^2 = \sum_{i=1}^{T} (y_i - \beta_1\, x_{1,i} - \beta_2\, x_{2,i} - \cdots - \beta_k\, x_{k,i})^2$$

In this case, we call this estimator the *Ordinary Least Squares* (OLS) estimator. (Ordinary = Linear functional form.)

Least Squares Estimation: Example

Example: We want to study the effect of a CEO's education (x) on a firm's CEO's compensation (y). We build a CEO's compensation model including a CEO's education (x) and other "*control variables*" (**W**: experience, gender, etc.), controlling for other features that make one CEO's compensation different from another. That is,

$$y_i = f(x_i, \mathbf{W}_i, \boldsymbol{\theta}) + \varepsilon_i, \qquad i = 1, 2, ..., T.$$

The term ε_i represents the effects of individual variation that have not been controlled for with $\boldsymbol{W}_i$ or x_i and $\boldsymbol{\theta}$ is a vector of parameters.

Usually, $f(x, \theta)$ is linear. Then, the compensation model becomes:

$$y_i = \alpha + \beta\ x_i + \gamma_1 W_{1,i} + \gamma_2 W_{2,i} + ... + \varepsilon_i$$

We are interested in estimating β, our parameter of interest, which measures the effect of a CEO's education on a CEO's compensation.

Least Squares Estimation: Linear Algebra

• We will use linear algebra notation. That is,

$$\boldsymbol{y} = f(\mathbf{X}, \theta) + \boldsymbol{\varepsilon}$$

Vectors will be column vectors: $\boldsymbol{y}$, $\boldsymbol{x}_k$, and $\boldsymbol{\varepsilon}$ are Tx1 vectors:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{y}' = [y_1\ y_2\ ....\ y_T]$$

$$\boldsymbol{x}_k = \begin{bmatrix} x_{k1} \\ \vdots \\ x_{kT} \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{x}_k' = [x_{k1}\ x_{k2}\ ....\ x_{kT}]$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{\varepsilon}' = [\varepsilon_1\ \varepsilon_2\ ....\ \varepsilon_T]$$

$\mathbf{X}$ is a Txk matrix. $\quad \Rightarrow \quad \mathbf{X} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ ....\ \boldsymbol{x}_k]$

Not to be posted/distributed online without written consent from author.

Least Squares Estimation: Linear Algebra

X is a Txk matrix. Its columns are the k Tx1 vectors $\boldsymbol{x}_k$. It is common to treat $\boldsymbol{x}_1$ as vector of ones:

$$\boldsymbol{x}_1 = \begin{bmatrix} x_{11} \\ \vdots \\ x_{1T} \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{x}_1\text{'}= [1\ 1\ ....\ 1] = \boldsymbol{\imath}\text{'}$$

Note: Pre-multiplying a vector (1xT) by $\boldsymbol{\imath}$ (or $\boldsymbol{\imath}$' $\boldsymbol{x}_k$) produces a scalar:

$$\boldsymbol{x}_k\text{'}\ \boldsymbol{\imath} = \boldsymbol{\imath}\text{'}\ \boldsymbol{x}_k = x_{k1} + x_{k2} + ... + x_{kT} = \textstyle\sum_j x_{kj}$$

Least Squares Estimation: Assumptions

• Typical Assumptions

(**A1**) DGP: $\boldsymbol{y} = f(\mathbf{X}, \theta) + \boldsymbol{\varepsilon}$ is correctly specified.

For example, $f(\mathrm{x}, \theta) = \mathbf{X}\ \boldsymbol{\beta}$

(**A2**) E[$\boldsymbol{\varepsilon}$|**X**] = 0

(**A3**) Var[$\boldsymbol{\varepsilon}$|**X**] = $\sigma^2\ \mathbf{I}_T$

(**A4**) **X** has full column rank –rank(**X**) = k–, where T ≥ k.

• Assumption (**A1**) is called *correct specification*. We know how the data is generated. We call $\boldsymbol{y} = f(\mathbf{X}, \theta) + \boldsymbol{\varepsilon}$ the *Data Generating Process* (DGP).

Note: The errors, $\boldsymbol{\varepsilon}$, are called *disturbances*. They are not something we add to $f(\mathbf{X}, \theta)$ because we don't know precisely $f(\mathbf{X}, \theta)$. No. The errors are part of the DGP.

Not to be posted/distributed online without written consent from author.

Least Squares Estimation: Assumptions

• Assumption (**A2**) is called *regression.*

From Assumption (**A2**) we get:

(i) $E[\boldsymbol{\varepsilon}|\mathbf{X}] = 0 \quad \Rightarrow E[\boldsymbol{y}|\mathbf{X}] = f(\mathbf{X}, \theta) + E[\boldsymbol{\varepsilon}|\mathbf{X}] = f(\mathbf{X}, \theta)$

That is, the observed $\boldsymbol{y}$ will equal $E[\boldsymbol{y}|\mathbf{X}]$ + random variation.

(ii) Using the Law of Iterated Expectations (LIE):

$$E[\boldsymbol{\varepsilon}] = E_{\mathbf{X}}[E[\boldsymbol{\varepsilon}|\mathbf{X}]] = E_{\mathbf{X}}[0] = 0$$

(iii) There is no information about $\boldsymbol{\varepsilon}$ in $\mathbf{X}$ $\Rightarrow \text{Cov}(\boldsymbol{\varepsilon}, \mathbf{X})=0$.

$$\text{Cov}(\boldsymbol{\varepsilon}, \mathbf{X}) = E[(\boldsymbol{\varepsilon} - 0)(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})] = E[\boldsymbol{\varepsilon}\mathbf{X}]$$

$$\Rightarrow E[\boldsymbol{\varepsilon}\mathbf{X}] = E_{\mathbf{X}}[E[\boldsymbol{\varepsilon}\mathbf{X}|\mathbf{X}]] = E_{\mathbf{X}}[\mathbf{X}\, E[\boldsymbol{\varepsilon}|\mathbf{X}]] = 0 \quad \text{(using LIE)}$$

$$\Rightarrow \text{That is,} \quad E[\boldsymbol{\varepsilon}\mathbf{X}] = 0 \quad \Rightarrow \boldsymbol{\varepsilon} \perp \mathbf{X}.$$

Least Squares Estimation: Assumptions

• From Assumption (**A3**)

$$\text{Var}[\boldsymbol{\varepsilon}|\mathbf{X}] = \boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\, \mathbf{I}_T$$

From (**A3**) we get

$$\text{Var}[\boldsymbol{\varepsilon}|\mathbf{X}] = \sigma^2 \mathbf{I}_T \quad \Rightarrow \text{Var}[\boldsymbol{\varepsilon}] = \sigma^2\, \mathbf{I}_T$$

Proof: $\text{Var}[\boldsymbol{\varepsilon}] = E_x[\text{Var}[\boldsymbol{\varepsilon}|\mathbf{X}]] + \text{Var}_x[E[\boldsymbol{\varepsilon}|\mathbf{X}]] = \sigma^2\, \mathbf{I}_T$. ▪

This assumption implies

(i) *homoscedasticity* $\quad \Rightarrow E[\varepsilon_i^2|\mathbf{X}] = \sigma^2 \quad$ for all i.

(ii) *no serial/cross correlation* $\quad \Rightarrow E[\varepsilon_i\, \varepsilon_j|\mathbf{X}] = 0 \quad$ for $i \neq j$.

Least Squares Estimation: Assumptions

• From Assumption (**A4**) ⇒ the k independent variables in **X** are linearly independent. Then, the kxk matrix **X'X** will also have full rank –i.e., rank(**X'X**) = k.

Thus, **X'X** is invertible. We will need this result to solve a system of equations given by the 1st-order conditions of Least Squares Estimation.

<u>Note</u>: To get asymptotic results we will need more assumptions about **X**.

Least Squares Estimation: F.o.c.

• General functional form:

$f(x_i, \boldsymbol{\theta})$ -$\boldsymbol{\theta}$ is a vector of k parameters.

• Model:

$$y_i = f(x_i, \boldsymbol{\theta}) + \varepsilon_i$$

• Objective function:

$$S(\boldsymbol{x};\, \theta) = \sum_{i=1}^{T} \varepsilon_i^2 = \sum_{i=1}^{T} \{y_i - f(x_i, \boldsymbol{\theta})\}^2$$
$$= \{y_1 - f(x_1, \boldsymbol{\theta})\}^2 + \{y_2 - f(x_2, \boldsymbol{\theta})\}^2 + \ldots + \{y_T - f(x_T, \boldsymbol{\theta})\}^2$$

• We minimize S($\boldsymbol{x}$, $\boldsymbol{\theta}$) with respect to $\boldsymbol{\theta}$:

$$\frac{\partial S(\boldsymbol{x}, \theta)}{\partial \theta} = 2\,\{y_1 - f(x_1, \boldsymbol{\theta})\}(-f'(x_1, \boldsymbol{\theta})) + \cdots + 2\,\{y_T - f(x_T, \boldsymbol{\theta})\}(-f'(x_T, \boldsymbol{\theta}))$$
$$= -2 \sum_i^T \{y_i - f(x_i, \boldsymbol{\theta})\}\, f'(x_i, \boldsymbol{\theta})\}$$

Least Squares Estimation: F.o.c.

• We minimize S($\boldsymbol{x}$, $\boldsymbol{\theta}$) with respect to $\boldsymbol{\theta}$:

$$\frac{\partial S(\boldsymbol{x},\theta)}{\partial\theta} = -2\sum_i^T \{y_i - f(x_i,\boldsymbol{\theta})\}\, f'(x_i,\boldsymbol{\theta})\}$$

• We set the f.o.c.'s:

$$-2\sum_i^T \{y_i - f(x_i,\hat{\boldsymbol{\theta}}_{LS})\}\, f'(x_i,\hat{\boldsymbol{\theta}}_{LS})\} = 0$$

$$\sum_i^T \{y_i - f(x_i,\hat{\boldsymbol{\theta}}_{LS})\}\, f'(x_i,\hat{\boldsymbol{\theta}}_{LS})\} = 0 \qquad \textit{(normal equations)}$$

• The *normal equations* (a kxk system) do not always have an analytic solution. When $f(x_i,\boldsymbol{\theta})$ is linear, we get an explicit solution, $\hat{\boldsymbol{\theta}}_{OLS} = \mathbf{b}$.

• When $f(x_i,\boldsymbol{\theta})$ is *non-linear*, we **do not have** an explicit solution for $\hat{\boldsymbol{\theta}}_{LS}$. The system can be solved numerically. In this case, the estimator is usually referred as *Non-linear Least Squares* estimator, $\hat{\boldsymbol{\theta}}_{NLLS}$.

CLM – OLS: Assumptions and Setup

• Suppose we assume a linear functional form for $f(x, \theta)$:

(**A1'**) DGP: $\boldsymbol{y} = f(\mathbf{X}, \theta) + \boldsymbol{\varepsilon} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Now, we have all the assumptions behind *classical linear regression model* (CLM):

(**A1**) DGP: $\boldsymbol{y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is correctly specified.

(**A2**) $E[\boldsymbol{\varepsilon}|\mathbf{X}] = 0$

(**A3**) $Var[\boldsymbol{\varepsilon}|\mathbf{X}] = \sigma^2\, \mathbf{I}_T$

(**A4**) $\mathbf{X}$ has full column rank –rank($\mathbf{X}$) = k, where T ≥ k.

Objective function: $S(\boldsymbol{x};\theta) = \sum_{i=1}^T \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'\,(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$

CLM – OLS: Rules for Vector Derivatives

• Recall the rules for vector differentiation of linear functions and quadratic forms:

(1) Linear function: $\boldsymbol{y} = f(\boldsymbol{x}) = \boldsymbol{x}'\,\beta + \omega$

where $\boldsymbol{x}$ and $\boldsymbol{\beta}$ are k-dimensional vectors and ω is a constant. Then,

$$\nabla f(\boldsymbol{x}) = \beta$$

(2) Quadratic form: $\mathrm{q} = f(\boldsymbol{x}) = \boldsymbol{x}'\,\mathbf{A}\,\boldsymbol{x}$

where $\boldsymbol{x}$ is kx1 vector and $\mathbf{A}$ is a kxk matrix, with a_{ji} elements. Then,

$$\nabla f(\boldsymbol{x}) = \mathbf{A}'\,\boldsymbol{x} + \mathbf{A}\,\boldsymbol{x} = (\mathbf{A}' + \mathbf{A})\,\boldsymbol{x}$$

If $\mathbf{A}$ is symmetric, then $\nabla f(\boldsymbol{x}) = 2\,\mathbf{A}\,\boldsymbol{x}$

Now, we apply them to $\mathrm{S}(\boldsymbol{x};\,\theta) = \sum_{i=1}^{T} \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'\,(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$

$$= (\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{\beta}'\mathrm{X}'\boldsymbol{y} - \boldsymbol{y}'\mathrm{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathrm{X}'\mathrm{X}\boldsymbol{\beta})$$

CLM – OLS: Derivation

• Objective function:

$$\mathrm{S}(\boldsymbol{x};\,\theta) = (\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{\beta}'\mathrm{X}'\boldsymbol{y} - \boldsymbol{y}'\mathrm{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathrm{X}'\mathrm{X}\boldsymbol{\beta})$$
$$= (\mathbf{c} - \beta'\mathbf{d} - \mathbf{d}'\beta + \beta'\mathbf{A}\beta)$$
$$= (\mathbf{c} - 2\,\mathbf{d}'\beta + \beta'\mathbf{A}\beta)$$

• First derivative w.r.t. β: $\nabla\,\mathrm{S}(\boldsymbol{x};\,\theta) = (-2\,\mathbf{d} + 2\,\mathbf{A}\,\boldsymbol{\beta})$ (kx1 vector)

• F.o.c. (*normal equations*): $-2\,(\mathbf{X}'\boldsymbol{y} - \mathbf{X}'\mathbf{X}\,\mathbf{b}) = \mathbf{0}$

$$\Rightarrow (\mathbf{X}'\mathbf{X})\,\mathbf{b} = \mathbf{X}'\boldsymbol{y}$$

• Assuming $(\mathbf{X}'\mathbf{X})$ is non-singular –i.e., invertible-, we solve for **b**:

$$\Rightarrow \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\boldsymbol{y} \qquad \text{(a } k\text{x1 vector)}$$

<u>Note</u>: **b** is called the Ordinary Least Squares (OLS) estimator.
(*Ordinary* = $f(\mathbf{X},\,\theta)$ is linear)

CLM – OLS

• **Example**: One explanatory variable model.

(**A1'**) DGP: $\boldsymbol{y} = \beta_1 + \beta_2 \boldsymbol{x} + \varepsilon$

Objective function: $S(x_i, \theta) = \sum_{i=1}^{T} \varepsilon_i^2 = \sum_i (y_i - \beta_1 - \beta_2 x_i)^2$

F.o.c. (2 equations, 2 unknowns):

(β_1): $-2\, \Sigma_i\, (y_i - b_1 - b_2 x_i)\, (-1) = 0 \quad \Rightarrow \Sigma_i\, (y_i - b_1 - b_2 x_i) = 0$ (1)

(β_2): $-2\, \Sigma_i\, (y_i - b_1 - b_2 x_i)\, (-x_i) = 0 \quad \Rightarrow \Sigma_i\, (y_i x_i - b_1 x_i - b_2 x_i^2) = 0$ (2)

From (1): $\sum_i y_i - \Sigma_i\, b_1 - b_2 \sum_i x_i = 0 \qquad \Rightarrow b_1 = \bar{y} - b_2 \bar{x}$

From (2): $\Sigma_i\, y_i\, x_i - (\bar{y} - b_2 \bar{x})\, \Sigma_i\, x_i - b_2\, \Sigma_i\, x_i^2 = 0 \Rightarrow b_2 = \frac{\sum_i (y_i - \bar{y}) x_i}{\sum_i (x_i - \bar{x}) x_i}$

or, more elegantly, $\quad b_2 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{cov(y_i, x_i)}{var(x_i)}$

OLS Estimation: Second Order Condition

• OLS estimator: $\quad \mathbf{b} = (\mathbf{X'X})^{-1}\, \mathbf{X'}\boldsymbol{y}$

<u>Note</u>: (i) $\mathbf{b} = \beta_{OLS}$. (Ordinary LS. *Ordinary* = linear)

(ii) $\mathbf{b}$ is a (linear) function of the data (y_i, x_i).

(iii) $\mathbf{X'}(\boldsymbol{y} - \mathbf{Xb}) = \mathbf{X'}\boldsymbol{y} - \mathbf{X'X}(\mathbf{X'X})^{-1}\mathbf{X'}\boldsymbol{y} = \mathbf{X'e} = \mathbf{0} \quad \Rightarrow \mathbf{e} \perp \mathbf{X}.$

• Q: Is $\mathbf{b}$ is a minimum? We need to check the s.o.c.

$$\frac{\partial (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})}{\partial \mathbf{b}} = -2\mathbf{X}'(\mathbf{y} - \mathbf{Xb})$$

$$\frac{\partial^2 (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})}{\partial \mathbf{b}\, \partial \mathbf{b}'} = \frac{\partial \left(\frac{\partial (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})}{\partial \mathbf{b}} \right)}{\partial \mathbf{b}'}$$

$$= \frac{\partial \text{ column vector}}{\partial \text{ row vector}}$$

$$= 2\mathbf{X}'\mathbf{X}$$

OLS Estimation: Second Order Condition

$$\frac{\partial^2 \mathrm{e}'\mathrm{e}}{\partial \mathrm{b} \partial \mathrm{b}'} = 2\boldsymbol{X}'\boldsymbol{X} = 2\begin{bmatrix} \sum_{i=1}^{T} x_{i1}^2 & \sum_{i=1}^{T} x_{i1}x_{i2} & \dots & \sum_{i=1}^{T} x_{i1}x_{iK} \\ \sum_{i=1}^{T} x_{i2}x_{i1} & \sum_{i=1}^{T} x_{i2}^2 & \dots & \sum_{i=1}^{T} x_{i2}x_{iK} \\ \dots & \dots & \dots & \dots \\ \sum_{i=1}^{T} x_{iK}x_{i1} & \sum_{i=1}^{T} x_{iK}x_{i2} & \dots & \sum_{i=1}^{T} x_{iK}^2 \end{bmatrix}$$

If there were a single **b**, we would require this to be positive, which it would be: 2 $\boldsymbol{x}'\boldsymbol{x} = 2\sum_{i=1}^{T} x_i^2 > 0.$

The matrix counterpart of a positive number is a positive definite (pd) matrix.

A square matrix (mxm) **A** "takes the sign" of the *quadratic form,* **z'A z**, where **z** is an mx1 vector. Then, **z'A z** is a scalar.

OLS Estimation: Second Order Condition

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} \sum_{i=1}^{T} x_{i1}^2 & \sum_{i=1}^{T} x_{i1}x_{i2} & \dots & \sum_{i=1}^{T} x_{i1}x_{iK} \\ \sum_{i=1}^{T} x_{i2}x_{i1} & \sum_{i=1}^{T} x_{i2}^2 & \dots & \sum_{i=1}^{T} x_{i2}x_{iK} \\ \dots & \dots & \dots & \dots \\ \sum_{i=1}^{T} x_{iK}x_{i1} & \sum_{i=1}^{T} x_{iK}x_{i2} & \dots & \sum_{i=1}^{T} x_{iK}^2 \end{bmatrix}$$

$$= \sum_{i=1}^{T} \begin{bmatrix} x_{i1}^2 & x_{i1}x_{i2} & \dots & x_{i1}x_{iK} \\ x_{i2}x_{i1} & x_{i2}^2 & \dots & x_{i2}x_{iK} \\ \dots & \dots & \dots & \dots \\ x_{iK}x_{i1} & x_{iK}x_{i2} & \dots & x_{iK}^2 \end{bmatrix}$$

<u>Definition</u>: A matrix **A** is *positive definite* (pd) if **z'A z** >0 for any **z**.

For some matrices, it is easy to check. Let **A** = **X'X** (a kxk matrix).

Then, **z'A z** = **z'X'X z** = $\boldsymbol{v}'\boldsymbol{v} = \sum_{i=1}^{T} v_i^2 > 0$. ($\boldsymbol{v}$=**Xz** is Tx1)

⇒ **X'X** is pd ⇒ **b** is a min!

OLS Estimation: Second Order Condition

• A typical pd matrix has positive diagonal positive elements and the off-diagonal elements are not too large in absolute value relative to the diagonal elements. Keep in mind for later, that the diagonal elements are positive.

• If **A** is pd, then $\mathbf{A}^{-1}$ is also pd. Thus, $(\mathbf{X'X})^{-1}$ is also pd.

• In multivariate calculus, the 2nd order condition requires the evaluation of the matrix of second derivatives, the Hessian. If all the leading principal minors are positive, then the critical point obtained is a minimum. In our case, this means that the Hessian is pd.

Note: In general, we need eigenvalues of **A** to check this. If all the eigenvalues are positive, then **A** is pd.

OLS Estimation – Properties

• The LS estimator of β_{LS} when $f(x, \theta) = \mathbf{X}\,\boldsymbol{\beta}$ is linear is

$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'}\,\boldsymbol{y}$ $\quad\Rightarrow$ **b** is a (linear) function of the data (y_i, x_i).

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'}\,\boldsymbol{y} = (\mathbf{X'X})^{-1}\,\mathbf{X'}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X'X})^{-1}\mathbf{X'}\boldsymbol{\varepsilon}$$

Under the typical assumptions, we can establish properties for **b**.

1) Expected value

$$E[\mathbf{b}|\mathbf{X}] = E[\boldsymbol{\beta}|\mathbf{X}] + E[(\mathbf{X'X})^{-1}\mathbf{X'}\boldsymbol{\varepsilon}\,|\mathbf{X}] = \boldsymbol{\beta}$$

That is, **b** is *unbiased* (on average, we get the population parameter).
Recall, bias of an estimator, $\hat{\theta}$, is defined as: $Bias(\hat{\theta}, \theta) = E[\hat{\theta}] - \theta$

2) Variance

$$\text{Var}[\mathbf{b}|\mathbf{X}] = E[(\mathbf{b} - \boldsymbol{\beta})\,(\mathbf{b} - \boldsymbol{\beta})'|\mathbf{X}] = (\mathbf{X'X})^{-1}\,\mathbf{X}'E[\boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon}'|\mathbf{X}]\,\mathbf{X}(\mathbf{X'X})^{-1}$$
$$= \sigma^2\,(\mathbf{X'X})^{-1}$$

OLS Estimation – Properties

3) BLUE (*Best Linear Unbiased Estimator*, or MVLUE).

Theorem: $\mathbf{b}$ is BLUE (*Best Linear Unbiased Estimator*, or MVLUE).

Proof**:**

Let $\mathbf{b}^* = \mathbf{C}\,\boldsymbol{y}$ (linear in $\boldsymbol{y}$)

$E[\mathbf{b}^*|\mathbf{X}] = E[\mathbf{Cy}|\mathbf{X}] = E[\mathbf{C}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})|\mathbf{X}] = \boldsymbol{\beta}$ (unbiased if $\mathbf{CX}=\mathbf{I}$)

$\text{Var}[\mathbf{b}^*|\mathbf{X}] = E[(\mathbf{b}^* - \boldsymbol{\beta})(\mathbf{b}^* - \boldsymbol{\beta})'|\mathbf{X}] = E[\mathbf{C}\boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon}'\mathbf{C}'|\mathbf{X}] = \sigma^2\,\mathbf{CC}'$

Now, let $\mathbf{D} = \mathbf{C} - (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'$ (note $\mathbf{DX}=0$ & $\mathbf{D}'\mathbf{D}$ a pd matrix)

Then, $\text{Var}[\mathbf{b}^*|\mathbf{X}] = \sigma^2\,(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\,(\mathbf{D}' + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1})$

$= \sigma^2\,\mathbf{DD}' + \sigma^2\,(\mathbf{X}'\mathbf{X})^{-1} = \text{Var}[\mathbf{b}|\mathbf{X}] + \sigma^2\,\mathbf{DD}'$. ▪

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

OLS Estimation – Properties

4) Normal Distribution (under additional assumptions for $\boldsymbol{\varepsilon}$)

If we make an additional assumption:

(**A5**) $\boldsymbol{\varepsilon}|\mathbf{X} \sim iid\, N(\mathbf{0}, \sigma^2\mathbf{I}_T)$

we can derive the distribution of $\mathbf{b}$.

Since $\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$, we have that $\mathbf{b}$ is a linear combination of normal variables

$\Rightarrow \mathbf{b}|\mathbf{X} \sim iid\, N(\boldsymbol{\beta}, \sigma^2\,(\mathbf{X}'\,\mathbf{X})^{-1})$

Note: From **(1)** & **(2)**, we compute the MSE (Mean square error)

$\text{MSE}[\mathbf{b}|\mathbf{X}] = E[\|(\mathbf{b} - \boldsymbol{\beta})\|^2] = E[(\mathbf{b} - \boldsymbol{\beta})'\,(\mathbf{b} - \boldsymbol{\beta})]$

After some algebra, we get:

$\Rightarrow \text{MSE}[\mathbf{b}|\mathbf{X}] = \text{tr(Variance)} + \text{squared bias} = \text{tr}[\sigma^2\,(\mathbf{X}'\mathbf{X})^{-1}]$

Not to be posted/distributed online without written consent from author.

OLS Estimation – MSE

• For a scalar estimator, $\hat{\theta}$, the MSE (Mean square error) is:

$$\begin{aligned}\text{MSE}[\hat{\theta}\,|\,\mathbf{X}] &= \text{E}[(\hat{\theta} - \theta)^2] = \text{E}[\{(\hat{\theta} - \text{E}[\hat{\theta}]) + (\text{E}[\hat{\theta}] - \theta)\}^2] \\ &= \text{E}[(\hat{\theta} - \text{E}[\hat{\theta}])^2 + (\text{E}[\hat{\theta}] - \theta)^2 + 2 * 0 \\ &= \text{Var}[\hat{\theta}] + (\textit{Bias}(\hat{\theta}, \theta))^2\end{aligned}$$

Note: The derivation can be done using $\text{Var}[Z] = \text{E}[Z^2] - (\text{E}[Z])^2$, which generalizes to vectors $\text{Var}[\boldsymbol{Z}] = \text{E}[\boldsymbol{Z}\,\boldsymbol{Z}'] - \text{E}[\boldsymbol{Z}]\,\text{E}[\boldsymbol{Z}]'$

• For a vector of estimators, $\hat{\boldsymbol{\theta}}$, we compute the MSE as:

$$\text{MSE}[\hat{\boldsymbol{\theta}}] = \text{E}[\|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\|^2] = \text{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})]$$

• Now, we compute the MSE of OLS **b** as:

$$\text{MSE}[\mathbf{b}\,|\,\mathbf{X}] = \text{E}[\|(\mathbf{b} - \boldsymbol{\beta})\|^2\,|\,\mathbf{X}] = \text{E}[(\mathbf{b} - \boldsymbol{\beta})'\,(\mathbf{b} - \boldsymbol{\beta})\,|\,\mathbf{X}]$$

OLS Estimation – MSE

• The MSE of OLS **b** is:

$$\text{MSE}[\mathbf{b}\,|\,\mathbf{X}] = \text{E}[\|(\mathbf{b} - \boldsymbol{\beta})\|^2\,|\,\mathbf{X}] = \text{E}[(\mathbf{b} - \boldsymbol{\beta})'\,(\mathbf{b} - \boldsymbol{\beta})\,|\,\mathbf{X}]$$

From Properties **(1)** & **(2)**, we can derive:

$\Rightarrow$ $\text{MSE}[\mathbf{b}\,|\,\mathbf{X}] = \text{tr}(\text{Var}[\mathbf{b}\,|\,\mathbf{X}]) + \text{squared bias} = \text{tr}[\sigma^2\,(\mathbf{X}'\mathbf{X})^{-1}]$

Note: In the derivation, we used the following result:

$$\text{E}[\boldsymbol{Z}'\,\mathbf{A}\,\boldsymbol{Z}] = \text{tr}(\mathbf{A}\,\text{Var}[\boldsymbol{Z}]) + \text{E}[\boldsymbol{Z}]'\,\mathbf{A}\,\text{E}[\boldsymbol{Z}]$$

where $\boldsymbol{Z}$, is a random vector and **A** a comformable non-random matrix

OLS Estimation – Variance

Example: One explanatory variable model.

(**A1**') DGP: $\boldsymbol{y} = \beta_1 + \beta_2\, \mathbf{x} + \boldsymbol{\varepsilon}$

$$\text{Var}[\mathbf{b}|\mathbf{X}] = \sigma^2\,(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \begin{bmatrix} \sum_i 1 & \sum_i 1x_i \\ \sum_i 1x_i & \sum_i x_i^2 \end{bmatrix}^{-1} = \sigma^2 \begin{bmatrix} T & T\bar{x} \\ T\bar{x} & \sum_i x_i^2 \end{bmatrix}^{-1}$$

$$\text{Var}[b_1|\mathbf{X}] = \sigma^2 \frac{\sum_i x_i^2}{T(\sum_i x_i^2 - T\bar{x}^2)} = \sigma^2 \frac{\sum_i x_i^2/T}{\sum_i (x_i - \bar{x})^2}$$

$$\text{Var}[b_2|\mathbf{X}] = \sigma^2 \frac{1}{(\sum_i x_i^2 - T\bar{x}^2)} = \sigma^2 \frac{1}{\sum_i (x_i - \bar{x})^2}$$

Algebraic Results

• Important Matrices

(1) "Residual maker" $\quad \mathbf{M} = \mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

$\mathbf{M}\boldsymbol{y} = \boldsymbol{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y} = \boldsymbol{y} - \mathbf{X}\mathbf{b} = \mathbf{e}$ (residuals)

$\mathbf{M}\mathbf{X} = \mathbf{0}$

- **M** is symmetric – $\mathbf{M} = \mathbf{M}'$
- **M** is idempotent – $\mathbf{M}^*\mathbf{M} = \mathbf{M}$
- **M** is singular – $\mathbf{M}^{-1}$ does not exist. $\Rightarrow$ rank($\mathbf{M}$) = $T - k$

(**M** does not have full rank. We have already proven this result.)

- Special case: $\mathbf{X} = \boldsymbol{\imath}$

$\mathbf{M}^0 = \mathbf{I} - \boldsymbol{\imath}(\boldsymbol{\imath}'\,\boldsymbol{\imath})^{-1}\boldsymbol{\imath}' = \mathbf{I} - \boldsymbol{\imath}\,\boldsymbol{\imath}'/T$

$\mathbf{M}^0\,\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{\imath}(\boldsymbol{\imath}'\,\boldsymbol{\imath})^{-1}\boldsymbol{\imath}'\,\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{\imath}\,\bar{\boldsymbol{y}}$

$\mathbf{M}^0$ = de-meaning matrix.

Algebraic Results

• Important Matrices

(2) "Projection matrix" $\mathbf{P} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$

$\mathbf{P}\boldsymbol{y} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}\,\boldsymbol{y} = \mathbf{Xb} = \hat{\boldsymbol{y}}$ (fitted values)

$\mathbf{P}\boldsymbol{y}$ is the projection of $\mathbf{y}$ into the column space of $\mathbf{X}$.

$\mathbf{PM} = \mathbf{MP} = \mathbf{0}$ (Projection matrix)

$\mathbf{PX} = \mathbf{X}$ s^2

- $\mathbf{P}$ is symmetric – $\mathbf{P} = \mathbf{P'}$
- $\mathbf{P}$ is idempotent – $\mathbf{P}^*\mathbf{P} = \mathbf{P}$
- $\mathbf{P}$ is singular – $\mathbf{P}^{-1}$ does not exist. $\Rightarrow$ rank($\mathbf{P}$) = k

Algebraic Results

• Disturbances and Residuals

In the population: $E[\mathbf{X'}\,\boldsymbol{\varepsilon}] = 0$.

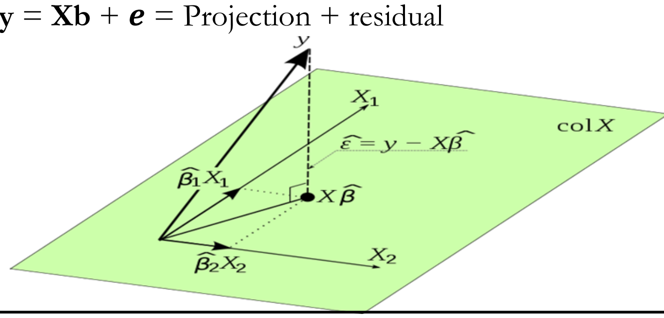
In the sample: $\mathbf{X'}\,\boldsymbol{e} = \mathbf{X'}(\mathbf{y} - \mathbf{Xb}) = \mathbf{X'y} - \mathbf{X'X}(\mathbf{X'X})^{-1}\mathbf{X'y}$

$= 1/T(\mathbf{X'}\,\boldsymbol{e}) = 0$.

• We have two ways to look at $\mathbf{y}$:

$\mathbf{y} = E[\mathbf{y}|\mathbf{X}] + \boldsymbol{\varepsilon}$ = Conditional mean + disturbance

$\mathbf{y} = \mathbf{Xb} + \boldsymbol{e}$ = Projection + residual



Results when X Contains a Constant Term

- Let the first column of **X** be a column of ones. That is

$$\mathbf{X} = [\boldsymbol{i}\ \boldsymbol{x}_2\ ....\ \boldsymbol{x}_k]$$

]

- Then,

(1) Since $\mathbf{X}'\,\boldsymbol{e} = \mathbf{0}$ $\quad\Rightarrow \mathbf{x}_1'\boldsymbol{e} = 0$ –the residuals sum to zero.

(2) Since $\boldsymbol{y} = \mathbf{Xb} + \boldsymbol{e}$ $\quad\Rightarrow \boldsymbol{i}'\boldsymbol{y} = \boldsymbol{i}'\mathbf{Xb} + \boldsymbol{i}'\,\boldsymbol{e} = \boldsymbol{i}'\mathbf{Xb}$

$\Rightarrow \overline{\boldsymbol{y}} = \overline{\boldsymbol{x}}\ \mathbf{b}$

That is, the regression line passes through the means.

Note: These results are only true if **X** contains a constant term!

OLS Estimation – Example in R

Example: 3 Factor Fama-French Model:

```
Returns <- read.csv("http://www.bauer.uh.edu/rsusmel/phd/K-DIS-IBM.csv", head=TRUE, sep=",")

y1 <- Returns$IBM; rf <- Returns$Rf; y <- y1 - rf
x1 <- Returns$Rm_Rf; x2 <- Returns$SMB; x3 <- Returns$HML
T <- length(x1)
x0 <- matrix(1,T,1)
x <- cbind(x0,x1,x2,x3)
k <- ncol(x)

b <- solve(t(x)%*% x)%*% t(x)%*%y          # b = (X'X)^-1 X' y  (OLS regression)
e <- y - x%*%b                             # regression residuals, e
RSS <- as.numeric(t(e)%*%e)                # RSS
Sigma2 <- as.numeric(RSS/(T-k))            # Estimated σ^2 = s^2 (See Chapter 2)
Var_b <- Sigma2*solve(t(x)%*% x)           # Estimated Var[b|X] = s^2 (X'X)^-1
SE_b <- sqrt(diag(Var_b))                  # SE[b|X]
```

Not to be posted/distributed online without written consent from author.

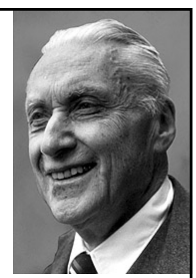
OLS Estimation – Example in R

```
> RSS
[1] 12.92964
> Sigma2
[1] 0.03894471
> t(b)
                  x1        x2         x3
[1,] -0.2258839 1.061934 0.1343667 -0.3574959
> SE_b
                x1          x2         x3
0.01095196 0.26363344 0.35518792 0.37631714
```

Note: You should get the same numbers using R's linear model command, *lm* (use summary(.) to print results):

```
fit <- lm(y ~ x -1)
summary(fit)
```

Frisch-Waugh (1933) Theorem



Ragnar Frisch (1895 – 1973)

• Context: Model contains two sets of variables:

$\mathbf{X}$ = [[1,*time*] | [*other variables*]]

= $[\mathbf{X}_1 \;\; \mathbf{X}_2]$

• Regression model:

$\boldsymbol{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$ (population)

$= \mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_2\mathbf{b}_2 + \boldsymbol{e}$ (sample)

• OLS solution: $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X}'\boldsymbol{y} = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix}$

Problem in 1933: Can we estimate $\boldsymbol{\beta}_2$ without inverting the (k_1+k_2) x (k_1+k_2) **X'X** matrix? The F-W theorem helps reduce computation, by getting simplified algebraic expression for OLS coefficient, $\mathbf{b}_2$.

F-W: Partitioned Solution

• We manipulate the normal equation, $(\boldsymbol{y} - \mathbf{Xb})'\,\mathbf{X} = 0$:

$$\begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{y} \\ \boldsymbol{X}_2'\boldsymbol{y} \end{bmatrix}$$

Then, focusing on the last row, we get

$$\boldsymbol{X}_2'\boldsymbol{X}_1\mathbf{b}_1 + \boldsymbol{X}_2'\boldsymbol{X}_2\mathbf{b}_2 = \boldsymbol{X}_2'\boldsymbol{y}$$

$$\Rightarrow \quad \mathbf{b}_2 = (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\,\boldsymbol{X}_2'(\boldsymbol{y} - \boldsymbol{X}_1\mathbf{b}_1)$$

Then, $\mathbf{b}_2$ is estimated with a regression of $(\boldsymbol{y} - \boldsymbol{X}_1\mathbf{b}_1)$ on $\boldsymbol{X}_2$

If $\boldsymbol{X}_2'\boldsymbol{X}_1 = \mathbf{0}$ ($\boldsymbol{X}_2$ & $\boldsymbol{X}_1$ are orthogonal)

$$\Rightarrow \mathbf{b}_2 = (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\,\boldsymbol{X}_2'\boldsymbol{y} \quad \text{(a regression of } \boldsymbol{y} \text{ on } \boldsymbol{X}_2\text{).}$$

F-W: Partitioned Solution

• Back to the estimation of $\mathbf{b}_2$, without inverting $(\mathbf{X'X})$. We start with

$$\begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{y} \\ \boldsymbol{X}_2'\boldsymbol{y} \end{bmatrix}$$

• To get $\mathbf{b}_2$, we use the partitioned inverse

$$\begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{bmatrix}^{-1} \qquad [\,]^{-1}_{(2,2)}$$

• With the partitioned inverse, we get:

$$\mathbf{b}_2 = [\,]^{-1}_{(2,1)}\,\boldsymbol{X}_1'\boldsymbol{y} + [\,]^{-1}_{(2,2)}\,\boldsymbol{X}_2'\boldsymbol{y}$$

We need the partitioned inverse of $(\boldsymbol{X}'\boldsymbol{X})$.

Not to be posted/distributed online without written consent from author.

F-W: Partitioned Solution

• Recall from the Linear Algebra Review:

1. $$\begin{bmatrix}\Sigma_{XX} & \Sigma_{XY} & I & 0\\ \Sigma_{YX} & \Sigma_{YY} & 0 & I\end{bmatrix} \xrightarrow{\Sigma_{XX}^{-1}R_1} \begin{bmatrix} I & \Sigma_{XX}^{-1}\Sigma_{XY} & \Sigma_{XX}^{-1} & 0\\ \Sigma_{YX} & \Sigma_{YY} & 0 & I\end{bmatrix}$$

2. $$\xrightarrow{R_2-\Sigma_{YX}R_1} \begin{bmatrix} I & \Sigma_{XX}^{-1}\Sigma_{XY} & \Sigma_{XX}^{-1} & 0\\ 0 & \Sigma_{YY}-\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} & -\Sigma_{YX}\Sigma_{XX}^{-1} & I\end{bmatrix}$$

3. $$\xrightarrow{[\Sigma_{YY}-\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}]^{-1}R_2} \begin{bmatrix} I & \Sigma_{XX}^{-1}\Sigma_{XY} & \Sigma_{XX}^{-1} & 0\\ 0 & I & D(-\Sigma_{YX}\Sigma_{XX}^{-1}) & D\end{bmatrix}$$

where $D=[\Sigma_{YY}-\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}]^{-1}$

4. $$\xrightarrow{R_1-\Sigma_{XX}^{-1}\Sigma_{XY}R_2} \begin{bmatrix} I & 0 & \Sigma_{XX}^{-1}+\Sigma_{XX}^{-1}\Sigma_{XY}D\Sigma_{YX}\Sigma_{XX}^{-1} & \Sigma_{XX}^{-1}\Sigma_{XY}D\\ 0 & I & -D(\Sigma_{YX}\Sigma_{XX}^{-1}) & D\end{bmatrix}$$

F-W: Partitioned Solution

• Then,

1. Matrix $X'X=\begin{bmatrix}X_1'X_1 & X_1'X_2\\ X_2'X_1 & X_2'X_2\end{bmatrix}$

2. Inverse $=\begin{bmatrix}(X_1'X_1)^{-1}+(X_1'X_1)^{-1}X_1'X_2DX_2'X_1(X_1'X_1)^{-1} & (X_1'X_1)^{-1}X_1'X_2D\\ -DX_2'X_1(X_1'X_1)^{-1} & D\end{bmatrix}$

where $D=[X_2'X_2-X_2'X_1(X_1'X_1)^{-1}X_1'X_2]^{-1}=[X_2'(I-X_1(X_1'X_1)^{-1}X_1)'X_2]^{-1}$

$=> D=[X_2'M_1X_2]^{-1}$

The algebraic result is: $[\,]^{-1}_{(2,1)} = -\mathbf{D}\,\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}$
$[\,]^{-1}_{(2,2)} = \mathbf{D} = [\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2]^{-1}$

• Then, continuing the algebraic manipulation:

$\mathbf{b}_2 = [\,]^{-1}_{(2,1)}\,\mathbf{X}_1'\boldsymbol{y} + [\,]^{-1}_{(2,2)}\,\mathbf{X}_2'\boldsymbol{y} =$
$= -\mathbf{D}\,\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\,\mathbf{X}_1'\boldsymbol{y} + \mathbf{D}\,\mathbf{X}_2'\boldsymbol{y} = [\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2]^{-1}\mathbf{X}_2'\mathbf{M}_1\boldsymbol{y}$

F-W: Partitioned Solution - Results

• Then, continuing the algebraic manipulation:

$$\begin{aligned}\mathbf{b}_2 &= [\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2]^{-1}\mathbf{X}_2'\mathbf{M}_1\boldsymbol{y} \\ &= [\mathbf{X}_2'\,\mathbf{M}_1'\mathbf{M}_1\mathbf{X}_2]^{-1}\mathbf{X}_2'\mathbf{M}_1'\mathbf{M}_1\boldsymbol{y} \\ &= [\mathbf{X}^*_2{}'\mathbf{X}^*_2]^{-1}\mathbf{X}^*_2{}'\boldsymbol{y}^*\end{aligned}$$

where $\mathbf{Z}^* = {}'\mathbf{M}_1\mathbf{Z}$ = residuals from a regression of $\mathbf{Z}$ on $\mathbf{X}_1$.

This is Frisch and Waugh's result - the *double residual regression*. We have a regression of residuals on residuals!

• Back to original context. Two ways to estimate $\mathbf{b}_2$:

(1) *Detrend* the other variables. Use detrended data in the regression.

(2) Use all the original variables, including constant and time trend.

Detrend: Compute the residuals from the regressions of the variables on a constant and a time trend.

Frisch-Waugh Result: Implications

• FW result:

$$\begin{aligned}\mathbf{b}_2 &= [\mathbf{X}^*_2{}'\,\mathbf{X}^*_2]^{-1}\mathbf{X}^*_2{}'\,\boldsymbol{y}^* \\ &= [\mathbf{X}_2'\,\mathbf{M}_1\mathbf{X}_2]^{-1}\mathbf{X}_2'\mathbf{M}_1\boldsymbol{y} = [\mathbf{X}_2'\mathbf{M}_1'\,\mathbf{M}_1\mathbf{X}_2]^{-1}\mathbf{X}_2'\,\mathbf{M}_1'\mathbf{M}_1\boldsymbol{y}\end{aligned}$$

• Implications

- We can isolate a single coefficient in a regression.

- It is not necessary to 'partial' the other $\mathbf{X}$s out of $\boldsymbol{y}$ ($\mathbf{M}_1$ is idempotent)

- Suppose $\mathbf{X}_1 \perp \mathbf{X}_2$ ($\Rightarrow \mathbf{X}_2'\mathbf{M}_1 = \mathbf{X}_2'$). Then, we have the orthogonal regression:

$$\begin{aligned}\mathbf{b}_2 &= (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\,\boldsymbol{y} \\ \mathbf{b}_1 &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\,\boldsymbol{y}\end{aligned}$$

Frisch-Waugh Result: Implications

Example: De-mean

Let $\mathbf{X}_1 = \mathbf{i}$ $\Rightarrow \mathbf{P}_1 = \mathbf{i}\,(\mathbf{i}'\,\mathbf{i})^{-1}\mathbf{i}' = \mathbf{i}\,(T)^{-1}\mathbf{i}' = \mathbf{i}\,\mathbf{i}'/T$

$\Rightarrow \mathbf{M}_1\,\mathbf{z} = \mathbf{z} - \mathbf{i}\,\mathbf{i}'\,\mathbf{z}\,/T = \mathbf{z} - \mathbf{i}\,\bar{z}$ ($\mathbf{M}_1$ demeans $\mathbf{z}$)

$$\mathbf{b}_2 = [\mathbf{X}_2'\mathbf{M}_1'\mathbf{M}_1\mathbf{X}_2]^{-1}\,\mathbf{X}_2'\,\mathbf{M}_1'\,\mathbf{M}_1\,\boldsymbol{y}$$

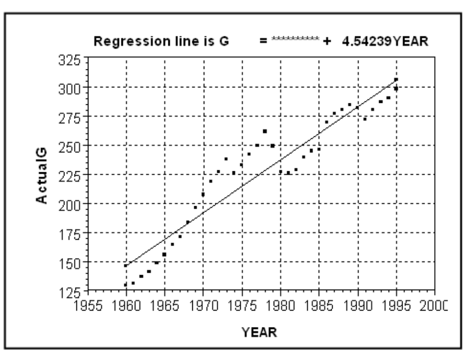
Note: We can do linear regression on data in mean deviation form.

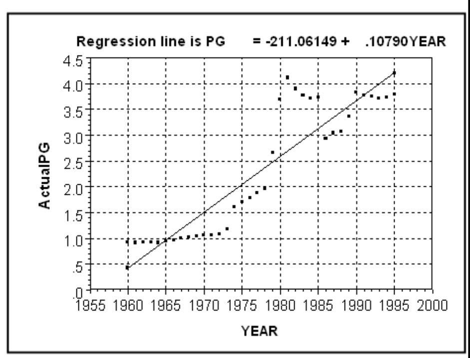
Application: Detrending G and PG

- Example taken from Greene

G: Consumption of Gasoline

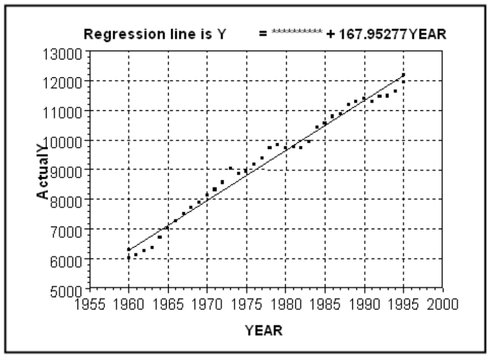
PG: Price of Gasoline





Not to be posted/distributed online without written consent from author.

Application: Detrending Y



Y: Income

Y* = Y – (******* + 167.95277* Year)

Application: Detrended Regression

Regression of detrended Gasoline (M_1G) on detrended Price of Gasoline (M_1PG) detrended Income (M_1Y)

```
namelist;x1=one,year$
regr;lhs=pg;rhs=x1;res=pgstar$
regr;lhs=y ;rhs=x1;res=ystar$
regr;lhs=g ;rhs=x1;res=gstar$
regr;lhs=gstar;rhs=pgstar,ystar$
+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error |t-ratio |P[|T|>t] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
 PGSTAR      -11.98265151       2.1171860   -5.660   .0000  .87954335E-14
 YSTAR     .4781809512E-01  .45261498E-02   10.565   .0000 -.48506384E-11
```

From the Gasoline data in Notes 3.

```
+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error |t-ratio |P[|T|>t] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
 Constant     4154.597719       1748.6561    2.376   .0237
 YEAR        -2.195824001        .90679770   -2.422   .0213      1977.5000
 PG          -11.98265151       2.1823454   -5.491   .0000      2.3166111
 Y         .4781809512E-01  .46654484E-02   10.249   .0000      9232.8611
```

Goodness of Fit of the Regression

• After estimating the model, we would like to judge the adequacy of the model. There are two ways to do this:

- Visual: plots of fitted values and residuals, histograms of residuals.
- Numerical measures: R^2, adjusted R^2, AIC, BIC, etc.

• Numerical measures. In general, they are simple and easy to compute. We call them *goodness-of-fit* measures. Most popular: R^2.

• Definition: Variation

In the context of a model, we consider the *variation* of a variable as the movement of the variable, usually associated with movement of another variable.

Goodness of Fit of the Regression

• Total variation = $\sum_i (y_i - \bar{y})^2$ = $\boldsymbol{y}'\mathbf{M}^0\boldsymbol{y}$.

where $\mathbf{M}^0 = \mathbf{I} - \boldsymbol{i}(\boldsymbol{i}'\,\boldsymbol{i})^{-1}\boldsymbol{i}'$ = the $\mathbf{M}$ de-meaning matrix.

• Decomposition of total variation (assume $\mathbf{X}_1 = \boldsymbol{i}$ –a constant.)

$\boldsymbol{y} = \mathbf{Xb} + \boldsymbol{e}$, so

$\mathbf{M}^0\boldsymbol{y} = \mathbf{M}^0\mathbf{Xb} + \mathbf{M}^0\boldsymbol{e} = \mathbf{M}^0\mathbf{Xb} + \boldsymbol{e}$ (deviations from means)

$\boldsymbol{y}'\mathbf{M}^0\boldsymbol{y} = \mathbf{b}'(\mathbf{X}'\,\mathbf{M}^0)(\mathbf{M}^0\mathbf{X})\mathbf{b} + \boldsymbol{e}'\boldsymbol{e}$

$= \mathbf{b}'\mathbf{X}'\,\mathbf{M}^0\mathbf{Xb} + \boldsymbol{e}'\boldsymbol{e}$ ($\mathbf{M}^0$ is idempotent & $\boldsymbol{e}'\,\mathbf{M}^0\mathbf{X} = \mathbf{0}$)

TSS = SSR + RSS

TSS: Total sum of squares

SSR: Regression Sum of Squares (also called ESS: *explained* SS)

RSS: Residual Sum of Squares (also called SSE: SS of errors)

Not to be posted/distributed online without written consent from author.

A Goodness of Fit Measure

- TSS = SSR + RSS

• We want to have a measure that describes the fit of a regression. Simplest measure: the standard error of the regression (SER)

SER = sqrt{RSS/(T - k)} $\Rightarrow$ SER depends on units. Not good!

• R-squared (R^2)

1 = SSR/TSS + RSS/TSS

R^2 = SSR/TSS = Regression variation/Total variation

$R^2 = \mathbf{b'X'M^0Xb}/\ \boldsymbol{y}'\mathbf{M^0}\boldsymbol{y} = 1 - \boldsymbol{e}'\boldsymbol{e}/\boldsymbol{y}'\mathbf{M^0}\boldsymbol{y}$

$= (\hat{\boldsymbol{y}} - \mathbf{i}\bar{y})'\ (\hat{\boldsymbol{y}} - \mathbf{i}\bar{y})/(\boldsymbol{y} - \mathbf{i}\bar{y})'\ (\boldsymbol{y} - \mathbf{i}\bar{y})$

$=[\ \hat{\boldsymbol{y}}'\ \hat{\boldsymbol{y}} - T\,\bar{y}^2]/[\boldsymbol{y}'\ \boldsymbol{y} - T\,\bar{y}^2]$

A Goodness of Fit Measure

- R^2 = SSR/TSS = $\mathbf{b'X'M^0Xb}/\boldsymbol{y}'\mathbf{M^0}\boldsymbol{y} = 1 - \mathbf{e'e}/\boldsymbol{y}'\mathbf{M^0}\boldsymbol{y}$

Note: R^2 is bounded by zero and one only if:

(a) There is a constant term in **X** –we need $\mathbf{e'\,M^0X=0}$!

(b) The line is computed by linear least squares.

• Adding regressors

R^2 never falls when regressors (say **z**) are added to the regression.

$$R_{Xz}^2 = R_X^2 + (1 - R_X^2)r_{yz}^{*2}$$

r_{yz}: partial correlation coefficient between **y** and **z**.

Problem: Judging a model based on R^2 tends to over-fitting.

A Goodness of Fit Measure

• Comparing Regressions

- Make sure the denominator in R^2 is the same - i.e., same left hand side variable.

Example: Linear vs. Loglinear. Loglinear will almost always appear to fit better because taking logs reduces variation.

• Linear Transformation of data

- Based on **X**, $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'}\boldsymbol{y}$.

Suppose we work with $\mathbf{X}^* = \mathbf{XH}$, instead (**H** is not singular).

$$\mathbf{P}^*\boldsymbol{y} = \mathbf{X}^*\mathbf{b}^* = \mathbf{XH}(\mathbf{H'X'XH})^{-1}\mathbf{H'X'}\boldsymbol{y} \quad \text{(recall } (\mathbf{ABC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1})$$
$$= \mathbf{XHH}^{-1}(\mathbf{X'X})^{-1}\mathbf{H'}^{-1}\mathbf{H'X'}\boldsymbol{y}$$
$$= \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}\boldsymbol{y} = \mathbf{P}\boldsymbol{y}$$

$\Rightarrow$ same fit, same residuals, same R^2!

Adjusted R-squared

• R^2 is modified with a penalty for number of parameters: *Adjusted*-R^2

$$\overline{R}^2 = 1 - \frac{(T-1)}{(T-k)}(1 - R^2) = 1 - \frac{(T-1)}{(T-k)}\frac{RSS}{TSS}$$

$\Rightarrow$ maximizing $\overline{R}^2 \Leftrightarrow$ minimizing $[RSS/(T-k)] = s^2$

• *Degrees of freedom* –i.e., $(T-k)$-- adjustment assumes something about "unbiasedness."

• $\overline{R}^2$ includes a penalty for variables that do not add much fit. Can fall when a variable is added to the equation.

• It will rise when a variable, say **z**, is added to the regression if and only if the t-ratio on **z** is larger than one in absolute value.

Adjusted R-squared

• Theil (1957) shows that, under certain assumptions (an important one: the true model is being considered), if we consider two linear models

M_1: $\boldsymbol{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1$
M_2: $\boldsymbol{y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2$

and choose the model with smaller s^2 (or, larger Adusted R^2), we will select the true model, M_1, on average.

• In this sense, we say that "maximizing Adjusted R^2" is an *unbiased* model-selection criterion.

• In the context of model selection, the Adjusted R^2 is also referred as *Theil's information criteria*.

Other Goodness of Fit Measures

• There are other goodness-of-fit measures that also incorporate penalties for number of parameters (degrees of freedom).

• Information Criteria

- *Amemiya*: $[\mathbf{e}'\mathbf{e}/(T-k)] \times (1 + k/T)$

- *Akaike Information Criterion* (AIC)

$\text{AIC} = -2 * \ln L + 2 * k$ $\qquad L$: Likelihood

$\Rightarrow$ if normality $\text{AIC} = T * \ln(\mathbf{e}'\mathbf{e}/T) + 2 * k$ $\qquad$ (+constants)

- *Bayes-Schwarz Information Criterion* (BIC)

$\text{BIC} = -2 \ln L + \ln(T) * k$

$\Rightarrow$ if normality $\text{BIC} = T * \ln(\mathbf{e}'\mathbf{e}/T) + \ln(T) * k$ $\qquad$ (+constants)

Not to be posted/distributed online without written consent from author.

Other Goodness of Fit Measures

• It is common to ignore constants and divide by T. For example:

$$\text{AIC} = \ln(\mathbf{e'e}/T) + (2/T) * k$$

• AIC and BIC are very popular for model selection (the lower, the better). AIC has a small penalty for larger models (large k), BIC has a larger penalty.

• For some specific model selection strategies, Mallows C_p statistic is used (where p= k):

$$C_p = \text{RSS}(k)/s^2 - T + 2 * k$$

where RSS(k) is the RSS for the model with k regressors. C_p is closely related to $\bar{R}^2$ (Kennard (1971)).

OLS Estimation – Example in R

Example: 3 Factor F-F Model (continuation) for IBM returns:

```
b <- solve(t(x)%*% x)%*% t(x)%*%y               # b = (X'X)-1X' y  (OLS regression)
e <- y - x%*%b                                  # regression residuals, e
RSS <- as.numeric(t(e)%*%e)                     # RSS
R2 <- 1 - as.numeric(RSS)/as.numeric(t(y)%*%y)  # R-squared
Adj_R2_2 <- 1 - (T-1)/(T-k)*(1-R2)              # Adjusted R-squared
AIC <- log(RSS/T)+2*k/T                         # AIC under N(.,.) –i.e.,  under (A5)
```

```
> R2
[1] 0.5679013        ⇒ The 3 factors explain 57% of the variation of IBM returns
> Adj_R2_2
[1] 0.5639968
> AIC
[1] -3.233779
```

Not to be posted/distributed online without written consent from author.

Maximum Likelihood Estimation (MLE)

• Idea: Assume a particular distribution with unknown parameters. Maximum likelihood (ML) estimation chooses the set of parameters that maximize the likelihood of drawing a particular sample.

• Consider a sample (X_1, ... , X_n) which is drawn from a pdf $f(\boldsymbol{X}|\,\theta)$ where θ are parameters. If the X_i's are independent with pdf $f(X_i\,|\,\theta)$ the joint probability of the whole sample is:

$$L(X\,|\,\theta) = f(X_1 \cdots X_n\,|\,\theta) = \prod_{i=1}^{n} f(X_i\,|\,\theta)$$

The function $L(X|\,\theta)$ --also written as $L(X;\,\theta)$-- is called the *likelihood function*. This function can be maximized with respect to θ to produce maximum likelihood estimates: $\hat{\theta}_{MLE}$.

Maximum Likelihood Estimation (MLE)

• It is often convenient to work with the Log of the likelihood function. That is,

$$\ln L(X\,|\,\theta) = \Sigma_i \ln f(X_i\,|\,\theta).$$

• The ML estimation approach is very general. Now, if the model is not correctly specified, the estimates are sensitive to the misspecification.

Ronald A. Fisher, England (1890 – 1962)

Not to be posted/distributed online without written consent from author.

Maximum Likelihood Estimation: Example I

Let the sample be $\boldsymbol{X}$ = {5, 6, 7, 8, 9, 10} drawn from a Normal(μ, 1). The probability of each of these points based on the unknown mean, μ, can be written as:

$$f(5|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(5-\mu)^2}{2}\right]$$
$$f(6|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(6-\mu)^2}{2}\right]$$
$$\vdots$$
$$f(10|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(10-\mu)^2}{2}\right]$$

Assume that the sample is independent.

Maximum Likelihood Estimation: Example I

Then, the joint pdf function can be written as:

$$L(X|\mu) = \frac{1}{(2\pi)^{6/2}} \exp\left[-\frac{(5-\mu)^2}{2} - \frac{(6-\mu)^2}{2} - \cdots - \frac{(10-\mu)^2}{2}\right]$$

The value of μ that maximize the likelihood function of the sample can then be defined by $\max_{\mu} L(X|\mu)$.

It easier, however, to maximize the *log likelihood*, ln L(X | μ). That is,

$$\max_{\mu} \ln\big(L(X|\mu)\big) = -\,{}^{6}/_{2}\ln(2\pi) + \left[-\frac{(5-\mu)^2}{2} - \frac{(6-\mu)^2}{2} - \cdots - \frac{(10-\mu)^2}{2}\right]$$

1st-derivative $\Rightarrow \frac{\partial}{\partial\mu}\left[K - \frac{(5-\mu)^2}{2} - \frac{(6-\mu)^2}{2} - \cdots - \frac{(10-\mu)^2}{2}\right]$

f.o.c. $\Rightarrow$ $(5-\hat{\mu}_{MLE}) + (6-\hat{\mu}_{MLE}) + \cdots + (10-\hat{\mu}_{MLE}) = 0$

Not to be posted/distributed online without written consent from author.

Maximum Likelihood Estimation: Example I

Then, the first order conditions:

$$(5-\hat{\mu}_{MLE})+(6-\hat{\mu}_{MLE})+\cdots+(10-\hat{\mu}_{MLE})=0$$

Solving for $\hat{\mu}_{MLE}$:

$$\hat{\mu}_{MLE}=\frac{5+6+7+8+9+10}{6}=\bar{x}$$

Maximum Likelihood Estimation (MLE)

• Under the assumed econometric model, the sample is the most likely. We will assume the errors, $\boldsymbol{\varepsilon}$, follow a normal distribution:

(**A5**) $\boldsymbol{\varepsilon}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_T)$

• Then, we can write the joint pdf of $\boldsymbol{y}$ as

$$f(y_t|\beta,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{1/2}}*\exp\left[-\frac{(y_t-x_t\beta)^2}{2\sigma^2}\right]$$

$$L=\prod_{t=1}^{T}\frac{1}{(2\pi\sigma^2)^{1/2}}*\exp\left[-\frac{(y_t-x_t\beta)^2}{2\sigma^2}\right]$$

$$=(2\pi\sigma^2)^{-T/2}*\exp\left[-\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{2\sigma^2}\right]$$

Taking logs, we have the log likelihood function

$$\ln L=-\frac{T}{2}\ln 2\pi-\frac{T}{2}\ln\sigma^2-\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{2\sigma^2}$$

Not to be posted/distributed online without written consent from author.

MLE: Cheat-Sheet for Vector Derivatives

- Consider the linear function: $\boldsymbol{y} = f(\boldsymbol{x}) = \boldsymbol{x}'\,\boldsymbol{\beta} + \omega$

where $\boldsymbol{x}$ and $\boldsymbol{\beta}$ are k-dimensional vectors and ω is a constant.

Then, $\nabla f(\boldsymbol{x}) = \boldsymbol{\beta}$

- Consider a quadratic form: $q = f(\boldsymbol{x}) = \boldsymbol{x}'\,\mathbf{A}\,\boldsymbol{x}$

where $\boldsymbol{x}$ is kx1 vector and $\mathbf{A}$ is a kxk matrix, with a_{ji} elements.

Then, $\nabla f(\boldsymbol{x}) = = \mathbf{A}'\,\boldsymbol{x} + \mathbf{A}\,\boldsymbol{x} = (\mathbf{A}' + \mathbf{A})\,\boldsymbol{x}$

If $\mathbf{A}$ is symmetric, then $\nabla f(\boldsymbol{x}) = 2\,\mathbf{A}\,\boldsymbol{x}$

67

MLE: Vector Foc & Solution

- Let $\boldsymbol{\theta} = (\beta, \sigma)$. Then, we want

$$\begin{aligned}\text{Max}_{\theta}\ \{\ln L &= -\tfrac{T}{2}\ln(2\pi) - \tfrac{T}{2}\ln(\sigma^2) - \frac{(y - \mathbf{X}\boldsymbol{\beta})'(y - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \\ &= -\tfrac{T}{2}\ln(\sigma^2) - \frac{(y'y - \boldsymbol{\beta}'\mathbf{X}'y - y'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \\ &= -\tfrac{T}{2}\ln(\sigma^2) - \frac{(y'y - 2\boldsymbol{\beta}'\mathbf{X}'y + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\}\end{aligned}$$

- Then, 1st derivatives of $\ln L$ with respect to $\boldsymbol{\beta}$ & σ^2:

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2}(2\mathbf{X}'\mathbf{y}' - 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}) = \frac{1}{\sigma^2}\mathbf{X}'\boldsymbol{\varepsilon}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{T}{2\sigma^2} - (-\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{2\sigma^4}) = (\frac{1}{2\sigma^2})[\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\sigma^2} - T]$$

MLE: Vector Foc & Solution

• Then, the f.o.c.:

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2}\mathbf{X}'\boldsymbol{e} = \frac{1}{\sigma^2}\mathbf{X}'\big(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{MLE}\big) = 0 \quad \Rightarrow \widehat{\boldsymbol{\beta}}_{MLE} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = \left(\frac{1}{2\hat{\sigma}^2_{MLE}}\right)\left[\frac{\mathbf{e}'\mathbf{e}}{\hat{\sigma}^2_{MLE}} - T\right] = 0 \quad \Rightarrow \hat{\sigma}^2_{MLE} = \frac{\mathbf{e}'\mathbf{e}}{T} = \frac{\sum_{i=1}^{T}(y_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_{MLE})^2}{T}$$

Note: The f.o.c. deliver the normal equations for β! The solution to the normal equation, β_{MLE}, is also the LS estimator, **b**.

• Nice result for **b**: ML estimators have very good properties!

ML: Score and Information Matrix

Definition: Score (or efficient score)

$$S(X;\theta) = \frac{\delta \log(L(X \mid \theta))}{\delta\theta} = \sum_{i=1}^{n} \frac{\delta \log(f(x_i \mid \theta))}{\delta\theta}$$

S(X; θ) is called the *score* of the sample. It is the vector of partial derivatives (the gradient), with respect to the parameter θ. If we have k parameters, the score will have a kx1 dimension.

Definition: Fisher information for a single sample:

$$E\left[\left(\frac{\partial \log(f(X \mid \theta))}{\partial\theta}\right)^2\right] = I(\theta)$$

$I(\theta)$ is sometimes just called *information*. It measures the shape of the log $f(X|\theta)$.

Not to be posted/distributed online without written consent from author.

ML: Score and Information Matrix

• The concept of information can be generalized for the *k*-parameter case. In this case:

$$E\left[\left(\frac{\partial \log L}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log L}{\partial \boldsymbol{\theta}}\right)^{\mathrm{T}}\right] = \mathbf{I}(\theta)$$

This is *k*x*k* matrix.

If *L* is twice differentiable with respect to *θ*, and under certain regularity conditions, then the information may also be written as9

$$E\left[\left(\frac{\partial \log L}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log L}{\partial \boldsymbol{\theta}}\right)^{\mathrm{T}}\right] = E\left[-\left(\frac{\delta^2 \log(L(X\,|\,\boldsymbol{\theta}))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right)\right] = \mathbf{I}(\boldsymbol{\theta})$$

I(*θ*) is called the *information matrix* (negative Hessian). It measures the shape of the likelihood function.

ML: Score and Information Matrix

• Properties of S(*X*; *θ*):

$$S(X;\theta) = \frac{\delta \log(L(X|\theta))}{\delta\theta} = \sum_{i=1}^{n} \frac{\delta \log(f(x_i\,|\,\theta))}{\delta\theta}$$

(1) E[S(*X*; *θ*)] = 0.

$$\int f(x;\theta)dx = 1 \quad \Rightarrow \int \frac{\partial f(x;\theta)}{\partial\theta} dx = 0$$

$$\int \frac{1}{f(x;\theta)} \frac{\partial f(x;\theta)}{\partial\theta} f(x;\theta)dx = 0$$

$$\int \frac{\partial \log f(x;\theta)}{\partial\theta} f(x;\theta)dx = 0 \quad \Rightarrow E[S(x;\theta)] = 0$$

ML: Score and Information Matrix

(2) Var[S(*X*; *θ*)]= *n I*(*θ*)

$$\int \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta)dx = 0$$

Let's differentiate the above integral once more:

$$\int \frac{\partial \log f(x;\theta)}{\partial \theta} \frac{\partial f(x;\theta)}{\partial \theta} dx + \int \frac{\partial^2 \log f(x;\theta)}{\partial \theta \partial \theta'} f(x;\theta)dx = 0$$

$$\int \frac{\partial \log f(x;\theta)}{\partial \theta} \left(\frac{1}{f(x;\theta)} \frac{\partial f(x;\theta)}{\partial \theta} \right) f(x;\theta)dx + \int \frac{\partial^2 \log f(x;\theta)}{\partial \theta \partial \theta'} f(x;\theta)dx = 0$$

$$\int \left(\frac{\partial \log f(x;\theta)}{\partial \theta} \right)^2 f(x;\theta)dx + \int \frac{\partial^2 \log f(x;\theta)}{\partial \theta \partial \theta'} f(x;\theta)dx = 0$$

$$E\left[\left(\frac{\partial \log f(x;\theta)}{\partial \theta} \right)^2 \right] = -E\left[\frac{\partial^2 \log f(x;\theta)}{\partial \theta \partial \theta'} \right] = I(\theta)$$

$$Var[S(X;\theta)] = n\,Var[\frac{\partial \log f(x;\theta)}{\partial \theta}] = n\; I(\theta)$$

ML: Score and Information Matrix

(3) Asymptotic Normality

If S(x_i; θ) are *i.i.d.* (with finite first and second moments), then we can apply the CLT to get:

$$S_n(X;\theta) = \Sigma_i\, S(x_i;\theta) \xrightarrow{a} N(0,\ [nI(\theta)])$$

Note: This an important result. It will drive the distribution of ML estimators.

ML: Score and Information Matrix – Example

• Again, we assume:

$$y_i = X_i\boldsymbol{\beta} + \varepsilon_i \qquad \varepsilon_i \sim N(0,\sigma^2)$$

$$or \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim N(0,\sigma^2\mathbf{I}_T)$$

• Taking logs, we have the log likelihood function:

$$\ln L = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma^2) - \frac{(y-\mathbf{X}\boldsymbol{\beta})'(y-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}$$

• The score function is –first derivatives of log L w.r.t. $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$:

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2}\mathbf{X}'\boldsymbol{\varepsilon}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{T}{2\sigma^2} - (-\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{2\sigma^4}) = (\frac{1}{2\sigma^2})[\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\sigma^2} - T]$$

ML: Score and Information Matrix – Example

• Then, we take second derivatives to calculate I(θ): :

$$\frac{\partial \ln L^2}{\partial\beta\partial\beta'} = -\sum_{i=1}^{T}\mathbf{x}_i\mathbf{x}_i'/\sigma^2 = -\frac{1}{\sigma^2}X'X$$

$$\frac{\partial \ln L}{\partial\beta\partial\sigma^{2}{'}} = -\frac{1}{\sigma^4}\sum_{i=1}^{T}\varepsilon_i x_i'$$

$$\frac{\partial \ln L}{\partial\sigma^2\partial\sigma^{2}{'}} = -\frac{1}{2\sigma^4}[\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\sigma^2} - T] + (\frac{1}{2\sigma^2})(-\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\sigma^4}) = -\frac{1}{2\sigma^4}[2\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\sigma^2} - T]$$

• Then,

$$I(\theta) = E[-\frac{\partial \ln L}{\partial\theta\partial\theta'}] = \begin{bmatrix} (\frac{1}{\sigma^2}X'X) & 0 \\ 0 & \frac{T}{2\sigma^4} \end{bmatrix}$$

ML: Regularity Conditions

• In deriving properties (1) and (2), we have made some implicit assumptions, which are called *regularity conditions*:

(i) θ lies in an open interval of the parameter space, Ω.

(ii) The 1st derivative and 2nd derivatives of $f(X;\theta)$ w.r.t. θ exist.

(iii) $L(X;\theta)$ can be differentiated w.r.t. θ under the integral sign.

(iv) $E[S(X;\theta)^2] > 0$, for all θ in Ω.

(v) $T(X)\ L(X;\theta)$ can be differentiated w.r.t. θ under the integral sign.

Recall: If $S(X;\theta)$ are *i.i.d.* and regularity conditions apply, then we can apply the CLT to get:

$$S(X;\theta) \xrightarrow{a} N(0, n\,I(\theta))$$

ML: Cramer-Rao inequality

Theorem: Cramer-Rao inequality

Let the random sample $(X_1, \ldots, X_n)$ be drawn from a pdf $f(\boldsymbol{X}|\,\theta)$ and let $T=T(X_1, \ldots, X_n)$ be a statistic such that $E[T]=u(\theta)$, differentiable in θ. Let $b(\theta)= u(\theta) - \theta$, the bias in T. Assume regularity conditions. Then,

$$Var(T) \geq \frac{[u'(\theta)]^2}{nI(\theta)} = \frac{[1+b'(\theta)]^2}{nI(\theta)}$$

Regularity conditions:

(1) θ lies in an open interval Ω of the real line.

(2) For all θ in Ω, $\delta f(\boldsymbol{X}|\,\theta)/\delta\theta$ is well defined.

(3) $\int L(\boldsymbol{X}|\,\theta)dx$ can be differentiated wrt. θ under the integral sign

(4) $E[S(\boldsymbol{X};\theta)^2]>0$, for all θ in Ω

(5) $\int T(\boldsymbol{X})\ L(\boldsymbol{X}|\,\theta)dx$ can be differentiated wrt. θ under the integral sign

ML: Cramer-Rao inequality

$$Var(T) \geq \frac{[u'(\theta)]^2}{nI(\theta)} = \frac{[1+b'(\theta)]^2}{nI(\theta)}$$

The lower bound for Var(T) is called the *Cramer-Rao* (CR) *lower bound*.

Corollary: If T(**X**) is an unbiased estimator of θ, then

$$Var(T) \geq (nI(\theta))^{-1}$$

<u>Note</u>: This theorem establishes the superiority of the ML estimate over all others. The CR lower bound is the smallest theoretical variance. It can be shown that ML estimates achieve this bound, therefore, any other estimation technique can at best only equal it.

Properties of ML Estimators

(1) Efficiency. Under general conditions, we have that $\hat{\theta}_{MLE}$

$$Var(\hat{\theta}_{MLE}) \geq [nI(\theta)]^{-1}$$

The right-hand side is the Cramer-Rao lower bound (CR-LB). If an estimator can achieve this bound, ML will produce it.

(2) Consistency. We know that E[S(X_i; θ)]=0 and Var[S(X_i; θ)]= $I(\theta)$.

The consistency of ML can be shown by applying Khinchine's LLN to S(X_i; θ) and then to $S_n(X; \theta) = \Sigma_i S(X_i; \theta)$.

Then, do a 1st-order Taylor expansion of $S_n(X; \theta)$ around $\hat{\theta}_{MLE}$

$$S_n(X;\theta) = S_n(X;\hat{\theta}_{MLE}) + S_n'(X;\theta_n^*)(\theta - \hat{\theta}_{MLE}) \qquad |\theta - \theta_n^*| \leq |\theta - \hat{\theta}_{MLE}| < \varepsilon$$

$$S_n(X;\theta) = S_n'(X;\theta_n^*)(\theta - \hat{\theta}_{MLE})$$

$S_n(X; \theta)$ and ($\hat{\theta}_{MLE}$ - θ) converge together to zero (i.e., expectation).

Properties of ML Estimators

(3) Asymptotic Normality - Theorem:

Let the likelihood function be $L(X_1, X_2,\ldots, X_n | \theta)$. Under general conditions, the MLE of θ is asymptotically distributed as

$$\hat{\theta}_{MLE} \xrightarrow{a} N(\theta, [nI(\theta)]^{-1})$$

Sketch of a proof. Using the CLT, we've already established

$S_n(X; \theta) \xrightarrow{a} \mathrm{N}(0, nI(\theta))$.

Then, using a first order Taylor expansion as before, we get

$$\mathrm{S_n}(\mathrm{X}; \theta)\frac{1}{\mathrm{n}^{1/2}} = \mathrm{S_n}'(\mathrm{X}; \theta_n^*)\frac{1}{\mathrm{n}^{1/2}}(\theta - \hat{\theta}_{MLE})$$

Notice that $E[S_n'(x_i; \theta)] = -I(\theta)$. Then, apply the LLN to get

$S_n'(X; \theta_n^*)/n \xrightarrow{p} -I(\theta)$. (using $\theta_n^* \xrightarrow{p} \theta$.)

Now, algebra and Slutzky's theorem for RV get the final result.

Properties of ML Estimators

(4) Sufficiency. If a single sufficient statistic exists for θ, the MLE of θ must be a function of it. That is, $\hat{\theta}_{MLE}$ depends on the sample observations only through the value of a sufficient statistic.

(5) Invariance. The ML estimate is invariant under functional transformations. That is, if $\hat{\theta}_{MLE}$ is the MLE of θ and if $g(\theta)$ is a function of θ, then $g(\hat{\theta}_{MLE})$ is the MLE of $g(\theta)$.