Lecture 1
Least Squares

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.”
What is Econometrics?

- Economic Theory:
  - The CAPM
    \[ R_i - R_f = \beta_i (R_{MP} - R_f) \]

- Mathematical Statistics:
  - Method to estimate CAPM. For example,
    Linear regression:
    \[ R_i - R_f = \alpha_i + \beta_i (R_{MP} - R_f) + \varepsilon_i \]
  - Properties of \( b_i \) (the LS estimator of \( \beta_i \))
  - Properties of different tests of CAPM. For example, a t-test for \( H_0: \alpha_i = 0 \).

- Data: \( R_i, R_f, \) and \( R_{MP} \)

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.

Idea:
- There is a functional form relating a dependent variable $Y$ and $k$ variables $X$. This function depends on unknown parameters, $\theta$. The relation between $Y$ and $X$ is not exact. There is an error, $\epsilon$. We have $T$ observations of $Y$ and $X$.
- We will assume that the functional form is known:
  \[ y_i = f(x_i, \theta) + \epsilon_i \quad i=1, 2, ..., T. \]
- We will estimate the parameters $\theta$ by minimizing a sum of squared errors:
  \[ \min_\theta \left\{ S(\theta) = \sum_i \epsilon_i^2 \right\} \]

Least Squares Estimation

- We will use linear algebra notation. That is,
  \[ y = f(X, \theta) + \epsilon \]
  Vectors will be column vectors: $y$, $x_k$, and $\epsilon$ are $T \times 1$ vectors:
  \[
  y = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix} \quad \Rightarrow \quad y' = [y_1 \ y_2 \ \ldots \ y_T]
  \]
  \[
  x_k = \begin{bmatrix} x_{k1} \\ \vdots \\ x_{kT} \end{bmatrix} \quad \Rightarrow \quad x_k' = [x_{k1} \ x_{k2} \ \ldots \ x_{kT}]
  \]
  \[
  \epsilon_k = \begin{bmatrix} \epsilon_{k1} \\ \vdots \\ \epsilon_{kT} \end{bmatrix} \quad \Rightarrow \quad \epsilon' = [\epsilon_1 \ \epsilon_2 \ \ldots \ \epsilon_T]
  \]
  $X$ is a $T \times k$ matrix.
Least Squares Estimation

\( X \) is a \( T \times k \) matrix. Its columns are the \( k \) \( T \times 1 \) vectors \( x_k \). It is common to treat \( x_1 \) as vector of ones:

\[
x_j = \begin{bmatrix} x_{11} \\ \vdots \\ x_{1T} \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \quad \Rightarrow \quad \mathbf{x}_1' = [1 \, 1 \, \ldots \, 1] = \mathbf{1}.
\]

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

\[
x_k' \mathbf{1} = \mathbf{1}' \mathbf{x}_k = x_{k1} + x_{k2} + \ldots + x_{kT} = \sum_j x_{kj}
\]

Least Squares Estimation - Assumptions

• Typical Assumptions

(A1) DGP: \( y = f(X, \theta) + \varepsilon \) is correctly specified. For example, \( f(x, \theta) = X \beta \)

(A2) \( E[\varepsilon | X] = 0 \)

(A3) \( \text{Var}[\varepsilon | X] = \sigma^2 I_T \)

(A4) \( X \) has full column rank – \( \text{rank}(X) = k \), where \( T \geq k \).

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

Note: The errors, \( \varepsilon \), are called disturbances. They are not something we add to \( f(X, \theta) \) because we don’t know precisely \( f(X, \theta) \). No. The errors are part of the DGP.
Least Squares Estimation - Assumptions

• Assumption (A2) is called *regression*.

From Assumption (A2) we get:

(i) \( E[\varepsilon | X] = 0 \) \implies E[y | X] = f(X, \theta) + E[\varepsilon | X] = f(X, \theta)

That is, the observed \( y \) will equal E[y | X] + random variation.

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

\[ E[\varepsilon] = E_X[E[\varepsilon | X]] = E_X[0] = 0 \]

(iii) There is no information about \( \varepsilon \) in \( X \) \implies Cov(\varepsilon, X) = 0.

\[ Cov(\varepsilon, X) = E[(\varepsilon - 0)(X - \mu_X)] = E[\varepsilon X] \]
\[ \implies E[\varepsilon X] = E_X[E[\varepsilon X | X]] = E_X[E[\varepsilon | X]X] = 0 \quad \text{(using LIE)} \]
\[ \implies \text{That is,} \quad E[\varepsilon X] = 0 \implies \varepsilon \perp X. \]

Least Squares Estimation - Assumptions

• From Assumption (A3) we get

\[ \text{Var}[\varepsilon | X] = \sigma^2 I_T \implies \text{Var}[\varepsilon] = \sigma^2 I_T \]

*Proof:* \[ \text{Var}[\varepsilon] = E_X[\text{Var}[\varepsilon | X]] + \text{Var}_X[E[\varepsilon | X]] = \sigma^2 I_T. \]

This assumption implies

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

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

• From Assumption (A4) \implies the \( k \) independent variables in \( X \) are linearly independent. Then, the \( k \times k \) matrix \( X'X \) will also have full rank – i.e., \( \text{rank}(X'X) = k \).

To get asymptotic results we will need more assumption about \( X \).
Least Squares Estimation – f.o.c.

• Objective function: \( S(x_i, \theta) = \sum_i \epsilon_i^2 \)

• We want to minimize w.r.t to \( \theta \). That is,
  \[
  \min_\theta \{ S(x_i, \theta) = \sum_i [y_i - f(x_i, \theta)]^2 \}
  \]
  \[
  \Rightarrow \frac{d S(x_i, \theta)}{d \theta} = -2 \sum_i [y_i - f(x_i, \theta)] f'(x_i, \theta)
  \]
  f.o.c. \( \Rightarrow -2 \sum_i [y_i - f(x_i, \theta_{LS})] f'(x_i, \theta_{LS}) = 0 \)

Note: The f.o.c. deliver the normal equations.

The solution to the normal equation, \( \theta_{LS} \), is the LS estimator. The estimator \( \theta_{LS} \) is a function of the data \( (y_i, x_i) \).

CLM - OLS

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

  (A1') DGP: \( y = f(X, \theta) + \epsilon = X\beta + \epsilon \)

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

  (A1) DGP: \( y = X\beta + \epsilon \) is correctly specified.
  (A2) \( \mathbb{E}[\epsilon | X] = 0 \)
  (A3) \( \text{Var}[\epsilon | X] = \sigma^2 I_T \)
  (A4) \( X \) has full column rank – \( \text{rank}(X) = k \), where \( T \geq k \).

Objective function: \( S(x_i, \theta) = \sum_i \epsilon_i^2 = \epsilon' \epsilon = (y - X\beta)' (y - X\beta) \)

Normal equations: \( -2 \sum_i [y_i - f(x_i, \theta_{LS})] f'(x_i, \theta_{LS}) = -2 (y - Xb)' X = 0 \)

Solving for \( b \) \( \Rightarrow b = (XX)^{-1} Xy \)
CLM - OLS

• Example: One explanatory model.

(A1') DGP: \( y = \beta_1 + \beta_2 x + \varepsilon \)

Objective function: \( S(x, \theta) = \sum_i \varepsilon_i^2 = \sum_i (y_i - \beta_1 - \beta_2 x_i)^2 \)

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

\[
\begin{align*}
(\beta_1): & \quad -2 \sum_i (y_i - b_1 - b_2 x_i) (-1) = 0 \quad \Rightarrow \sum_i (y_i - b_1 - b_2 x_i) = 0 \quad (1) \\
(\beta_2): & \quad -2 \sum_i (y_i - b_1 - b_2 x_i) (-x_i) = 0 \quad \Rightarrow \sum_i (y_i x_i - b_1 x_i - b_2 x_i^2) = 0 \quad (2)
\end{align*}
\]

From (1):
\[
\sum_i y_i - \sum_i b_1 - b_2 \sum_i x_i = 0 \quad \Rightarrow b_1 = \bar{y} - b_2 \bar{x}
\]

From (2):
\[
\sum_i y_i x_i - (\bar{y} - b_2 \bar{x}) \sum_i x_i - b_2 \sum_i x_i^2 = 0 \quad \Rightarrow b_2 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{\text{cov}(y_i, x_i)}{\text{var}(x_i)}
\]
or, more elegantly,
\[
b_2 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{\text{cov}(y_i, x_i)}{\text{var}(x_i)}
\]

OLS Estimation: Second Order Condition

• OLS estimator: \( \hat{b} = (X'X)^{-1} X'y \)

Note: (i) \( b = \hat{b}_{OLS} \) (Ordinary LS. \( \text{Ordinary} = \text{linear} \))

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

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

• Q: Is \( b \) a minimum? We need to check the s.o.c.

\[
\frac{\partial(y - Xb)'(y - Xb)}{\partial b} = -2X'(y - Xb)
\]

\[
\frac{\partial^2(y - Xb)'(y - Xb)}{\partial b \partial b'} = \frac{\partial}{\partial b} \left( \frac{\partial(y - Xb)'(y - Xb)}{\partial b} \right) = \frac{\partial}{\partial b'} \text{ column vector} = 2X'X
\]
If there were a single $b$, we would require this to be positive, which it would be; $2\mathbf{x}'\mathbf{x} = 2\sum_{i=1}^{n} x_i^2 > 0$.

The matrix counterpart of a positive number is a positive definite matrix.
OLS Estimation - Properties

- The LS estimator of $\beta_{LS}$ when $f(x, \theta) = X\beta$ is linear is

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

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

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

1) $E[b | X] = E[\beta | X] + E[(XX')^{-1}X'\varepsilon | X] = \beta$ (unbiased if $CX=I$)

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

We can show that $b$ is BLUE (or MVLUE).

**Proof.**

Let $b^* = Cy$ (linear in $y$)

$$E[b^* | X] = E[Cy | X] = E[C(X\beta + \varepsilon) | X] = \beta$$

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

This result is known as the Gauss-Markov theorem.

3) If we make an additional assumption:

$$(A5) \varepsilon | X \sim iid N(0, \sigma^2 I_d)$$

we can derive the distribution of $b$.

Since $b = \beta + (XX')^{-1}X'\varepsilon$, we have that $b$ is a linear combination of normal variables $\Rightarrow b | X \sim iid N(\beta, \sigma^2 (X'X)^{-1})$
OLS Estimation - Variance

- Example: One explanatory model.

(A1') DGP: \( y = \beta_1 + \beta_2 x + \varepsilon \)

\[
\text{Var}[b \mid X] = \sigma^2 (XX)^{-1} = \sigma^2 \left[ \frac{\sum_i 1}{\sum_i x_i^2} \frac{\sum_i 1x_i}{\sum_i x_i^2} \right]^{-1} = \left[ T \bar{x} \ T \bar{x} \right]^{-1}
\]

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

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

Algebraic Results

- Important Matrices

(1) “Residual maker” \[ M = I_T - XX^{-1}X \]

\[ My = y - XX^{-1}X'y = y - Xb = e \quad \text{(residuals)} \]

\[ MX = 0 \]

- \( M \) is symmetric \(- M = M' \)
- \( M \) is idempotent \(- M^2 = M \)
- \( M \) is singular \(- M^{-1} \) does not exist. \( \Rightarrow \text{rank}(M) = T-k \)

(\( M \) does not have full rank. We have already proven this result.)
Algebraic Results

- Important Matrices

(2) “Projection matrix” \[ P = X(X'X)^{-1}X' \]

\[ Py = X(XX)^{-1}X'y = Xb = \hat{y} \quad \text{(fitted values)} \]

Py is the projection of y into the column space of X.

PM = MP = 0 (Projection matrix)

PX = X

- P is symmetric \[ - P = P' \]
- P is idempotent \[ - P^2 = P \]
- P is singular \[ - P^{-1} \text{ does not exist.} \quad \Rightarrow \text{rank}(P)=k \]

Algebraic Results

- Disturbances and Residuals

In the population: \[ E[X' \varepsilon] = 0. \]
In the sample: \[ X'e = X'(y-Xb) = X'y - X'X(XX)^{-1}X'y = 1/T(X'e) = 0. \]

- We have two ways to look at y:

\[ y = E[y | X] + \varepsilon = \text{Conditional mean + disturbance} \]
\[ y = Xb + e = \text{Projection + residual} \]
Results when X Contains a Constant Term

• Let the first column of X be a column of ones. That is 
  \( X = [1, x_2, \ldots, x_k] \)

• Then,
  (1) Since \( X'e = 0 \)  \( \Rightarrow x_1'e = 0 \) – the residuals sum to zero.
  (2) Since \( y = Xb + e \)  \( \Rightarrow \text{'}y = \text{'}Xb + \text{'}e = \text{'}Xb \)

  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)
  c <- y - x%*%b  # regression residuals, e
  RSS <- as.numeric(t(c)%*%c)  # RSS
  Sigma2 <- as.numeric(RSS/(T-k))  # Estimated \( \sigma^2 = \hat{\sigma}^2 \) (See Chapter 2)
  Var_b <- Sigma2*solve(t(x)%*%x)  # Estimated \( \text{Var}[b|X] = \hat{\sigma}^2 (X'X)^{-1} \)
  SE_b <- sqrt(diag(Var_b))  # SE[b|X]
OLS Estimation – Example in R

```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):

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

---

Frisch-Waugh (1933) Theorem

- **Context:** Model contains two sets of variables:
  
  \[ X = \begin{bmatrix} [1, time] \mid \text{other variables} \end{bmatrix} = [X_1, X_2] \]
  
- **Regression model:**
  \[
  y = X_1\beta_1 + X_2\beta_2 + \varepsilon \quad \text{(population)}
  \]
  \[
  y = X_1b_1 + X_2b_2 + e \quad \text{(sample)}
  \]

- **OLS solution:** \( b = (XX)^{-1}X'\ y = \begin{bmatrix} X_1'X_1 & X_1'X_2 \end{bmatrix}^{-1}[X_1'y] \begin{bmatrix} X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1}[X_2'y] \)

Problem in 1933: Can we estimate \( \beta_2 \) without inverting the \((k_1+k_2)x(k_1+k_2)\) \( XX \) matrix? The F-W theorem helps reduce computation, by getting simplified algebraic expression for OLS coefficient, \( b_2 \).
F-W: Partitioned Solution

• Direct manipulation of normal equations produces

\[(X'X)b = X'y\]

\[X = [X_1, X_2] \text{ so } X'X = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \text{ and } X'y = \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix}\]

\[(X'X)b = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix}\]

\[X_1'X_1b_1 + X_1'X_2b_2 = X_1'y\]

\[X_2'X_1b_1 + X_2'X_2b_2 = X_2'y \Rightarrow X_2'X_2b_2 = X_2'y - X_2'X_1b_1\]

\[= X_2'(y - X_1b_1)\]

Note: if \(X_2'X_1 = 0\) \(\Rightarrow b_2 = (X_2'X_2)^{-1}X_2'y\)

• Use of the partitioned inverse result produces a fundamental result:

What is the southeast element in the inverse of the moment matrix?

\[\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1}\]

\[\begin{bmatrix} \cdot \cdot^{1}_{(2,2)} \cdot \cdot \cdot^{1}_{(2,2)} \cdot \cdot \end{bmatrix}\]

• With the partitioned inverse, we get:

\[b_2 = [\cdot^{1}_{(2,1)} X_1'y + [\cdot^{1}_{(2,2)} X_2'y\]

F-W: Partitioned Solution
**F-W: Partitioned Solution**

- Recall from the Linear Algebra Review:

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

2. \[
\begin{bmatrix}
R_2 - \Sigma_{YX}R_1 \\
\Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}
\end{bmatrix}^{-1}R_2
\xrightarrow{\Sigma_{XX}^{-1}\Sigma_{XY}}
\begin{bmatrix}
I & \Sigma_{XX}^{-1}\Sigma_{XY} & \Sigma_{XX}^{-1} & 0 \\
0 & I & D(-\Sigma_{XX}^{-1}\Sigma_{XY}) & D
\end{bmatrix}
\]

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

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

**F-W: Partitioned Solution**

- Then,

1. Matrix \( XX = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \)

2. Inverse: \[
\begin{vmatrix}
(X_1'X_1)^{-1} + (X_1'X_1)^{-1}X_1'X_2DX_2'(X_1'X_1)^{-1}X_1'X_2D \\
-DX_2'(X_1'X_1)^{-1}X_1'X_1X_1'X_1X_1'X_1X_1'X_2D
\end{vmatrix}
\]

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

\( \Rightarrow D = [X_2'M_1X_2]^{-1} \)

The algebraic result is: \([ ]^{-1}_{(2,1)} = -D X_2'X_1(IX_1X_1')^{-1} \)

\([ ]^{-1}_{(2,2)} = D = [X_2'M_1X_2]^{-1} \)

Interesting cases: \( X_1 \) is a single variable & \( X_1 = i \).

- Then, continuing the algebraic manipulation:

\( b_2 = [ ]^{-1}_{(2,1)} X_1'y + [ ]^{-1}_{(2,2)} X_2'y = [X_2'M_1X_2]^{-1}X_2'M_1y \)
F-W: Partitioned Solution - Results

• Then, continuing the algebraic manipulation:
\[
\begin{align*}
    b_2 &= \left[X_2'M_1X_2\right]^{-1}X_2'M_1y \\
    &= \left[X_2'M_1'M_1X_2\right]^{-1}X_2'M_1'M_1y \\
    &= \left[X_2'^*X_2'^*\right]^{-1}X_2'^*y^*
\end{align*}
\]
where \(Z^* = M_1Z\) = residuals from a regression of \(Z\) on \(X_1\).

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 \(b_2\):
  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:
\[
\begin{align*}
    b_2 &= \left[X_2'^*X_2'^*\right]^{-1}X_2'^*y^* \\
    &= \left[X_2'M_1X_2\right]^{-1}X_2'M_1'y = \left[X_2'M_1'M_1X_2\right]^{-1}X_2'M_1'M_1'y
\end{align*}
\]

• Implications
  - We can isolate a single coefficient in a regression.
  - It is not necessary to ‘partial’ the other \(Xs\) out of \(y\) (\(M_1\) is idempotent)
  - Suppose \(X_1 \perp X_2\). Then, we have the orthogonal regression
  \[
  \begin{align*}
  \Rightarrow b_2 &= (X_2'X_2)^{-1}X_2'y \\
  b_1 &= (X_1'X_1)^{-1}X_1'y
  \end{align*}
  \]
Frisch-Waugh Result: Implications

Example: De-mean
Let \( X_1 = i \)

\[ P_1 = i (i' i)^{-1} i' = (T)^{-1} i' = i' / T \]

\[ M_1 z = z - i' z / T = z - i \bar{z} \quad (M_1 \text{ demeans } z) \]

\[ b_2 = [X_2' M_1' M_1 X_2]^{-1} X_2' M_1' M_1 y \]

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
Application: Detrending Y

\[ Y^* = Y - (\text{*****} + 167.95277 \times \text{Year}) \]

Application: Detrended Regression

Regression of detrended Gasoline (M1G) on detrended Price of Gasoline (M1PG) detrended Income (M1Y)

```
name list x1=y year5
regr x1=pq y rh=x1 rem=pqtor
regr x1=y rh=x1 rem=ytor
regr x1=g rh=x1 rem=gtor
```

| Variable | Coefficient | Standard Error | t-ratio | P(|T|>t) | Mean of X |
|----------|-------------|----------------|---------|---------|-----------|
| PQTOR    | -11.98265151 | 2.1171860       | -5.660  | 0.000   | 0.87954355E-14 |
| YSTOR    | 0.4781809512E-01 | 0.45281498E-02 | 10.565  | 0.000   | -0.85563848E-11 |

From the Gasoline data in Note 3.

| Variable | Coefficient | Standard Error | t-ratio | P(|T|>t) | Mean of X |
|----------|-------------|----------------|---------|---------|-----------|
| YEAR     | -2.195624001 | 0.90679770     | -2.422  | 0.0213  | 1997.5000 |
| PQ       | -11.98265151 | 2.18232568     | -5.491  | 0.000   | 2.3166111 |
| Y        | 0.4701309512E-01 | 0.46654904E-02 | 10.249  | 0.000   | 9232.9611 |
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.

\[
\text{Total variation} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = y'M^0y.
\]

where $M^0 = I - \bar{i}'\bar{i}^{-1}\bar{i} = \text{the M de-meaning matrix.}$

- Decomposition of total variation (assume $X_i = i$ – a constant.)
  \[
  y = Xb + e, \quad \text{so} \quad M^0y = M^0Xb + M^0e = M^0Xb + e \quad \text{(deviations from means)}
  \]
  \[
  y'M^0y = b'(X'M^0(M^0X)b + e'e
  \]
  \[
  = b'X'M^0Xb + e'e. \quad \text{(M}^0\text{ is idempotent & e' M}^0X = 0)
  \]

\[
\text{TSS} = \text{SSR} + \text{RSS}
\]

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)
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)
  \[ \text{SER} = \sqrt{\frac{\text{RSS}}{T-k}} \]
  \( \Rightarrow \) SER depends on units. Not good!

• R-squared \((R^2)\)
  \[ 1 = \frac{\text{SSR}}{\text{TSS}} + \frac{\text{RSS}}{\text{TSS}} \]
  \[ R^2 = \frac{\text{SSR}}{\text{TSS}} = \frac{\text{Regression variation}}{\text{Total variation}} \]
  \[ R^2 = \frac{b'X'M^0Xb}{y'M^0y} = 1 - \frac{e'e}{y'M^0y} \]
  \[ = (\hat{y} - \bar{y})'(\hat{y} - \bar{y}) / (y - \bar{y})' (y - \bar{y}) = [\hat{y}'\hat{y} - T\bar{y}^2] / [y'y - T\bar{y}^2] \]

Note: \(R^2\) is bounded by zero and one only if:
(a) There is a constant term in \(X\) --we need \(e' M^0 X = 0!\)
(b) The line is computed by linear least squares.

• Adding regressors
  \(R^2\) never falls when regressors (say \(z\)) are added to the regression.
  \[ R^2_{xz} = R^2_X + (1 - R^2_X) r_{yz}^2 \]
  \(r_{yz}\) 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 R² 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 \( X, b = (X'X)^{-1}X'y \).
  Suppose we work with \( X^* = XH \), instead \( (H \) is not singular).
  \[ P^*y = X^*b^* = XH(H'X'H)^{-1}H'X'y \] (recall \((ABCD)^{-1} = C^{-1}B^{-1}A^{-1}\))
  \[ = XHH^{-1}(X'X)^{-1}H'H'X'y \]
  \[ = X(X'X)^{-1}X'y = Py \]
  \[ \Rightarrow \text{same fit, same residuals, same } R^2! \]

Adjusted R-squared

• \( R^2 \) is modified with a penalty for number of parameters: \( Adjusted-R^2 \)
  \[ R^2 = 1 - \frac{(T-1)/(T-k)}{(1 - R^2)} = 1 - \frac{RSS}{TSS} \]
  \[ = 1 - \frac{RSS/(T-k)}{TSS} \]
  \[ \Rightarrow \text{maximizing } R^2 \Leftrightarrow \text{minimizing } \frac{RSS}{T-k} = \hat{R}^2 \]

• Degrees of freedom - i.e., \( T-k \) - adjustment assumes something about “unbiasedness.”

• \( \hat{R}^2 \) 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
  \[ M_1: \quad y = X_1 \beta_1 + \varepsilon_1 \]
  \[ M_2: \quad y = X_2 \beta_2 + \varepsilon_2 \]

  and choose the model with smaller \( s^2 \) (or, larger Adjusted \( R^2 \)), 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*: \( [e'e/(T - K)] \times (1 + k/T) \)

  - *Akaike Information Criterion* (AIC)
    \[
    \text{AIC} = -2/T \ln L - k \quad L: \text{Likelihood} \\
    \Rightarrow \text{if normality AIC} = \ln(e'e/T) + (2/T) k \quad (+\text{constants})
    \]

  - *Bayes-Schwarz Information Criterion* (BIC)
    \[
    \text{BIC} = -(2/T) \ln L - [\ln(T)/T] k \\
    \Rightarrow \text{if normality AIC} = \ln(e'e/T) + [\ln(T)/T] k \quad (+\text{constants})
    \]
OLS Estimation – Example in R

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

```r
b <- solve(t(x)%*% x)%*% t(x)%*%y # \( \hat{\beta} = (XX')^{-1}X'y \) (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)
```

```r
c > R2
[1] 0.5679013
c > Adj_R2_2
c [1] 0.5639968
c > AIC
[1] -3.233779
```

Maximum Likelihood Estimation (MLE)

- We will assume the errors, \( \varepsilon \), follow a normal distribution:

\[(A5) \varepsilon | X \sim N(0, \sigma^2 I_T)\]

- Then, we can write the joint pdf of \( y \) as

\[
f(y_1, y_2, ..., y_T | \beta, \sigma^2) = \Pi_{t=1}^{T} \left( \frac{1}{2\pi\sigma^2} \right)^{1/2} \exp\left[ -\frac{1}{2\sigma^2} (y_t - x_t'\beta)^2 \right]
\]

\[
L = f(y_1, y_2, ..., y_T | \beta, \sigma^2) = \Pi_{t=1}^{T} \left( \frac{1}{2\pi\sigma^2} \right)^{1/2} \exp\left[ -\frac{1}{2\sigma^2} (y_t - x_t'\beta)^2 \right]
\]

\[
= \frac{1}{(2\pi\sigma^2)^{T/2}} \exp\left( -\frac{1}{2\sigma^2} \varepsilon'\varepsilon \right)
\]

Taking logs, we have the log likelihood function

\[
\ln L = -\frac{T}{2} \ln 2\pi - \frac{T}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \varepsilon'\varepsilon
\]
Maximum Likelihood Estimation (MLE)

• Let \( \theta = (\beta, \sigma) \). Then, we want to

\[
\max_0 \ln L(\theta \mid y, X) = -\frac{T}{2} \ln 2\pi \sigma^2 - \frac{T}{2} \sigma^2 (y - X\beta)'(y - X\beta)
\]

• Then, the f.o.c.:

\[
\frac{\partial \ln L}{\partial \beta} = - \frac{1}{2\sigma^2} (-2X'y - 2X'\beta) = \frac{1}{\sigma^2} (X'y - X'\beta) = 0
\]

\[
\frac{\partial \ln L}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta) = 0
\]

**Note:** The f.o.c. deliver the normal equations for \( \beta \)! The solution to the normal equation, \( \hat{\beta}_{MLE} \), is also the LS estimator, \( \hat{b} \). That is,

\[
\hat{\beta}_{MLE} = \hat{b} = (X'X)^{-1} Xy; \quad \hat{\sigma}^2_{MLE} = \frac{\hat{e}'\hat{e}}{T}
\]

• Nice result for \( b \): ML estimators have very good properties!

ML: Score and Information Matrix

**Definition:** Score (or efficient score)

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

\( S(X; \theta) \) is called the score of the sample. It is the vector of partial derivatives (the gradient), with respect to the parameter \( \theta \). If we have \( k \) parameters, the score will have a \( k \times 1 \) dimension.

**Definition:** Fisher information for a single sample:

\[
E \left[ \left( \frac{\delta \ln(f(X \mid \theta))}{\delta \theta} \right)^2 \right] = I(\theta)
\]

\( I(\theta) \) is sometimes just called information. It measures the shape of the log \( f(X \mid \theta) \).
ML: Score and Information Matrix

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

$$E \left[ \begin{pmatrix} \frac{\partial \log L}{\partial \theta} \\ \frac{\partial \log L}{\partial \theta} \end{pmatrix}^{\top} \right] = I(\theta)$$

This is $k \times k$ matrix.

If $L$ is twice differentiable with respect to $\theta$, and under certain regularity conditions, then the information may also be written as:

$$E \left[ \begin{pmatrix} \frac{\partial \log L}{\partial \theta} \\ \frac{\partial \log L}{\partial \theta} \end{pmatrix}^{\top} \right] = E \left[ \begin{pmatrix} \frac{\partial^2 \log L(X|\theta)}{\partial \theta \partial \theta} \end{pmatrix} \right] = I(\theta)$$

$I(\theta)$ is called the information matrix (negative Hessian). It measures the shape of the likelihood function.

ML: Score and Information Matrix

• Properties of $S(X; \theta)$:

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

(1) $E[S(X; \theta)] = 0$.

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

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

$$\int \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta)dx = 0 \Rightarrow E[S(x;\theta)] = 0$$
ML: Score and Information Matrix

(2) $\text{Var}[S(X; \theta)] = n I(\theta)$

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

Let's differentiate the above integral once more:

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

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

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

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

ML: Score and Information Matrix

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

\[
S_n(X; \theta) = \sum_i S(x_i; \theta) \rightarrow \mathcal{N}(0, n I(\theta)).
\]

Note: This an important result. It will drive the distribution of ML estimators.
• Again, we assume:
  \[ y_i = X_i \beta + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2) \]
  or
  \[ y = X\beta + \varepsilon \quad \varepsilon \sim N(0, \sigma^2 I_T) \]

• Taking logs, we have the log likelihood function:
  \[
  \ln L = -\frac{T}{2} \ln 2\pi \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{T} \varepsilon_i^2 = -\frac{T}{2} \ln 2\pi - \frac{T}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta)
  \]

• The score function is –first derivatives of log L wrt \( \theta = (\beta, \sigma^2) \):
  \[
  \frac{\partial \ln L}{\partial \beta} = +\frac{T}{2} \sum_{i=1}^{T} 2\varepsilon_i X_i / \sigma^2 = +\frac{1}{\sigma^2} X'\varepsilon
  \]
  \[
  \frac{\partial \ln L}{\partial \sigma^2} = -\frac{T}{2\sigma^2} - \left(-\frac{1}{2\sigma^4}\right) \sum_{i=1}^{T} \varepsilon_i^2 = \left(-\frac{1}{2\sigma^2}\right) \left[ \frac{(X'X)^{-1}}{\sigma^2} \right] - T
  \]

• Then, we take second derivatives to calculate \( I(\theta) \):
  \[
  \frac{\partial^2 \ln L}{\partial \beta \partial \beta'} = +\sum_{i=1}^{T} x_i x_i' / \sigma^2 = \frac{1}{\sigma^2} X'X
  \]
  \[
  \frac{\partial^2 \ln L}{\partial \beta \partial \sigma^2} = -\frac{1}{\sigma^4} \sum_{i=1}^{T} \varepsilon_i x_i'
  \]
  \[
  \frac{\partial^2 \ln L}{\partial \sigma^2 \partial \sigma^2} = \frac{1}{2\sigma^4} \left[ \frac{X'X}{\sigma^2} \right] - T + \frac{1}{2\sigma^2} \left( \frac{X'X}{\sigma^4} \right) = -\frac{1}{2\sigma^4} \left[ 2\frac{X'X}{\sigma^2} \right] - T
  \]

• Then,
  \[
  I(\theta) = \mathbb{E}[\frac{\partial^2 \ln L}{\partial \theta \partial \theta'}] = \begin{bmatrix}
  \frac{1}{\sigma^2} X'X & 0 \\
  0 & T/2\sigma^4
  \end{bmatrix}
  \]
ML: Score and Information Matrix

- In deriving properties (1) and (2), we have made some implicit assumptions, which are called regularity conditions:
  
  (i) \( \theta \) lies in an open interval of the parameter space, \( \Omega \).
  
  (ii) The 1st derivative and 2nd derivatives of \( f(X; \theta) \) w.r.t. \( \theta \) exist.
  
  (iii) \( L(X; \theta) \) can be differentiated w.r.t. \( \theta \) under the integral sign.
  
  (iv) \( E[S(X; \theta)^2] > 0 \), for all \( \theta \) in \( \Omega \).
  
  (v) \( T(X) L(X; \theta) \) can be differentiated w.r.t. \( \theta \) 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 \( f(X|\theta) \) and let \( T=T(X_1, \ldots, X_n) \) be a statistic such that \( E[T]=u(\theta) \), differentiable in \( \theta \). Let \( b(\theta) = u(\theta) - \theta \), the bias in \( T \). Assume regularity conditions. Then,

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

Regularity conditions:

(1) \( \theta \) lies in an open interval \( \Omega \) of the real line.

(2) For all \( \theta \) in \( \Omega \), \( \delta f(X|\theta) / \delta \theta \) is well defined.

(3) \( \int L(X|\theta)dx \) can be differentiated wrt. \( \theta \) under the integral sign

(4) \( E[S(X;\theta)^2] > 0 \), for all \( \theta \) in \( \Omega \)

(5) \( \int T(X) L(X; \theta)dx \) can be differentiated wrt. \( \theta \) under the integral sign
The lower bound for \( \text{Var}(T) \) is called the \textit{Cramer-Rao (CR) lower bound}.

\textbf{Corollary}: If \( T(\mathbf{X}) \) is an unbiased estimator of \( \theta \), then

\[
\text{Var}(T) \geq (nI(\theta))^{-1}
\]

\textbf{Note}: 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.

\textbf{Properties of ML Estimators}

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

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

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

(2) \textit{Consistency}. We know that \( \text{E}[S(X; \theta)]=0 \) and \( \text{Var}[S(X; \theta)]=I(\theta) \).

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

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

\[
S_n(X; \theta)=S_n(X; \hat{\theta}_{MLE})+S_n'(X; \hat{\theta}_{n}) (\theta-\hat{\theta}_{MLE}) \quad |\theta-\hat{\theta}_{n}^*| \leq |\theta-\hat{\theta}_{MLE}|<\varepsilon
\]

\[
S_n(X; \theta)=S_n'(X; \hat{\theta}_{n}^*) (\theta-\hat{\theta}_{MLE})
\]

\( S_n(X; \theta) \) and (\( \hat{\theta}_{MLE}^* \ \theta \)) converge together to zero (i.e., expectation).
Properties of ML Estimators

(3) Theorem: Asymptotic Normality

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

\[ \hat{\theta}_{MLE} \overset{a}{\longrightarrow} N\left(\theta, \left[nI(\theta)\right]^{-1}\right) \]

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

\[ S_n(X; \theta) \overset{a}{\longrightarrow} N(0, nI(\theta)). \]

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

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

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

\[ S_n'(X; \theta^*) / n \overset{a}{\longrightarrow} -I(\theta). \quad \text{(using } \theta^* \overset{a}{\longrightarrow} \theta_0) \]

Now, algebra and Slutsky’s theorem for RV get the final result.

Properties of ML Estimators

(4) Sufficiency. If a single sufficient statistic exists for \( \theta \), the MLE of \( \theta \) 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 \( \theta \) and if \( g(\theta) \) is a function of \( \theta \), then \( g(\hat{\theta}_{MLE}) \) is the MLE of \( g(\theta) \).