L3. Classical Linear Regression Models

Yonghui Zhang

School of Economics, RUC

October 19, 2020

Readings

- Reading Chapter 1 of *Econometrics* (Hayashi).
- Reading Chapters 3-5 of *Econometrics* (Hansen, 2020)

Outline

1. Assumptions for the Ordinary Least Squares Regression

Basic setup

- Assume that the **population** model is

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + \varepsilon = \beta' X + \varepsilon$$

where $X = (1, X_2, \cdots, X_k)'$ and $\beta = (\beta_1, \beta_2, \cdots, \beta_k)'$.

- A **random sample** of n observations is drawn from the population:

$$Y_i = \beta' \mathbf{X}_i + \varepsilon_i \text{ for } i = 1, \cdots, n,$$

where $\mathbf{X}_i = (1, X_{i2}, \cdots, X_{ik})'$ is a $k \times 1$ vector.

Basic setup

- In matrix form

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

where

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

and

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{pmatrix} = \begin{pmatrix} 1 & X_{12} & \cdots & X_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n2} & \cdots & X_{nk} \end{pmatrix}.$$

The classical assumptions on the model

- **Assumption A.1 (Linearity)** $\{Y_i, \mathbf{X}_i\}_{i=1}^{n}$ *satisfies the* **linear** *relationship:*

$$Y_i = \beta' \mathbf{X}_i + \varepsilon_i$$

where β is a $k \times 1$ unknown parameter vector, $\mathbf{X}_i$ is a $k \times 1$ vector of independent variables (regressors, explanatory variables), ε_i is an unobservable disturbance/error term, and Y_i is the dependent variable (regressand).

- **Remarks.**
  1. The key notion of **linearity** is that the regression model is **linear in** β **rather than in** X_i.
  2. If the above LRM is **correctly specified** for $E(Y_i|\mathbf{X}_i)$, then

$$\beta = \frac{\partial E(Y_i|\mathbf{X}_i)}{\partial \mathbf{X}_i}.$$

The classical assumptions on the model

- **Assumption A.2 (Strict exogeneity)**

$$E\left(\varepsilon_i|\mathbf{X}\right) = E\left(\varepsilon_i|\mathbf{X}_1, \cdots, \mathbf{X}_n\right) = 0 \ for \ i = 1, \ldots, n.$$

- **Remarks.**
  - By LIE, strict exogeneity implies that:
    - (i) $E\left(\varepsilon_i\right) = 0$;
    - (ii) $E\left(\varepsilon_i \mathbf{X}_j\right) = 0$;
    - (iii) $E\left[\varepsilon_i g\left(\mathbf{X}_1, \cdots, \mathbf{X}_n\right)\right] = 0$ for any function $g\left(\cdot\right)$
  - Note that
    - (i) & (ii) are necessary conditions for strict exogeneity;
    - (iii) implies strict exogeneity, or (iii) $\Leftrightarrow$ strict exogeneity.

The classical assumptions on the model

- **Remarks.**

1. For time series (TS) data, A.2 requires that ε_i does not depend on the past, current, or future values of the regressors $\mathbf{X}_i$. It rules out the **dynamic** model such as

$$Y_t = \beta_1 + \beta_2 Y_{t-1} + \varepsilon_t \text{ for } t = 1, \ldots, T$$

where $\varepsilon_t \backsim IID\left(0, \sigma^2\right)$. Let $\mathbf{X}_t = (1, Y_{t-1})'$. Note that $E\left(\mathbf{X}_t \varepsilon_t\right) = 0$ but $E\left(\mathbf{X}_{t+1} \varepsilon_t\right) \neq 0$. (Check $E\left(Y_t \varepsilon_t\right) \neq 0$). It follows that

$$E\left(\varepsilon_t | \mathbf{X}\right) \neq 0.$$

2. A.2 says nothing about higher order conditional moments. It allow for conditional heteroskedasticity:

$$E\left(\varepsilon_i^2 | \mathbf{X}_i\right) = \sigma^2\left(\mathbf{X}_i\right).$$

3. If $\mathbf{X}_i$, $i = 1, \ldots, n$ are non-stochastic, then A.2 becomes $E(\varepsilon_i) = 0$.

The classical assumptions on the model

- If $\{Y_i, X_i\}_{i=1}^n$ is an independent sample, then

$$E(\varepsilon_i|\mathbf{X}) = E(\varepsilon_i|\mathbf{X}_i) = 0.$$

- Test strict exogeneity: $H_0 : E[\varepsilon_i g(\mathbf{X}_i)] = 0$ for any $g(\cdot)$ with $E|\varepsilon_i g(\mathbf{X}_i)| < \infty$.
    - Let $p_1(x), p_2(x), \cdots, p_K(x)$ be basis functions such as $1, x, x^2, \ldots, x^K$ (or Fourier series, cubic spline, wavelets) with

$$K = K_n \to \infty \text{ as } n \to \infty$$

    - For any function $g(x)$ in some space of functions $\mathcal{G}$, we have $g(x) = \sum_{k=1}^K p_k(x)\beta_k + e^K(x)$, where $e^K(x)$ is the sieve approximation error with $\sup_{x\in\mathcal{X}}\sup_{g\in\mathcal{G}}\left|e^K(x)\right| \to 0$ as $K \to \infty$
    - Under H_0, for $\forall g$, $0 = E\left(\varepsilon_i g\left(\mathbf{X}_j\right)\right) \approx \sum_{k=1}^K \beta_k E[\varepsilon_i p_k(\mathbf{X}_i)]$, which implies K moments: $E\left[\varepsilon_i p_k\left(\mathbf{X}_j\right)\right] = 0$ for $k = 1, \cdots, K$.
    - Let $\hat{\varepsilon}_i$ be the residuals under H_0, we can check whether $n^{-1}\sum\hat{\varepsilon}_i p_k(\mathbf{X}_i)$ is close to 0 for $k = 1, \cdots, K$.

The classical assumptions on the model

- **Assumption A.3 (Nonsingularity)** *The rank of* $\mathbf{X}'\mathbf{X}$ *is* k *with probability 1.*
- **Assumption A.3* (Nonsingularity)** *The minimum eigenvalue of* $\mathbf{X}'\mathbf{X} = \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'$ *satisfies*

$$\lambda_{\min}(\mathbf{X}'\mathbf{X}) \to \infty \text{ as } n \to \infty \text{ with probability tending to 1.}$$

- **Remarks.**
  - Recall that the eigenvalues of a square matrix A are defined as the solution to $|A - \lambda I_k| = 0$. Let $\lambda_1, \cdots, \lambda_k$ denote the k eigenvalues with possible multiplicity.
  - A.3 rules out **perfect collinearity** among regressors in finite samples;
  - A.3* rules **asymptotic multicollinearity** in large samples.
  - Noting that $\mathbf{X}'\mathbf{X}$ is positive semi-definite (p.s.d.), A.3 also implies that $\lambda_{\min}(\mathbf{X}'\mathbf{X}) > 0$ in finite samples.

The classical assumptions on the model

- **Assumption A.4 (Spherical error variance)**

$$\begin{aligned} E\left(\varepsilon_i^2|\mathbf{X}\right) &= \sigma^2 \textit{ for all } i = 1, \ldots, n \textit{ (conditional homoskedasticity)} \\ E\left(\varepsilon_i\varepsilon_j|\mathbf{X}\right) &= 0 \textit{ for all } i, j = 1, \ldots, n \\ & \quad \textit{(conditional spatial/serial uncorrelatedness)} \end{aligned}$$

- **Remarks**
  - Together with A.2, A.4 implies that

$$\begin{aligned} \mathtt{Var}\left(\varepsilon_i\right) &= \mathtt{Var}\left[E\left(\varepsilon_i|\mathbf{X}\right)\right] + E\left[\mathtt{Var}\left(\varepsilon_i|\mathbf{X}\right)\right] = 0 + \sigma^2 = \sigma^2 \\ \mathtt{Cov}\left(\varepsilon_i, \varepsilon_j\right) &= \mathtt{Cov}\left[E\left(\varepsilon_i|\mathbf{X}\right), E\left(\varepsilon_j|\mathbf{X}\right)\right] + E\left[\mathtt{Cov}\left(\varepsilon_i, \varepsilon_j|\mathbf{X}\right)\right] \\ &= 0 + 0 = 0. \end{aligned}$$

    (Please verify the covariance decomposition formula by yourself.)
  - In matrix notation, we can express A.2 and A.4 as follows: $E\left(\varepsilon|\mathbf{X}\right) = \mathbf{0}_{n\times 1}$ and $E\left(\varepsilon\varepsilon'|\mathbf{X}\right) = I_n\sigma^2$.

2. Ordinary Least Squares Estimation

Estimation of coefficients

Definition (Ordinary least squares (OLS) estimator)

Let the residual sum of squares (RSS) of the LRM $Y_i = \beta' \mathbf{X}_i + \varepsilon_i$ as

$$RSS(\beta) = (\mathbf{Y} - \mathbf{X}\beta)' (\mathbf{Y} - \mathbf{X}\beta) = \sum_{i=1}^{n} \left(Y_i - \beta' \mathbf{X}_i \right)^2 .$$

Then the ordinary least squares (OLS) estimator $\hat{\beta}$ of β is given by

$$\hat{\beta} = \hat{\beta}_{OLS} = \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}} \, RRS(\beta) .$$

Theorem (OLS estimator)

Under A.1 and A.3 the OLS estimator $\hat{\beta}$ of β exists and is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} = \left(\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i' \right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i Y_i$$

Estimation of coefficients

Proof.

Noting that $RSS(\beta) = \sum_{i=1}^{n} \left(Y_i - \beta'\mathbf{X}_i\right)^2$, the FOC is given by

$$\begin{aligned}
\frac{\partial RRS(\beta)}{\partial \beta} &= \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \left(Y_i - \beta'\mathbf{X}_i\right)^2 = -2\sum_{i=1}^{n} \mathbf{X}_i \left(Y_i - \beta'\mathbf{X}_i\right) \\
&= -2\sum_{i=1}^{n} \mathbf{X}_i Y_i + 2\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i' \beta \\
&= -2\mathbf{X}'\mathbf{Y} + \mathbf{2X}'\mathbf{X}\beta = 0 \text{ when } \beta = \hat{\beta}.
\end{aligned}$$

It follows that $\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\beta$ and $\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Y}$ by A.3.
For SOC, $\frac{\partial^2 RRS(\beta)}{\partial \beta \partial \beta'} = 2\mathbf{X}'\mathbf{X}$ is p.d. under A.3. So SOC holds and $\hat{\beta}$ is the global minimizer. □

Estimation of coefficients

Remarks.

- $\hat{Y}_i = \mathbf{X}_i'\hat{\beta}$ is called the (in-sample) **fitted value** or **predicted value** of Y_i;
- $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ is called the (estimated) **residual** for Y_i.
- Let $\hat{\mathbf{Y}} = \left(\hat{Y}_1, \ldots, \hat{Y}_n\right)' = \mathbf{X}\hat{\beta}$ and $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)'$. Then $\mathbf{Y}$ has the following **orthogonal decomposition**:

$$\mathbf{Y} = \hat{\mathbf{Y}} + \hat{\boldsymbol{\varepsilon}}.$$

- The FOC implies that a very important equation, i.e., **normal equation**, holds:

$$\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\hat{\beta} = 0 \Rightarrow \mathbf{X}'\hat{\boldsymbol{\varepsilon}} = \sum_{i=1}^{n} \mathbf{X}_i\hat{\varepsilon}_i = 0$$

  by noting that $\mathbf{X}'\hat{\boldsymbol{\varepsilon}} = \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\hat{\beta} = 0$.
- The normal equation always hold no matter whether $E(\varepsilon_i|\mathbf{X}) = 0$ or not.

Estimation of coefficients

Remarks (Cont.)

- If $X_1 = 1$ (the model has the intercept), then $\sum_{i=1}^{n} \hat{\varepsilon}_i = \sum_{i=1}^{n} X_{i1} \cdot \hat{\varepsilon}_i = 0$.
- Exercise:
  - (i) Let $\overline{\hat{Y}} = n^{-1} \sum_{i=1}^{n} \hat{Y}_i$, and $\overline{Y} = n^{-1} \sum_{i=1}^{n} Y_i$. Demonstrate that $\overline{\hat{Y}} = \overline{Y}$ if an intercept is included in the LRM.
  - (ii) Show that $\hat{\mathbf{Y}}'\hat{\boldsymbol{\varepsilon}} = 0$ in the orthogonal decomposition of $\mathbf{Y} = \hat{\mathbf{Y}} + \hat{\boldsymbol{\varepsilon}}$.

Estimation of variance

- Recall that $\sigma^2 = E(\varepsilon_i^2)$ under Assumption A.4.
- We can estimate it by method of moments (MOM)

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = n^{-1} \hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}}.$$

- In finite samples, the above estimator is biased for σ^2. (We will see later)
- An unbiased estimator is given by

$$s^2 = \frac{1}{n-k} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = (n-k)^{-1} \hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}}.$$

Alternative Interpretation of OLS Estimator (MLE)

- Assume that $\varepsilon_i|\mathbf{X}_i \sim N(0,\sigma^2)$ in the LRM: $Y_i = \beta'\mathbf{X}_i + \varepsilon_i$ and ε_i's are independent given $\mathbf{X}$.
- Consider the *maximum likelihood estimator (MLE)* of σ^2 and β: Note that the likelihood of $\varepsilon_1, \ldots, \varepsilon_n$ conditional on $\mathbf{X}$ is given by

$$
\begin{aligned}
f\left(\varepsilon_1,\ldots,\varepsilon_n|\mathbf{X},\beta,\sigma^2\right) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(Y_i - \mathbf{X}_i'\beta\right)^2}{2\sigma^2}\right) \\
&= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \mathbf{X}_i'\beta\right)^2\right)
\end{aligned}
$$

The log-likelihood function is

$$
\begin{aligned}
L_n\left(\beta,\sigma^2\right) &= \log f\left(\varepsilon_1,\ldots,\varepsilon_n|\mathbf{X},\beta,\sigma^2\right) \\
&= -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \mathbf{X}_i'\beta\right)^2
\end{aligned}
$$

Alternative Interpretation of OLS Estimator

- To maximize the above log-likelihood function, we can obtain the FOCs:

$$\begin{cases} \frac{\partial L_n\left(\beta,\sigma^2\right)}{\partial \beta} = -\frac{1}{2\sigma^2}\sum_{i=1}^n \mathbf{X}_i\left(Y_i - \mathbf{X}_i'\beta\right) = 0 \\ \frac{\partial L_n(\beta,\sigma^2)}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n \left(Y_i - \mathbf{X}_i'\beta\right)^2 = 0 \end{cases}$$

$\Rightarrow$

$$\begin{cases} \hat{\beta}_{ML} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} = \hat{\beta}_{OLS} \\ \hat{\sigma}^2_{ML} = n^{-1}\sum_{i=1}^n \left(Y_i - \mathbf{X}_i'\hat{\beta}_{ML}\right)^2 = \hat{\sigma}^2 \end{cases}$$

Projection Matrices

- Recall that

$$\begin{aligned}\hat{\beta} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \\ \hat{\mathbf{Y}} &= \mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \equiv P\mathbf{Y} \\ \hat{\boldsymbol{\varepsilon}} &= \mathbf{Y} - \hat{\mathbf{Y}} = \left(I_n - P\right)\mathbf{Y} \equiv M\mathbf{Y}\end{aligned}$$

where

$$P \equiv \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$$

and

$$M \equiv I_n - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' = I_n - P.$$

Projection Matrices

Here are some important properties of P and M

(1) P and M are symmetric and idempotent so that they are projection matrices. The symmetry is obvious. We now check the idempotence:

$$\begin{aligned} P^2 &= \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' = P \\ M^2 &= (I_n - P)(I_n - P) = I_n - 2P + P^2 = I_n - M \end{aligned}$$

(2) $P\mathbf{X} = \mathbf{X}$, $M\mathbf{X} = 0$ and $PM = 0$.

(3) About the Trace

$$\begin{aligned} \mathtt{tr}(P) &= \mathtt{tr}\left(\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right) = \mathtt{tr}\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\right) = \mathtt{tr}\left(I_k\right) = k \\ \mathtt{tr}\left(M\right) &= \mathtt{tr}(I_n - P) = n - k \end{aligned}$$

Projection Matrices

Here are some important properties of P and M (Cont.)

(4) Residuals

$$\begin{aligned}\hat{\varepsilon} &= M\mathbf{Y} = M\left(\mathbf{X}\beta + \varepsilon\right) = M\varepsilon \\ \mathbf{Y} &= \left(P + M\right)\mathbf{Y} = P\mathbf{Y} + M\mathbf{Y} = \hat{\mathbf{Y}} + \hat{\varepsilon}\end{aligned}$$

Note that $\hat{\mathbf{Y}}'\hat{\varepsilon} = \mathbf{Y}'PM\mathbf{Y} = \mathbf{Y}'0\mathbf{Y} = \mathbf{0}$. Noting that $\hat{\mathbf{Y}} = P\mathbf{Y}$ and $M\mathbf{X} = 0$, so P is known as the "**hat matrix**" and M is called an **orthogonal projection** matrix or an "**annihilator matrix**".

(5) $\hat{\varepsilon}'\hat{\varepsilon} = \varepsilon' M\varepsilon = \mathbf{Y}'M\mathbf{Y}$.

Goodness of Fit

Define

$$\begin{aligned} TSS &= \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 : \text{total sum of squares} \\ ESS &= \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{\hat{Y}}\right)^2 : \text{explained sum of squares} \\ RSS &= \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 : \text{residual sum of squares} \end{aligned}$$

where $\overline{\hat{Y}} = n^{-1} \sum_{i=1}^{n} \hat{Y}_i$, and $\overline{Y} = n^{-1} \sum_{i=1}^{n} Y_i$. We consider two measures of coefficient of determination (R^2):

$$R_1^2 = 1 - \frac{RSS}{TSS} \text{ and } R_2^2 = \frac{ESS}{TSS}$$

Theorem

If an intercept is included in the regression, then
(i) $TSS = ESS + RSS$;
(ii) $R_1^2 = R_2^2 \in [0, 1]$.

Goodness of Fit

Proof. (i) When an intercept is included in the regression, we have $\sum_{i=1}^n \hat{\varepsilon}_i = 0$ which implies that $\overline{\hat{Y}} = \overline{Y}$. It follows that

$$
\begin{aligned}
TSS &= \sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^n \left(Y_i - \hat{Y}_i + \hat{Y}_i - \overline{\hat{Y}}\right)^2 \\
&= \sum_{i=1}^n \left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^n \left(\hat{Y}_i - \overline{\hat{Y}}\right)^2 \\
&\quad +2\sum_{i=1}^n \left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \overline{\hat{Y}}\right) \\
&= RSS + ESS + 2\sum_{i=1}^n \hat{\varepsilon}_i \left(\hat{Y}_i - \overline{\hat{Y}}\right) = RSS + ESS
\end{aligned}
$$

because of

$$
\begin{aligned}
\sum_{i=1}^n \hat{\varepsilon}_i \left(\hat{Y}_i - \overline{\hat{Y}}\right) &= \sum_{i=1}^n \hat{\varepsilon}_i \hat{Y}_i - \sum_{i=1}^n \hat{\varepsilon}_i \cdot \overline{\hat{Y}} \\
&= \sum_{i=1}^n \hat{\varepsilon}_i \mathbf{X}_i' \beta - \sum_{i=1}^n \hat{\varepsilon}_i \cdot \overline{\hat{Y}} = 0
\end{aligned}
$$

the normal equation. (ii) This follows from (ii).

Goodness of Fit

Remarks.

- Without an intercept, $R_1^2 \leq 1$ but can be negative, and $R_2^2 \geq 0$ but can be greater than 1.
- With an intercept, we can write

$$R^2 = R_1^2 = R_2^2$$

- In this case,

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS} = \frac{\hat{\mathbf{Y}}' M_0 \hat{\mathbf{Y}}}{\mathbf{Y}' M_0 \mathbf{Y}},$$

  where $M_0 = I_n - \iota\iota' / n = I_n - \iota \left(\iota'\iota\right) \iota'$ is the demeaned matrix and ι is an $n \times 1$ vector of ones.

Goodness of Fit

Remarks.(Cont.)

- It is easy to verify that M_0 is a symmetric idempotent matrix and thus a projection matrix. In addition,

$$\hat{\mathbf{Y}}-\iota\overline{\hat{Y}} = \begin{pmatrix} Y_1 - \overline{\hat{Y}} \\ \vdots \\ Y_n - \overline{\hat{Y}} \end{pmatrix} = \hat{\mathbf{Y}}-\iota\frac{\iota'\overline{\hat{\mathbf{Y}}}}{n} = \left(I_n - \frac{1}{n}\iota\iota'\right)\hat{\mathbf{Y}} = M_0\hat{\mathbf{Y}}$$

$$\hat{\mathbf{Y}}'M_0\hat{\mathbf{Y}} = \left(M_0\hat{\mathbf{Y}}\right)' M_0\hat{\mathbf{Y}} = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

- Similarly,

$$\mathbf{Y}'M_0\mathbf{Y} = \left(M_0\mathbf{Y}\right)' M_0\mathbf{Y} = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 \text{ and}$$

$$\mathbf{X}'M_0\mathbf{Y} = \left(M_0\mathbf{X}\right)' M_0\mathbf{Y} = \sum_{i=1}^{n}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\left(Y_i - \bar{Y}\right).$$

Goodness of Fit

Remarks.(Cont...)

- R^2 **never decreases** when we include some additional regressors to the LRM. (Note that $\hat{\beta}_{OLS} = \text{argmin}_{\beta \in \mathbb{R}^k} RRS(\beta)$)
  - High (low) R^2 does not necessarily imply a good (bad) model.
  - In macroeconomics, R^2 can be as high as 0.99 but in microeconomics or finance, R^2 can be as low as 0.1 or 0.2 but the model is still fine.
- When an intercept is included, R^2 indicates the sample correlation between Y_i and $\hat{Y}_i$:

$$
\begin{aligned}
R^2 &= \left[\widehat{\texttt{Corr}}\left(Y, \hat{Y}\right)\right]^2 = \frac{\left[\sum_{i=1}^n \left(\hat{Y}_i - \overline{\hat{Y}}\right)\left(Y_i - \overline{Y}\right)\right]^2}{\sum_{i=1}^n \left(\hat{Y}_i - \overline{\hat{Y}}\right)^2 \sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2} \\
&= \frac{\left(\hat{\mathbf{Y}}' M_0 \mathbf{Y}\right)^2}{\left(\hat{\mathbf{Y}}' M_0 \hat{\mathbf{Y}}\right)\left(\mathbf{Y}' M_0 \mathbf{Y}\right)}
\end{aligned}
$$

Goodness of Fit

Remarks.(Cont...)

Proof. With an intercept, we have $\sum_{i=1}^{n}\hat{\varepsilon}_i = 0$ and $\overline{\hat{Y}} = \overline{Y}$. Then it follows that

$$\begin{aligned} ESS &= \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{\hat{Y}}\right)^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{\hat{Y}}\right)\left(Y_i - \hat{\varepsilon}_i - \overline{Y}\right) \\ &= \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{\hat{Y}}\right)\left(Y_i - \overline{Y}\right) - \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{\hat{Y}}\right)\hat{\varepsilon}_i \\ &= \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{\hat{Y}}\right)\left(Y_i - \overline{Y}\right) = \hat{\mathbf{Y}}' M_0 \mathbf{Y} \end{aligned}$$

where we use $\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{\hat{Y}}\right)\hat{\varepsilon}_i = \sum_{i=1}^{n}\hat{Y}_i\hat{\varepsilon}_i - \overline{\hat{Y}}\sum_{i=1}^{n}\hat{\varepsilon}_i = 0 - 0 = 0$. Then

$$R^2 = \frac{ESS}{TSS} = \frac{ESS^2}{ESS \cdot TSS} = \frac{\left(\hat{\mathbf{Y}}' M_0 \mathbf{Y}\right)^2}{\left(\hat{\mathbf{Y}}' M_0 \hat{\mathbf{Y}}\right)\left(\mathbf{Y}' M_0 \mathbf{Y}\right)} = \left[\widehat{\mathtt{Corr}}\left(Y, \hat{Y}\right)\right]^2$$

Goodness of Fit

Definition ($\bar{R}^2$: R-bar-squared)

A better measure of goodness-of-fit is given by the adjusted coefficient of determination:

$$\bar{R}^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)} = 1 - \frac{n-1}{n-k}\left(1 - R^2\right)$$

- Note that $\bar{R}^2$ is **not a monotone** function of k. It may rise or fall as one adds one additional regressor to the regression model.
- Exercise.
  - (i) Assume that the linear model only include an intercept. If we add one more regressor into the model, please specify when $\bar{R}^2$ increases.
  - (ii) Assume that the linear model include k_0 regressors. If we add one more regressor into the model, please specify when $\bar{R}^2$ increases.

3. Finite Sample Properties of the OLS Estimators

Efficiency

Definition

An unbiased estimator $\hat{\beta}$ is more efficient than another unbiased estimator $\tilde{\beta}$ if $\mathtt{Var}(\tilde{\beta}) - \mathtt{Var}(\hat{\beta})$ is p.s.d.

Remarks.

- An important implication of the above definition is: for any $k \times 1$ vector C s.t. $C'C = 1$, we have

$$C' \left[\mathtt{Var}(\tilde{\beta}) - \mathtt{Var}(\hat{\beta})\right] C \geq 0.$$

  For example, taking $C = (1, 0, \ldots, 0)'$ yields $\mathtt{Var}(\tilde{\beta}_1) - \mathtt{Var}(\hat{\beta}_1) \geq 0$.

Efficiency

Remarks (Cont.)

- **Sufficient condition.** In the case where $E(\tilde{\beta}|\mathbf{X}) = E(\hat{\beta}|\mathbf{X}) = \beta$, it is sufficient to have

$$\mathtt{Var}(\tilde{\beta}|\mathbf{X}) - \mathtt{Var}(\hat{\beta}|\mathbf{X}) \text{ is p.s.d. with probability 1.}$$

To see why, recall that $\mathtt{Var}(Y) = \mathtt{Var}\left[E\left(Y|\mathbf{X}\right)\right] + E\left[\mathtt{Var}\left(Y|\mathbf{X}\right)\right]$. Then

$$\begin{aligned}
\mathtt{Var}(\tilde{\beta}) &= \mathtt{Var}\left[E\left(\tilde{\beta}|\mathbf{X}\right)\right] + E\left[\mathtt{Var}\left(\tilde{\beta}|\mathbf{X}\right)\right] = E\left[\mathtt{Var}\left(\tilde{\beta}|\mathbf{X}\right)\right] \\
\mathtt{Var}(\hat{\beta}|\mathbf{X}) &= \mathtt{Var}\left[E\left(\hat{\beta}|\mathbf{X}\right)\right] + E\left[\mathtt{Var}\left(\hat{\beta}|\mathbf{X}\right)\right] = E\left[\mathtt{Var}\left(\hat{\beta}|\mathbf{X}\right)\right]
\end{aligned}$$

and thus $\mathtt{Var}(\tilde{\beta}) - \mathtt{Var}(\hat{\beta}) = E\left[\mathtt{Var}(\tilde{\beta}|\mathbf{X}) - \mathtt{Var}(\hat{\beta}|\mathbf{X})\right]$ which is p.s.d. provided $\mathtt{Var}(\tilde{\beta}|\mathbf{X}) - \mathtt{Var}(\hat{\beta}|\mathbf{X})$ is p.s.d. with probability 1.

Finite sample properties of OLS

Theorem

Assume that the classical Assumptions A.1-A.4 hold. Then:
(a) (Unbiasedness) $E\left(\hat{\beta}|\mathbf{X}\right) = \beta$ *and* $E\left(\hat{\beta}\right) = \beta$;
(b) (Variance-covariance matrix) $\mathtt{Var}\left(\hat{\beta}|\mathbf{X}\right) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$;
(c) (Gauss-Markov Theorem) $\hat{\beta}$ *is the best linear unbiased estimator (BLUE) of* β*. That is, for any unbiased estimator* $\tilde{\beta}$ *that is linear in* $\mathbf{Y}$,

$$\mathtt{Var}(\tilde{\beta}|\mathbf{X}) \geq \mathtt{Var}(\hat{\beta}|\mathbf{X});$$

(d) (Unbiased estimator of variance) $E\left(s^2|\mathbf{X}\right) = \sigma^2$;
(e) (Orthogonality between $\hat{\beta}$ *and* $\hat{\varepsilon}$**)**
$\mathtt{Cov}\left(\hat{\beta}, \hat{\varepsilon}|\mathbf{X}\right) = E\left[\left(\hat{\beta} - \beta\right)\hat{\varepsilon}'|\mathbf{X}\right] = 0.$

Finite sample properties of OLS

Proof of (a) and (b)

(a) By Assumptions A.1 and A.3, we have

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{X}\beta+\varepsilon\right) = \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon.$$

By Assumption A.2, we have

$$E\left(\hat{\beta}|\mathbf{X}\right) = \beta + E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon|\mathbf{X}\right] = \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'E\left[\varepsilon|\mathbf{X}\right] = \beta.$$

By LIE, we have $E\left(\hat{\beta}\right) = \beta$.

(b) By Assumption A.4,

$$\begin{aligned} \mathtt{Var}\left(\hat{\beta}|\mathbf{X}\right) &= \mathtt{Var}\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon|\mathbf{X}\right) \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathtt{Var}\left(\varepsilon|\mathbf{X}\right)\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]' \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'I_n\sigma^2\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}. \end{aligned}$$

Finite sample properties of OLS

Proof of (c)

(c) Let $\tilde{\beta} = AY$ be a linear estimator of β where A is a $k \times n$ "weight" matrix, which may be constant or functions of $\mathbf{X}$. It is unbiased iff

$$E\left(\tilde{\beta}|\mathbf{X}\right) = E\left[AY|\mathbf{X}\right] = E\left[A\mathbf{X}\beta|\mathbf{X}\right] + E\left[A\varepsilon|\mathbf{X}\right] = A\mathbf{X}\beta = \beta$$

It follows that $A\mathbf{X} = I_k$. Then

$$\begin{aligned}
\mathtt{Var}\left(\tilde{\beta}|\mathbf{X}\right) &= A\mathtt{Var}\left(\varepsilon|\mathbf{X}\right)A' = \sigma^2 AA' \\
&= \sigma^2 A\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'A' + \sigma^2 AMA' \\
&= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} + \sigma^2 AM\left(AM\right)' \\
&\geq \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \mathtt{Var}\left(\hat{\beta}|\mathbf{X}\right).
\end{aligned}$$

Note that $A'MA = 0$ implies that $0 = AM = A\left(I_n - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)$ $= A - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$, which means that $A = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$, ie., $\tilde{\beta} = \hat{\beta}$. Thus suggests the **uniqueness** of the BLUE.

Finite sample properties of OLS

Proof of (d) and (e)

(d)

$$\begin{aligned} E\left(s^2|\mathbf{X}\right) &= \frac{1}{n-k}E\left(\hat{\varepsilon}'\hat{\varepsilon}|\mathbf{X}\right) = \frac{1}{n-k}E\left(\varepsilon' M\varepsilon|\mathbf{X}\right) \\ &= \frac{1}{n-k}E\left[\mathtt{tr}\left(\varepsilon' M\varepsilon\right)|\mathbf{X}\right] = \frac{1}{n-k}E\left[\mathtt{tr}\left(M\varepsilon\varepsilon'\right)|\mathbf{X}\right] \\ &= \frac{1}{n-k}\mathtt{tr}\left\{M\cdot E\left(\varepsilon\varepsilon'|\mathbf{X}\right)\right\} = \frac{1}{n-k}\mathtt{tr}\left(M\cdot\sigma^2 I_n\right) \\ &= \sigma^2\frac{1}{n-k}\mathtt{tr}\left(M\right) = \sigma^2. \end{aligned}$$

(e) Recall that $\hat{\beta} = \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon$ and $\hat{\varepsilon} = M\varepsilon$. We have

$$\begin{aligned} \mathtt{Cov}(\hat{\beta}, \hat{\varepsilon}|\mathbf{X}) &= E[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon\left(M\varepsilon\right)'|\mathbf{X}] \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'E\left(\varepsilon\varepsilon'|\mathbf{X}\right)M \\ &= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'M = 0. \end{aligned}$$

Finite sample properties of OLS

Remarks.

1. By (a) and (b), we have the conditional MSE of $\hat{\beta}$ given $\mathbf{X}$

$$\begin{aligned} MSE\left(\hat{\beta}|\mathbf{X}\right) &= E\left\{\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)'|\mathbf{X}\right\} \\ &= \mathtt{Var}\left(\hat{\beta}|\mathbf{X}\right)+\left[\mathtt{Bias}\left(\hat{\beta}|\mathbf{X}\right)\right]^2 \\ &= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}+0=\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}. \end{aligned}$$

   As $n\to\infty$, $\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\to 0$ under Assumption A.3* such that $MSE\left(\hat{\beta}|\mathbf{X}\right)\to 0$, which implies the *consistency* of $\hat{\beta}$ (We will see later).
2. The Gauss-Markov theorem **makes no assumption on the distribution of the error term** except that $E\left(\varepsilon_i|\mathbf{X}\right)=0$ and $\mathtt{Var}\left(\varepsilon_i|\mathbf{X}\right)=\sigma^2$.
3. Remember that it says noting about **nonlinear** estimators and compares only **linear unbiased** estimators. We can have biased or nonlinear estimators that have smaller MSE than the OLS estimators.

4. Sampling Distribution

Normality Assumption

To obtain the finite sample distribution of $\hat{\beta}$ we impose the following assumptions.

- **Assumption A.5 (Normality)** $\varepsilon | \mathbf{X} \sim N\left(0, \sigma^2 I_n\right)$.

Remarks.

1. Because the conditional pdf of ε given $\mathbf{X}$ is

$$f\left(\varepsilon | \mathbf{X}\right) = \left(2\pi\sigma^2\right)^{n/2} \exp\left(-\frac{\varepsilon'\varepsilon}{2}\right)$$

   which has nothing to do with $\mathbf{X}$, and then implies that $\varepsilon \perp \mathbf{X}$.

2. Assumption A.5 implies A.2 and A.4.

Normality Assumption

The following lemma is very useful and will be used frequently.

Lemma

(i) If $\varepsilon \sim N(0, \Sigma)$ where Σ is nonsingular, then $\varepsilon' \Sigma^{-1} \varepsilon \sim \chi^2(n)$;
(ii) If $\varepsilon \sim N(0, \sigma^2 I_n)$ and A is an $n \times n$ projection matrix, then $\varepsilon' A \varepsilon \sim \chi^2(\texttt{rank}(A))$ $(\texttt{rank}(A) = \texttt{tr}(A))$
(iii) If $\varepsilon \sim N(0, \sigma^2 I_n)$, A is an $n \times n$ projection matrix, and $A'B = 0$, then $\varepsilon' A \varepsilon \perp B' \varepsilon$ (Note that $A\varepsilon$ and $B'\varepsilon$ are independent).
(iv) If $\varepsilon \sim N(0, \sigma^2 I_n)$, A and B are both symmetric, then $\varepsilon' A \varepsilon \perp \varepsilon' B \varepsilon$ iff $AB = 0$.

In fact, if $\varepsilon \sim N(0, \sigma^2 I_n)$ and A is symmetric, then

$$\varepsilon' A \varepsilon / \sigma^2 \sim \chi^2(\texttt{rank}(A))$$

iff A is idempotent.

Sampling distribution

The following theorem states the sampling distribution of $\hat{\beta}$ and s^2.

Theorem

Suppose that Assumptions A.1, A.3 and A.5 hold. Then

(a) $\hat{\beta} - \beta | \mathbf{X} \sim N\left(0, \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$;

(b) $\frac{(n-k)s^2}{\sigma^2} | \mathbf{X} \sim \chi^2(n-k)$;

(c) $\hat{\beta} \perp s^2 | \mathbf{X}$.

Sampling distribution

Proof.

(a) $\hat{\beta}-\beta=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}\varepsilon=C\varepsilon=\sum_{i=1}^{n}C_i\varepsilon_i$ is a linear combination of ε_i's, where $C=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}$ is a $k\times n$ matrix and $C_i=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_i$. Conditional on $\mathbf{X}$, $\hat{\beta}-\beta$ is also normally distributed with mean $E\left(\hat{\beta}-\beta|\mathbf{X}\right)=\sum_{i=1}^{n}C_iE\left(\varepsilon_i|\mathbf{X}\right)=0$ and

$$\begin{aligned}\mathtt{Var}\left(\hat{\beta}-\beta|\mathbf{X}\right) &= C\mathtt{Var}\left(\varepsilon|\mathbf{X}\right)C'=\sigma^2CI_nC' \\ &= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}\mathbf{X}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}=\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}.\end{aligned}$$

Then $\hat{\beta}-\beta|\mathbf{X}\sim N\left(0,\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$;

(b) By (ii) in the previous lemma, we have

$$\frac{(n-k)\,s^2}{\sigma^2}|\mathbf{X}=\frac{\varepsilon'M\varepsilon}{\sigma^2}|\mathbf{X}\sim\chi^2\left(\mathtt{rank}(M)\right)=\chi^2\left(n-k\right).$$

(c) Note that $\hat{\beta}-\beta=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon(\equiv B'\varepsilon)$ and $s^2=\frac{1}{n-k}\varepsilon'M\varepsilon$. We have $MB=M\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}=0$ and then $\hat{\beta}\perp s^2|\mathbf{X}$.

5. Hypothesis Testing

Hypothesis testing

- Hypothesis testing is frequently needed when we conduct statistical inference in the regression framework. It can be used to evaluate the validity of economic theory, to detect absence of structure, among many other things.

Example (Production Function)

Given the Cobb-Douglas production function $Y = AK^{\beta_2}L^{\beta_3}$, we want to test

$$H_0 : \beta_2 + \beta_3 = 1 \text{ (constant return to scale)}$$

versus

$$H_1 : \beta_2 + \beta_3 < 1 \text{ (decreasing return to scale).}$$

To carry out the test, one can take log on both sides of the CD production function and add an error term

$$\ln(Y) = \beta_1 + \beta_2 \ln(K) + \beta_3 \ln(L) + \varepsilon.$$

Hypothesis testing

Examples (Structural Change)

Let GDP_i be the gross domestic product at year i. We are interested in whether there is a structural change in GDP around the year 2008. Define a dummy variable $D_i = 1(i \geq 2008)$ and consider the following regression model

$$\ln(GDP_i) = (\beta_1 + \beta_3 D_i) + (\beta_2 + \beta_4 D_i)\, i + \varepsilon_i .$$

The null hypothesis is $H_0 : \beta_3 = \beta_4 = 0$ (no structural change) versus $H_1 : \beta_3 \neq 0$ or $\beta_4 \neq 0$ (having structural change).

- Note that the first example has only restriction while the second example has two restrictions.
- We will discuss tests with a *single linear restriction* and *multiple linear restrictions* separately.

Single Linear Restriction: t-test

- For clarity, we assume Assumptions A.1-A.5 hold. The normality assumption A.5 is crucial in deriving the **exact distribution** of the t and F tests defined below.
- Consider testing a **single** linear restriction

$$H_0 : c'\beta = r \text{ vs } H_0 : c'\beta \neq r$$

where c is a $k \times 1$ vector and r is a scalar.
- Under A.5, $\hat{\beta}|\mathbf{X} \sim N\left(\beta, \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$ and

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

- It follows that under H_0

$$Z \equiv \frac{c'\hat{\beta} - r}{\sqrt{\sigma^2 c' \left(\mathbf{X}'\mathbf{X}\right)^{-1} c}} \sim N(0,1) \text{ conditional on } \mathbf{X}.$$

Single Linear Restriction: t-test

- Replacing σ^2 by its OLS estimators s^2, we get the feasible test statistic

$$T_n = \frac{c'\hat{\beta} - r}{\sqrt{s^2 c' \left(\mathbf{X}'\mathbf{X}\right)^{-1} c}}$$

Definition (Student t distribution)

A random variable T follows the student t distribution with q degrees of freedom, written as $T \sim t(q)$, if

$$T = \frac{U}{\sqrt{V/q}}$$

where $U \sim N\left(0,1\right)$, $V \sim \chi^2\left(q\right)$, and $U \perp V$.

Single Linear Restriction: t-test

Theorem (t-test)

Under Assumptions A.1-A.5 and H_0, $T_n \sim t(n-k)$.

Proof. Let $S_n = (n-k)\, s^2/\sigma^2$. Under H_0, $c'\beta = r$,

$$
\begin{aligned}
T_n &= \frac{c'\left(\hat{\beta}-\beta\right)}{\sqrt{s^2 c' \left(\mathbf{X}'\mathbf{X}\right)^{-1} c}} = \frac{c'\left(\hat{\beta}-\beta\right)/\sqrt{\sigma^2 c' \left(\mathbf{X}'\mathbf{X}\right)^{-1} c}}{\sqrt{\frac{(n-k)s^2/\sigma^2}{n-k}}} \\
&= \frac{Z_n}{\sqrt{S_n/\left(n-k\right)}}
\end{aligned}
$$

The result follows because $Z_n \sim N(0,1)$, $S_n \sim \chi^2(n-k)$, and $Z_n \perp S_n$. (see the sampling theory for OLS)

Single Linear Restriction: t-test

Remark.

- Suppose we are interested in testing $H_0 : \beta_j = \beta_{j0}$ versus $H_1 : \beta_j \neq \beta_{j0}$. The test statistic

$$T_n = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{s^2 \left[\left(\mathbf{X}'\mathbf{X} \right)^{-1} \right]_{jj}}} \sim t(n-k) \text{ under } H_0,$$

where $[A]_{jj}$ is the j-th diagonal element of A. Reject H_0 if $|T_n| > t_{\alpha/2}(n-k)$, where $t_{\alpha/2}(n-k)$ is the upper $\alpha/2$-percentile of $t(n-k)$ distribution. Let $\mathtt{SE}\left(\hat{\beta}_j\right) = \sqrt{s^2[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}$. A two-sided $1-\alpha$ confidence interval (CI) for β_j is given by

$$CI(\alpha) \equiv \left[T_n - t_{\alpha/2}(n-k)\,\mathtt{SE}\left(\hat{\beta}_j\right),\, T_n + t_{\alpha/2}(n-k)\,\mathtt{SE}\left(\hat{\beta}_j\right) \right].$$

By the duality between hypothesis tests and confidence intervals, we also reject the null at the significance level α if $\hat{\beta}_j \subsetneq CI(\alpha)$.

Single Linear Restriction: t-test

Remark.

- In modern econometrics, more attention has been given to the use of *p*value which is the **smallest significance level** at which we can reject the null hypothesis.
- Note that the p-value for a one-sided test is different from that for a two-sided test. For example, in the above two-sided test, if the statistic takes value t_n (a fixed number), then its p-value is defined by

$$p\text{-value} = 2P\left(t\left(n-k\right) > \left|t_n\right|\right)$$

  where $t\left(n-k\right)$ is the student t random variable with $n-k$ degrees of freedom. We reject the null if p-value $< \alpha$, the prescribed level of significance.
- In testing $H_0 : \beta_j = \beta_{j0}$ vs $H_1 : \beta_j > \beta_{j0}$, we can obtain the t statistic value t_n as above, but the p-value is defined as

$$p\text{-value} = P\left(t\left(n-k\right) > t_n\right).$$

Multiple Linear Restrictions: F-test

Consider testing the q linear restrictions on β

$$H_0 : R\beta = r \text{ vs } H_1 : R\beta = r,$$

where R is a known $q \times k$ matrix with $q < k$ and r is a known $q \times 1$ vector. We assume that $\mathtt{rank}(R) = q$.

Example

(a) $R = [1, 0, \ldots, 0]$, $r = 0$, $q = 1$. This is to test $H_0 : \beta_1 = 0$.
(b) $R = (0, I_{k-1})$, $r = \mathbf{0}_{k-1}$, $q = k - 1$. This is to test $H_0 : \beta_2 = \cdots = \beta_k = 0$. (all regressors are not significant)
(c) $R = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & 1 & \cdots & 0 \end{pmatrix}$, $r = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. This is to test $H_0 : \beta_1 = \beta_2$ and $\beta_2 + \beta_3 = 1$.

Multiple Linear Restrictions: F-test

Definition (F-distribution)

A random variable F follows the F distribution with (p, q) degrees of freedom, written as $F \sim F(p, q)$ if

$$F = \frac{U/p}{V/q}$$

where $U \sim \chi^2(p)$, $V \sim \chi^2(q)$, and $U \perp V$.

Theorem (F-test)

Suppose Assumptions A.1-A.5 hold. Then under H_0

$$F_n \equiv \frac{1}{q}\left(R\hat{\beta} - r\right)'\left[s^2 R\left(X'X\right)^{-1} R'\right]^{-1}\left(R\hat{\beta} - r\right) \sim F(q, n-k)$$

conditional on $\mathbf{X}$.

Multiple Linear Restrictions: F-test

Proof.

- Since $\hat{\beta}|\mathbf{X} \sim N\left(\beta, \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$, under H_0, $R\hat{\beta} - r|\mathbf{X} \sim N\left(0, \sigma^2 R \left(\mathbf{X}'\mathbf{X}\right)^{-1} R'\right)$. Then

$$A_n \equiv \left(R\hat{\beta} - r\right)' \left[\sigma^2 R \left(\mathbf{X}'\mathbf{X}\right)^{-1} R'\right]^{-1} \left(R\hat{\beta} - r\right) \sim \chi^2\left(q\right)$$

  conditional on $\mathbf{X}$ by noting that $R\left(\mathbf{X}'\mathbf{X}\right)^{-1} R'$ is a p.d. matrix with rank q.
- In addition, we have

$$S_n = \frac{\left(n-k\right) s^2}{\sigma^2} \sim \chi^2\left(n-k\right)$$

  conditional on $\mathbf{X}$. Note that $s^2 \perp \hat{\beta}|\mathbf{X}$, which implies that $S_n \perp A_n|\mathbf{X}$.
- We combine these results to get

$$F_n = \frac{A_n / q}{S_n / \left(n-k\right)} \sim F\left(q, n-k\right) \text{ conditional on } \mathbf{X}.$$

Multiple Linear Restrictions: F-test

Remarks on F test.

- The above theorem implies that $F_n \sim F(q, n-k)$ unconditionally under H_0.
- Suppose that we are still interested in testing $H_0 : \beta_j = \beta_{j0}$.
  - In this case, $q = 1, r = \beta_{j0}$, and $R = e_j'$, where e_j is a $k \times 1$ vector with 1 in its jth place and 0 elsewhere.
  - $R\hat{\beta} - r = \hat{\beta}_j - \beta_{j0}$. $R(\mathbf{X}'\mathbf{X})^{-1}R' = \left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{jj}$.
  - (Link with t test) The test statistic is given by

$$F_n = \left\{ \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{s^2 \left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{jj}}} \right\}^2 \sim F(1, n-k) \text{ under } H_0.$$

  - Note that the expression inside the curly bracket is just the t-statistic. The result is not surprising since $t(n-k)^2 = F(1, n-k)$.

Multiple Linear Restrictions: F-test

Remarks on F test

- (Decision rule) The F test is usually used to test for **multiple restrictions** and one rejects the null only when the F statistic takes **sufficiently large** value.
  - We reject the null if $F_n > F_\alpha(q, n-k)$, the upper α-percentile of the $F(q, n-k)$ distribution.
  - Alternatively, we reject the null at the prescribed α level of significance if

    $$p\text{-value} = P\left(F\left(q, n-k\right) > f_n\right) < \alpha,$$

    where f_n is the value of the F_n test statistic (a fixed number).

6. Constrained Least Squares

Constrained LS

Consider testing the linear restrictions

$$H_0 : R\beta = r \text{ vs } H_1 : R\beta = r.$$

We use RSS_{ur} to denote the **unrestricted sum of squared residuals** and RSS_r to denote the **restricted sum of squared residuals** under the restriction H_0.
The following theorem shows an *alternative expression* for the F test statistic.

Theorem

Suppose Assumptions A.1-A.5 hold. Then under H_0

$$F_n = \frac{\left(RSS_r - RSS_{ur}\right)/q}{RSS_{ur}/\left(n-k\right)} \sim F\left(q, n-k\right).$$

Constrained LS

Proof.

Consider the following minimization problem under the null restrictions:

$$\min_{\beta} (\mathbf{Y} - \mathbf{X}\beta)' (\mathbf{Y} - \mathbf{X}\beta) \text{ s.t. } R\beta = r.$$

The Lagrangian is given by

$$\mathcal{L}(\beta, \lambda) = (\mathbf{Y} - \mathbf{X}\beta)' (\mathbf{Y} - \mathbf{X}\beta) + \lambda' (R\beta - r)$$

where λ is a $q \times 1$ vector of Lagrangian multiplier. The solution, $(\tilde{\beta}, \tilde{\lambda})$ should satisfy the FOCs

$$\begin{aligned} \frac{\partial \mathcal{L}(\tilde{\beta}, \tilde{\lambda})}{\partial \beta} &= -2\mathbf{X}' (\mathbf{Y} - \mathbf{X}\tilde{\beta}) + R'\tilde{\lambda} = 0 \\ \frac{\partial \mathcal{L}(\tilde{\beta}, \tilde{\lambda})}{\partial \lambda} &= R\tilde{\beta} - r = 0 \end{aligned}$$

Constrained LS

Proof. (Cont.)
By the first FOC, $\tilde{\beta} = \hat{\beta} - \frac{1}{2}\left(\mathbf{X}'\mathbf{X}\right)^{-1} R'\tilde{\lambda}$. By the second FOC, $r = R\tilde{\beta} = R\hat{\beta} - \frac{1}{2}R\left(\mathbf{X}'\mathbf{X}\right)^{-1} R'\tilde{\lambda}$. Then

$$\tilde{\lambda} = 2\left[R\left(\mathbf{X}'\mathbf{X}\right)^{-1} R'\right]^{-1}\left(R\hat{\beta} - r\right).$$

It follows that $\tilde{\beta} = \hat{\beta} - \left(\mathbf{X}'\mathbf{X}\right)^{-1} R'\left[R\left(\mathbf{X}'\mathbf{X}\right)^{-1} R'\right]^{-1}\left(R\hat{\beta} - r\right)$.
Write $\tilde{\varepsilon} \equiv \mathbf{Y} - \mathbf{X}\tilde{\beta} = \hat{\varepsilon} + \mathbf{X}\left(\hat{\beta} - \tilde{\beta}\right)$ and

$$\begin{aligned} RSS_r &= \tilde{\varepsilon}'\tilde{\varepsilon} = \left[\hat{\varepsilon} + \mathbf{X}\left(\hat{\beta} - \tilde{\beta}\right)\right]'\left[\hat{\varepsilon} + \mathbf{X}\left(\hat{\beta} - \tilde{\beta}\right)\right] \\ &= \hat{\varepsilon}'\hat{\varepsilon} + \left(\hat{\beta} - \tilde{\beta}\right)'\mathbf{X}'\mathbf{X}\left(\hat{\beta} - \tilde{\beta}\right) \\ &= RSS_{ur} + \left(R\hat{\beta} - r\right)'\left[R\left(\mathbf{X}'\mathbf{X}\right)^{-1} R'\right]^{-1}\left(R\hat{\beta} - r\right). \end{aligned}$$

Therefore, $F_n = \frac{(RSS_r - RSS_{ur})/q}{RSS_{ur}/(n-k)} \sim F\left(q, n-k\right)$ by the definition of F_n.(p.53)

Example

Consider testing $H_0 : \beta_2 = \beta_3$ in the linear regression model

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

for $i = 1, \ldots, n$. The restricted model is

$$Y_i = \beta_1 + \beta_2 X_i^* + \varepsilon_i$$

where $X_i^* = X_{i2} + X_{i3}$. We can run the long and short regressions respectively, and obtain the RSS form each model to construct the F-test.

Example (The significance test of regression)

Consider testing

$$H_0 : \beta_2 = \cdots = \beta_k \text{ vs } H_1 : \beta_j \neq 0 \text{ for some } j \geq 2$$

in the regression model: $Y_i = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$.
The restricted model is $Y_i = \beta_1 + \varepsilon_i$, the restricted estimator is $\tilde{\beta}_1 = \bar{Y}$, and $RSS_r = \sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2 = TSS$.
RSS_{ur} can be obtained from the long regression. As remarked earlier on, R^2 (in the unrestricted model) is closely related to a F-test statistic to test for the above null. The F-statistic in this case is

$$\begin{aligned} F_n &= \frac{\left(RSS_r - RSS_{ur} \right) / \left(k - 1 \right)}{RSS_{ur} / \left(n - k \right)} \\ &= \frac{\left(1 - \frac{RSS_{ur}}{TSS} \right) / \left(k - 1 \right)}{\frac{RSS_{ur}}{TSS} / \left(n - k \right)} = \frac{R^2 / \left(k - 1 \right)}{\left(1 - R^2 \right) / \left(n - k \right)} \uparrow \text{ as } R^2 \uparrow . \end{aligned}$$

7. Generalized Least Squares

GLS

What may go wrong if Assumptions A.1-A.5 do not hold? Here we relax Assumption A.5 a little bit by imposing A.5*.

- **Assumption A.5* (Normality)** $\varepsilon|\mathbf{X} \sim N\left(0, \sigma^2 V\right)$ where $V = V\left(\mathbf{X}\right)$ is a **known** finite p.d. matrix.
- The above assumption means that $\mathtt{Var}(\varepsilon|\mathbf{X}) = \sigma^2 V$ is known up to a finite constant σ^2 and it allows for conditional heteroskedasticity of known form. Written explicitly, this assumption indicates that

$$
\begin{aligned}
E\left(\varepsilon_i|\mathbf{X}\right) &= \mathbf{0} \\
E\left(\varepsilon_i^2|\mathbf{X}\right) &= \sigma^2 V_{ii}\left(\mathbf{X}\right) \\
E\left(\varepsilon_i\varepsilon_j|\mathbf{X}\right) &= \sigma^2 V_{ij}\left(\mathbf{X}\right)
\end{aligned}
$$

where $V_{ij}\left(\mathbf{X}\right)$ denotes the (i,j)-th element of $V\left(\mathbf{X}\right)$.

GLS

Here are the main finite sample properties of the usual OLS.

Theorem

Suppose Assumptions A.1, A.3 and A.5 hold. Then*
(a) (Unbiasedness) $E\left(\hat{\beta}|\mathbf{X}\right) = \beta$ *and* $E\left(\hat{\beta}\right) = \beta$;
(b) (Variance-covariance matrix)
$\mathtt{Var}\left(\hat{\beta}|\mathbf{X}\right) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}' V \mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1}$;
(c) (Normality) $\hat{\beta} - \beta|\mathbf{X} \sim N\left(0, \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}' V \mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$;
(d) (Orthogonality) $\mathtt{Cov}\left(\hat{\beta}, \hat{\varepsilon}|\mathbf{X}\right) = 0$.

- (i) The proof is simple and thus omitted. (ii) The OLS estimator is still unbiased, but is not BLUE generally. (iii) The classical t and F tests are not valid any more because they are based on the incorrect variance-covariance estimator. One can estimate $\sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}' V \mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1}$ by $\breve{\sigma}^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}' V \mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1}$ with $\breve{\sigma}^2$ being a *consistent* estimator of σ^2.

GLS

How to obtain an efficient (BLUE) estimator?

- Recall that for any symmetric p.d. matrix V, we can write $V^{-1} = C'C$ where C is a nonsingular matrix.
- Pre-multiplying both sides of the following equation by C

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

gives that

$$C\mathbf{Y} = C\mathbf{X}\beta + C\varepsilon \text{ or } \mathbf{Y}^* = \mathbf{X}^*\beta + \varepsilon^*$$

where $\mathbf{A}^* = C\mathbf{A}$. Now, $\varepsilon^*|\mathbf{X} \sim N\left(0, \sigma^2 I_n\right)$ by construction under Assumption A.5* and the OLS based on the previous equation gives that

$$\begin{aligned} \hat{\beta}^* &= \left(\mathbf{X}^{*\prime}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*\prime}\mathbf{Y}^* = \left(\mathbf{X}'C'C\mathbf{X}\right)^{-1}\mathbf{X}'C'C\mathbf{Y} \\ &= \left(\mathbf{X}'V^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'V^{-1}\mathbf{Y} \equiv \hat{\beta}_{GLS}, \end{aligned}$$

which is called the *generalized least squares* (GLS) estimator of β.

GLS

Here are the properties of GLS estimator:

Theorem

Suppose Assumptions A.1, A.3 and A.5 hold. Then*
(a) (Unbiasedness) $E\left(\hat{\beta}_{GLS}|\mathbf{X}\right)=\beta$;
(b) (Variance-covariance matrix) $\texttt{Var}\left(\hat{\beta}_{GLS}|\mathbf{X}\right)=\sigma^2\left(\mathbf{X}'V^{-1}\mathbf{X}\right)^{-1}$;
(c) (Normality) $\hat{\beta}_{GLS}-\beta|\mathbf{X}\sim N\left(0,\sigma^2\left(\mathbf{X}'V^{-1}\mathbf{X}\right)^{-1}\right)$
(d) (Unbiasedness of s^{2*}*)* $E\left(s^{2*}|\mathbf{X}\right)=\sigma^2$.
*(e) (Orthogonality***)** $\texttt{Cov}\left(\hat{\beta}_{GLD},\hat{\varepsilon}^*|\mathbf{X}\right)=0$.

- The proof of is straightforward. **Please complete the proof by yourself.** (**Hints**: Note that $\hat{\beta}_{GLS}$ is the OLS estimator of β in the transformed model which satisfies Assumptions A.1, A.3, and A.5 with $\varepsilon^*|\mathbf{X}\sim N\left(0,\sigma^2 I_n\right)$)

GLS

Remarks.

- Classical t and F tests are applicable for inference procedure based on $\hat{\beta}_{GLS}$.
- But in practice, V is generally **unknown** so that $\hat{\beta}_{GLS}$ is usually **infeasible**.
- One has to estimate V in order to obtain a feasible GLS estimator of β.
- If we can estimate V consistently by $\hat{V}$, then we can use the feasible GLS (FGLS) estimate

$$\hat{\beta}_{FGLS} = \left(\mathbf{X}'\hat{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{V}^{-1}\mathbf{Y}.$$

Then one has to rely on the **large sample theory** for justification.

GLS

Remarks. (Cont.)

- Alternatively, we can continue to use $\hat{\beta}_{OLS}$ but obtain the correct variance-covariance formula

$$\texttt{Var}\left(\hat{\beta}|\mathbf{X}\right) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}' V \mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

  as well as a consistent estimator for it.
- In this case, the classical t and F tests cannot be used because they are based on an incorrect formula for $\texttt{Var}\left(\hat{\beta}|\mathbf{X}\right)$.
- Nevertheless, modified t and F tests (or Wald tests, to be introduced in the next section) are valid by using the correct estimator of $\texttt{Var}\left(\hat{\beta}|\mathbf{X}\right)$.
- Again, one has to rely on the large sample theory for justification.
- (for your Interest) Read "The HAC Emperor Has No Clothes: Parts 1-3" by Francis X. Diebold.