Lecture 3
Multiple Linear Regression: Inference

CHUNG-MING KUAN

Department of Finance & CRETA

National Taiwan University

March 3, 2022

Lecture Outline

1. Multiple Linear Regression: Inference
   - Distribution of the OLS Estimators
   - Testing A Single Hypothesis
   - Testing Multiple Hypotheses
   - Regressions with Dummy Variables
   - Digression: The Causal Analysis

Distribution of the OLS Estimators

To derive the distribution of the OLS estimators, the assumption below further imposes normality on data.

Classical Assumption III

The random variables y_i, $i = 1, \ldots, n$, are independently normally distributed with $\mathbb{E}(y_i) = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}$, where $x_{i1}, \ldots, x_{ik}$ are non-random, and $\text{var}(y_i) = \sigma_o^2$.

Remark: Classical Assumption III is equivalent to:

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{b}_o,\, \sigma_o^2 \boldsymbol{I}),$$

where $\boldsymbol{b}_o = (b_0 \ b_1 \ \ldots \ b_k)'$ and $\boldsymbol{X}$ is non-random. Writing $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b}_o + \boldsymbol{\varepsilon}$, we have $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0},\, \sigma_o^2 \boldsymbol{I})$.

Given that the OLS estimators $\hat{\beta}_j$ are linear in y, the result below is immediate.

Distributions of the OLS Estimators

Under Classical Assumption III, $\hat{\beta}_j$ are jointly normally distributed with

$$\hat{\beta}_j \sim \mathcal{N}\left(b_j,\ \sigma_o^2 \frac{1}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 (1 - R_j^2)}\right), \quad j = 1, \ldots, k,$$

where R_j^2 is the coefficient of determination of the regression of x_j on 1 and other regressors x_h, $h \neq j$.

Remark: A more complete result of the distribution of $\hat{\boldsymbol{\beta}}$ is:

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{b}_o,\ \sigma_o^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}),$$

where $\hat{\beta}_j$, $j = 0, 1, \ldots, k$, are pairwise correlated.

More specifically, the covariance matrix of $\hat{\boldsymbol{\beta}}$ is:

$$\begin{aligned}
&\sigma_o^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \begin{bmatrix}
\text{var}(\hat{\beta}_0) & \text{cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{cov}(\hat{\beta}_0, \hat{\beta}_2) & \cdots & \text{cov}(\hat{\beta}_0, \hat{\beta}_k) \\
\text{cov}(\hat{\beta}_1, \hat{\beta}_0) & \text{var}(\hat{\beta}_1) & \text{cov}(\hat{\beta}_1, \hat{\beta}_2) & \cdots & \text{cov}(\hat{\beta}_1, \hat{\beta}_k) \\
\text{cov}(\hat{\beta}_2, \hat{\beta}_0) & \text{cov}(\hat{\beta}_2, \hat{\beta}_1) & \text{var}(\hat{\beta}_2) & \cdots & \text{cov}(\hat{\beta}_2, \hat{\beta}_k) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\text{cov}(\hat{\beta}_k, \hat{\beta}_0) & \text{cov}(\hat{\beta}_k, \hat{\beta}_1) & \text{cov}(\hat{\beta}_k, \hat{\beta}_2) & \cdots & \text{var}(\hat{\beta}_k)
\end{bmatrix}.
\end{aligned}$$

Let m^{ij} denote the (i, j) th element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, $i, j = 1, \ldots, k+1$. The j th diagonal term of $\text{var}(\hat{\boldsymbol{\beta}})$ is $\text{var}(\hat{\beta}_{j-1}) = \sigma_o^2 \times m^{jj}$, and its square root is the standard deviation:

$$\text{sd}(\hat{\beta}_{j-1}) = \sigma_o \sqrt{m^{jj}}, \quad j = 1, \ldots, k+1.$$

The (i, j) th off-diagonal element is $\text{cov}(\hat{\beta}_{i-1}, \hat{\beta}_{j-1}) = \sigma_o^2 \times m^{ij}$.

Replacing σ_o^2 with $\hat{\sigma}^2$, we obtain the following variance estimators:

$$\widehat{\mathrm{var}(\hat{\boldsymbol{\beta}})} = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1},$$

which is unbiased for $\mathrm{var}(\hat{\boldsymbol{\beta}})$. The jth diagonal term of $\widehat{\mathrm{var}(\hat{\boldsymbol{\beta}})}$ is $\hat{\sigma}^2 \times m^{jj}$, and its square root is the standard error of $\hat{\beta}_{j-1}$:

$$\mathrm{se}(\hat{\beta}_{j-1}) = \hat{\sigma}\sqrt{m^{jj}}, \quad j = 1, \ldots, k+1.$$

For example,

$$\mathrm{se}(\hat{\beta}_{j-1}) = \hat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2(1 - R_j^2)}}, \quad j = 2, \ldots, k+1,$$

as shown in Lecture 2.

Distribution of $\hat{\sigma}^2$

Under Classical Assumption III, $(n-k-1)\hat{\sigma}^2/\sigma_o^2 \sim \chi^2(n-k-1)$.

Proof (Optional): Writing $\hat{\boldsymbol{u}} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y} = (\boldsymbol{I}_n - \boldsymbol{P})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_o)$, we have

$$\begin{aligned}(n-k-1)\hat{\sigma}^2/\sigma_o^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}/\sigma_o^2 &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_o)'(\boldsymbol{I}_n - \boldsymbol{P})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_o)/\sigma_o^2 \\ &= \boldsymbol{y}^{*\prime}(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y}^*,\end{aligned}$$

where we write $\boldsymbol{y}^* = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_o)/\sigma_o$. Note that $\boldsymbol{I}_n - \boldsymbol{P}$ is symmetric and hence can be orthogonally diagonalized as $\boldsymbol{C}'(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{C} = \boldsymbol{\Lambda}$, where $\boldsymbol{C}$ is an orthogonal matrix, and $\boldsymbol{\Lambda}$ is the diagonal matrix with the eigenvalues of $\boldsymbol{I}_n - \boldsymbol{P}$ on the main diagonal. As $\boldsymbol{C}\boldsymbol{C}' = \boldsymbol{I}$, we can write

$$\boldsymbol{y}^{*\prime}(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y}^* = \boldsymbol{y}^{*\prime}\boldsymbol{C}[\boldsymbol{C}'(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{C}]\boldsymbol{C}'\boldsymbol{y}^* = \eta'\boldsymbol{\Lambda}\eta,$$

where $\eta = \boldsymbol{C}'\boldsymbol{y}^*$.

Proof (Cont'd): Note by definition that the eigenvalue λ is such that $(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{v} = \lambda\boldsymbol{v}$, where $\boldsymbol{v}$ is its corresponding eigenvector. Given that $\boldsymbol{I}_n - \boldsymbol{P}$ is idempotent,

$$(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{v} = (\boldsymbol{I}_n - \boldsymbol{P})(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{v} = (\boldsymbol{I}_n - \boldsymbol{P})\lambda\boldsymbol{v} = \lambda^2\boldsymbol{v};$$

that is, λ^2 is also the eigenvalue for $\boldsymbol{v}$. This shows that $\lambda = \lambda^2$, so that λ must be either one or zero. It can be shown that there are $n-k-1$ eigenvalues equal to one (see next slide). We have

$$\boldsymbol{\eta}'\boldsymbol{\Lambda}\boldsymbol{\eta} = \boldsymbol{\eta}'\begin{bmatrix} \boldsymbol{I}_{n-k-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\boldsymbol{\eta} = \sum_{i=1}^{n-k-1} \eta_i^2.$$

As $\boldsymbol{y}^* \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $\boldsymbol{\eta} = \boldsymbol{C}'\boldsymbol{y}^* \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, and its elements η_i are independent $\mathcal{N}(0,1)$ variables. Consequently, η_i^2 are independent $\chi^2(1)$, and

$$\boldsymbol{\eta}'\boldsymbol{\Lambda}\boldsymbol{\eta} \sim \chi^2(n-k-1).$$

To verify that $\boldsymbol{\Lambda}$ has $n-k-1$ eigenvalues equal to one, note that trace($\boldsymbol{\Lambda}$) is the number of non-zero eigenvalues and also its rank. Using the fact that $\text{trace}(\boldsymbol{AB}) = \text{trace}(\boldsymbol{BA})$, we have

$$\begin{aligned}\text{trace}(\boldsymbol{\Lambda}) &= \text{trace}(\boldsymbol{C}'(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{C}) = \text{trace}(\boldsymbol{C}\boldsymbol{C}'(\boldsymbol{I}_n - \boldsymbol{P})) \\ &= \text{trace}(\boldsymbol{I}_n - \boldsymbol{P}),\end{aligned}$$

where $\text{trace}(\boldsymbol{I}_n) = n$, and

$$\begin{aligned}\text{trace}(\boldsymbol{P}) &= \text{trace}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = \text{trace}(\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) \\ &= \text{trace}(\boldsymbol{I}_{k+1}) = k+1.\end{aligned}$$

Consequently,

$$\text{trace}(\boldsymbol{\Lambda}) = \text{trace}(\boldsymbol{I}_n) - \text{trace}(\boldsymbol{P}) = n-k-1,$$

the number of eigenvalues equal to one in $\boldsymbol{\Lambda}$.

Efficiency of the OLS Estimators

With the normality assumption, the OLS estimators $\hat{\beta}_j$ are also the maximum likelihood estimators (MLEs) of b_j and hence the most efficient among all unbiased (not necessarily linear) estimators, because $\text{var}(\hat{\beta}_j)$ achieves the Cramér-Rao lower bound. It turns out that $\hat{\sigma}^2$, though not an MLE, is also the best unbiased estimator for σ_o^2.

Efficiency of the OLS Estimators

Under Classical Assumption III, $\hat{\beta}_j$, $j = 0, 1, \ldots, k$, and $\hat{\sigma}^2$ are the best unbiased estimators.

Remark: Compared with the Gauss-Markov Theorem, the OLS estimators are now the most efficient in a larger class of estimators (i.e., all unbiased estimators) when the data satisfy the normality condition.

Testing A Single Hypothesis with One Parameter

Consider the null hypothesis: $b_j = c$, where c is a given, hypothetical value. For example, we may test if $b_j = 0$ or $b_j = 1$. If the hypothesis is true, we would expect $\hat{\beta}_j$ to be "close" to c. It is then natural to construct a test statistic that compares $\hat{\beta}_j$ and c. The "closeness" between $\hat{\beta}_j$ and c is determined by the underlying distribution of $\hat{\beta}_j$.

Given Classical Assumption III, $\hat{\beta}_j \sim \mathcal{N}(b_j, \mathrm{var}(\hat{\beta}_j))$. Then under the null hypothesis,

$$(\hat{\beta}_j - c)/\mathrm{sd}(\hat{\beta}_j) = (\hat{\beta}_j - b_j)/\mathrm{sd}(\hat{\beta}_j) \sim \mathcal{N}(0, 1).$$

The left-hand side is not readily used as a test statistic because the standard deviation involves unknown σ_o.

Replacing $\mathrm{sd}(\hat{\beta}_j)$ with $\mathrm{se}(\hat{\beta}_j)$, we have the following statistic:

$$\frac{\hat{\beta}_j - c}{\mathrm{se}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - c}{\mathrm{sd}(\hat{\beta}_j)} \bigg/ \frac{\hat{\sigma}}{\sigma_o} = \frac{\hat{\beta}_j - c}{\mathrm{sd}(\hat{\beta}_j)} \bigg/ \sqrt{\frac{(n-k-1)\hat{\sigma}^2}{\sigma_o^2(n-k-1)}},$$

where the numerator is distributed as $\mathcal{N}(0,1)$. We have seen that

$$(n-k-1)\frac{\hat{\sigma}^2}{\sigma_o^2} \sim \chi^2(n-k-1).$$

The denominator above is thus the square root of $\chi^2(n-k-1)$ divided by its degrees of freedom $n-k-1$. It can also be shown that the numerator and denominator are independent under the normality assumption (proof omitted). It follows that the ratio above has a t distribution.

The ratio $(\hat{\beta}_j - c)/\text{se}(\hat{\beta}_j)$ is thus referred to as the *t* statistic.

Distribution of the *t* Statistic

Given Classical Assumption III, the t statistic is:

$$\frac{\hat{\beta}_j - c}{\text{se}(\hat{\beta}_j)} \sim t(n-k-1), \quad j = 0, 1, \ldots, k,$$

under the hypothesis $b_j = c$.

For the hypothesis $b_j = 0$, the t statistic $\hat{\beta}_j/\text{se}(\hat{\beta}_j)$ is known as the *t* ratio. Most econometric packages report the t ratios for all coefficient estimates and their p values.

Remarks

- A t test of $b_j = c$ is one-sided when the alternative hypothesis is $b_j < c$ (or $b_j > c$), or two-sided when the alternative is $b_j \neq c$.
- When discussing a test result, we must be specific about the significance level α. For example, we say a parameter estimate is significantly different from c at α level when the null hypothesis of $b_j = c$ is rejected using the critical value at α level.
- It is common to set $\alpha = 5\%$. For a larger α (say 10%), the critical values is smaller (in magnitude), and the test is more liberal (easier to reject); for a smaller α (say 1%), the test becomes more conservative (more difficult to reject).

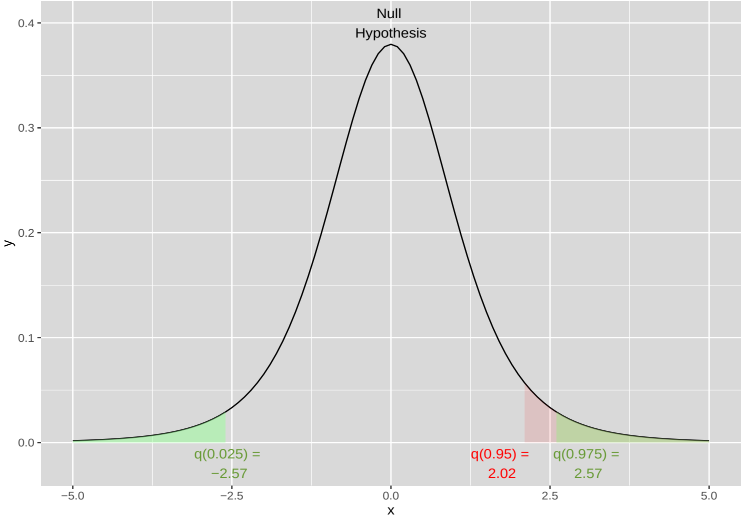

Figure: The null distribution $t(5)$ with critical values at 5% level

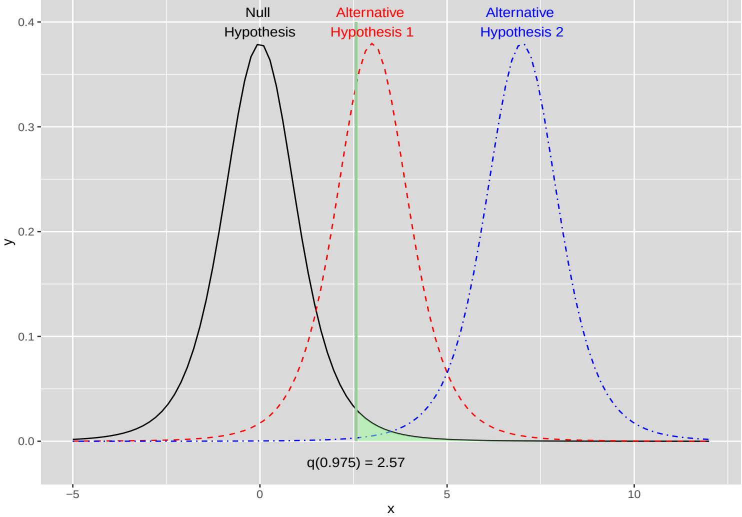

Figure: The null and two alternative distributions

Testing A Single Hypothesis with Several Parameters

Consider now the null hypothesis with two parameters: $b_2 + b_3 = c$; for example, $b_2 + b_3 = 1$. To construct a test statistic, it is natural to compare $\hat{\beta}_2 + \hat{\beta}_3$ with c. Clearly,

$$\hat{\beta}_2 + \hat{\beta}_3 \sim \mathcal{N}(b_2 + b_3,\ \text{var}(\hat{\beta}_2 + \hat{\beta}_3)),$$

where

$$\begin{aligned}\text{var}(\hat{\beta}_2 + \hat{\beta}_3) &= \text{var}(\hat{\beta}_2) + \text{var}(\hat{\beta}_3) + 2\,\text{cov}(\hat{\beta}_2, \hat{\beta}_3) \\ &= \sigma_o^2(m^{33} + m^{44} + 2m^{34}).\end{aligned}$$

Under the null hypothesis,

$$\frac{(\hat{\beta}_2 + \hat{\beta}_3) - c}{\text{sd}(\hat{\beta}_2 + \hat{\beta}_3)} = \frac{(\hat{\beta}_2 + \hat{\beta}_3) - (b_2 + b_3)}{\sigma_o\sqrt{m^{33} + m^{44} + 2m^{34}}} \sim \mathcal{N}(0, 1).$$

Replacing σ_o with $\hat{\sigma}$, we have the following t statistic:

$$\frac{(\hat{\beta}_2+\hat{\beta}_3)-c}{\textcolor{red}{\text{se}}(\hat{\beta}_2+\hat{\beta}_3)} = \frac{(\hat{\beta}_2+\hat{\beta}_3)-c}{\textcolor{red}{\hat{\sigma}}\sqrt{m^{33}+m^{44}+2m^{34}}} \sim t(n-k-1).$$

Similarly, consider the hypothesis: $2b_2 - b_3 = c$; for example $2b_2 - b_3 = 0$. Note that

$$2\hat{\beta}_2 - \hat{\beta}_3 \sim \mathcal{N}(2b_2 - b_3,\ \text{var}(2\hat{\beta}_2 - \hat{\beta}_3)).$$

where $\text{var}(2\hat{\beta}_2 - \hat{\beta}_3) = \sigma_o^2(4m^{33} + m^{44} - 4m^{34})$. Under the null hypothesis,

$$\frac{(2\hat{\beta}_2-\hat{\beta}_3)-c}{\text{sd}(2\hat{\beta}_2-\hat{\beta}_3)} = \frac{(2\hat{\beta}_2-\hat{\beta}_3)-(2b_2-b_3)}{\sigma_o\sqrt{4m^{33}+m^{44}-4m^{34}}} \sim \mathcal{N}(0,1),$$

and hence

$$\frac{(2\hat{\beta}_2-\hat{\beta}_3)-c}{\text{se}(2\hat{\beta}_2-\hat{\beta}_3)} = \frac{(2\hat{\beta}_2-\hat{\beta}_3)-c}{\hat{\sigma}\sqrt{4m^{33}+m^{44}-4m^{34}}} \sim t(n-k-1).$$

t Statistics in Matrix Notations

Consider the general linear hypothesis $\boldsymbol{R}\boldsymbol{b}_o = c$, where $\boldsymbol{R}$ is $1 \times (k+1)$; that is, there is only one hypothesis. For example,

$$\boldsymbol{R} = \begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 0 & \cdots & 0 \end{pmatrix}, \quad \boldsymbol{R}\boldsymbol{b}_o = b_2 + b_3 = c,$$

$$\boldsymbol{R} = \begin{pmatrix} 0 & 0 & 2 & -1 & 0 & 0 & \cdots & 0 \end{pmatrix}, \quad \boldsymbol{R}\boldsymbol{b}_o = 2b_2 - b_3 = c,$$

$$\boldsymbol{R} = \begin{pmatrix} 0 & 1 & 0 & -2 & 1 & 0 & \cdots & 0 \end{pmatrix}, \quad \boldsymbol{R}\boldsymbol{b}_o = b_1 - 2b_3 + b_4 = c.$$

To compare $\boldsymbol{R}\hat{\beta}$ with c, note that given Classical Assumption III,

$$\boldsymbol{R}\hat{\beta} \sim \mathcal{N}(\boldsymbol{R}\boldsymbol{b}_o,\ \boldsymbol{R}[\mathrm{var}(\hat{\beta})]\boldsymbol{R}'),$$

where $\boldsymbol{R}[\mathrm{var}(\hat{\beta})]\boldsymbol{R}'$ is a scalar. Then under the null hypothesis,

$$\frac{\boldsymbol{R}\hat{\beta} - c}{\sqrt{\boldsymbol{R}[\mathrm{var}(\hat{\beta})]\boldsymbol{R}'}} = \frac{\boldsymbol{R}\hat{\beta} - \boldsymbol{R}\boldsymbol{b}_o}{\sqrt{\boldsymbol{R}[\mathrm{var}(\hat{\beta})]\boldsymbol{R}'}} \sim \mathcal{N}(0,1).$$

Replacing $\text{var}(\hat{\beta}) = \sigma_o^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ with $\widehat{\text{var}(\hat{\beta})} = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, we have the following t statistic:

$$\frac{\boldsymbol{R}\hat{\beta} - c}{\hat{\sigma}\sqrt{\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'}} = \frac{\boldsymbol{R}\hat{\beta} - c}{\sigma_o\sqrt{\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'}} \Bigg/ \frac{\hat{\sigma}}{\sigma_o} \sim t(n-k-1).$$

Distribution of the t Statistic

Given Classical Assumption III, the t statistic is:

$$\frac{\boldsymbol{R}\hat{\beta} - c}{\hat{\sigma}\sqrt{\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'}} \sim t(n-k-1),$$

under the hypothesis $\boldsymbol{R}\boldsymbol{b}_o = c$, where $\boldsymbol{R}$ is $1 \times (k+1)$.

Testing Multiple Hypotheses

Suppose we would like to jointly test q hypotheses: $\boldsymbol{R}\boldsymbol{b}_o = \boldsymbol{c}$, where $\boldsymbol{R}$ is $q \times (k+1)$ with full row rank q, and $\boldsymbol{c}$ is a vector of q hypothetical values. For example, the joint hypotheses that $b_1 = 0$ and $b_2 = 0$ and $b_3 = 0$ can be expressed as:

$$\boldsymbol{R} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}, \quad \boldsymbol{R}\boldsymbol{b}_o = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

and the hypothesis that $b_1 = 1$ and $b_2 = b_3$ is:

$$\boldsymbol{R} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & -1 & 0 & \cdots & 0 \end{pmatrix}, \quad \boldsymbol{R}\boldsymbol{b}_o = \begin{pmatrix} b_1 \\ b_2 - b_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

F Statistic in Matrix Notations

Following previous discussion, we would like to construct a statistic that compares $\boldsymbol{R}\hat{\beta}$ with $\boldsymbol{c}$. Note that

$$\boldsymbol{R}\hat{\beta} \sim \mathcal{N}(\boldsymbol{R}\boldsymbol{b}_o,\ \boldsymbol{R}[\mathrm{var}(\hat{\beta})]\boldsymbol{R}'),$$

where $\boldsymbol{R}[\mathrm{var}(\hat{\beta})]\boldsymbol{R}'$ is $q \times q$. Then under the null hypothesis $\boldsymbol{R}\boldsymbol{b}_o = \boldsymbol{c}$,

$$\{\boldsymbol{R}[\mathrm{var}(\hat{\beta})]\boldsymbol{R}'\}^{-1/2}(\boldsymbol{R}\hat{\beta} - c) \sim \mathcal{N}(\mathbf{0},\ \boldsymbol{I}_q).$$

Taking inner product of the left-hand side, we have:

$$\begin{aligned}(\boldsymbol{R}\hat{\beta} - c)'\{\boldsymbol{R}[\mathrm{var}(\hat{\beta})]\boldsymbol{R}'\}^{-1/2}\{\boldsymbol{R}[\mathrm{var}(\hat{\beta})]\boldsymbol{R}'\}^{-1/2}(\boldsymbol{R}\hat{\beta} - c)& \\ = (\boldsymbol{R}\hat{\beta} - c)'\{\boldsymbol{R}[\mathrm{var}(\hat{\beta})]\boldsymbol{R}'\}^{-1}(\boldsymbol{R}\hat{\beta} - c)& \\ = (\boldsymbol{R}\hat{\beta} - c)'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\beta} - c)/\sigma_o^2,&\end{aligned}$$

which is the sum of q squared independent $\mathcal{N}(0,\ 1)$ variables and hence distributed as $\chi^2(q)$.

When the right-hand side is divided by q and σ_o^2 is replaced with $\hat{\sigma}^2$, the resulting statistic is:

$$\begin{aligned}
&\frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - c)'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - c)}{\hat{\sigma}^2 q} \\
&\qquad = \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - c)'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - c)}{\sigma_o^2 q} \bigg/ \frac{\hat{\sigma}^2}{\sigma_o^2} \\
&\qquad = \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - c)'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - c)}{\sigma_o^2 q} \bigg/ \frac{(n-k-1)\hat{\sigma}^2}{\sigma_o^2(n-k-1)}.
\end{aligned}$$

As shown earlier, the numerator term above is $\chi^2(q)$ divided by its degrees of freedom q, and the denominator is $\chi^2(n-k-1)$ divided by its degrees of freedom $n-k-1$. It can also be shown that these two terms are independent (proof omitted). Thus, this statistic has an F distribution under the null hypothesis and is known as the F statistic.

Distribution of the F Statistic

Given Classical Assumption III, the F statistic is:

$$\frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - c)'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - c)}{\hat{\sigma}^2 q} \sim F(q, n-k-1),$$

under the hypothesis $\boldsymbol{R}\boldsymbol{b}_o = c$, where $\boldsymbol{R}$ is $q \times (k+1)$ with full row rank q,

Note: When $q = 1$, the F statistic is

$$\frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - c)'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - c)}{\hat{\sigma}^2} \sim F(1, n-k-1),$$

where the left-hand side is nothing but the square of the t statistic, so that its F distribution is the square of $t(n-k-1)$, as it ought to be.

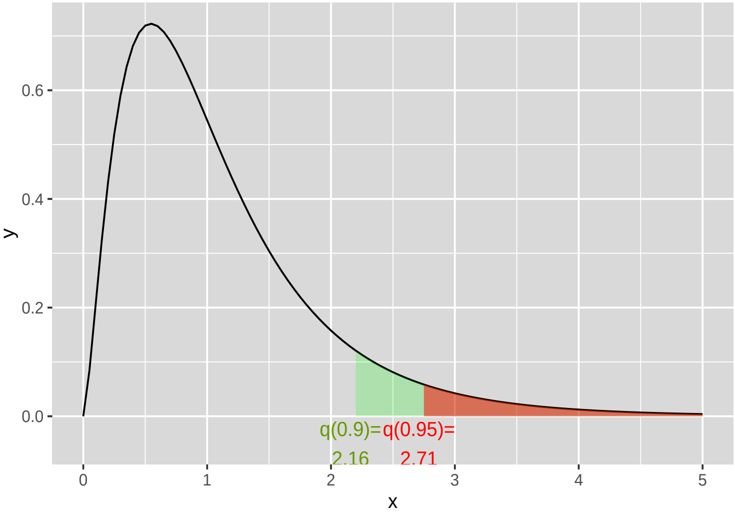

Figure: The null distribution ($F(5,20)$) with critical values at 5% & 10% level

Remarks

- Checking if all regressors (except the intercept) are useful in explaining y amounts to testing the hypothesis $b_1 = 0$ and $b_2 = 0 \ldots$ and $b_k = 0$. In this case, $q = k$, and the resulting F statistic is known as the regression F statistic and is distributed as $F(k, n-k-1)$. This statistic is also a standard output of most econometric packages.
- When the joint hypotheses of multiple restrictions is rejected by an F test, it suggests that there is at least one false restriction; as such, some of the restrictions may still be correct.
- Note that the inference of an F test of multiple restrictions does not necessarily agree with those of individual t tests. For example, when an F test does not reject the null hypothesis of 5 restrictions, it is possible that some of the t tests reject the corresponding restrictions.

Alternative Forms of F Statistic

Consider the joint hypotheses:

$$\boldsymbol{R}\boldsymbol{b}_o = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{R} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}.$$

Given Classical Assumption III,

$$y_i = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik} + u_i,$$

and under the null hypothesis,

$$y_i = b_0 + b_4 x_{i4} + b_5 x_{i5} + \cdots + b_k x_{ik} + u_i.$$

It turns out that the F statistic for this hypothesis can also be obtained by comparing the performance of the unrestricted regression of y on $1, x_1, x_2, \ldots, x_k$ and the restricted regression of y on $1, x_4, x_5, \ldots, x_k$.

It can be shown that the F statistic presented earlier is algebraically equivalent to:

$$\frac{(\mathrm{SSR}_r - \mathrm{SSR}_{ur})/3}{\mathrm{SSR}_{ur}/(n-k-1)} \sim F(3, n-k-1),$$

where SSR_r and SSR_{ur} are the residual sums of squares of the restricted and unrestricted regressions, respectively. This F statistic can also be expressed as:

$$\frac{(R_{ur}^2 - R_r^2)/3}{(1-R_{ur}^2)/(n-k-1)} \sim F(3, n-k-1),$$

where R_r^2 and R_{ur}^2 are the coefficients of determination of the restricted and unrestricted regressions, respectively. That is, the F statistic in effect compares the fitness of the restricted and unrestricted models.

Under the null (when the restrictions are correct), we expect these two SSRs (or R^2s) are close to each other and the statistic is small. Otherwise, the restricted regression must perform poorly (with much larger SSR and much smaller R^2), so that the statistic is large, leading to rejection of the null hypothesis.

More generally, when the null hypothesis imposes q restrictions, the resulting F statistic can also be computed as

$$\frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/(n-k-1)} \sim F(q, n-k-1),$$

where R_r^2 now is the coefficient of determination of the restricted regression obtained under q restrictions.

Example: Wage Regressions

The estimated wage model based on Taiwan's 2010 male data (11561 obs):

$$
\begin{array}{cccc}
3.8939 & +\ 0.0800\ \text{educ} & +\ 0.0166\ \text{exper} & \\
(0.0198) & (0.0012) & (0.0003) & \\
(197.05) & (65.41) & (50.45) & \\
\bar{R}^2 = 0.2893 & \hat{\sigma} = 0.3595 & \text{Reg } F = 2354 & \\
3.790 & +\ 0.0779\ \text{educ} & +\ 0.0365\ \text{exper} & -\ 0.0005\ \text{exper}^2 \\
(0.0199) & (0.0012) & (0.0009) & (0.00002) \\
(190.60) & (64.77) & (38.72) & (-22.47) \\
\bar{R}^2 = 0.319 & \hat{\sigma} = 0.3519 & \text{Reg } F = 1806 &
\end{array}
$$

The numbers in the first and second parentheses above are the standard error and t ratio of the OLS estimate, respectively. The regression F statistic suggests that some of these coefficients are significantly different from zero, even at 0.1% level.

An F Test for Model Misspecification: RESET

Given the linear model:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + u_i,$$

there may be neglected nonlinearity which may be captured using nonlinear functions of some regressors. For example, we consider in Lecture 2 a wage regression with 3 regressors: educ, exper, and exper^2. More generally, one may consider quadratic and cubic functions of regressors.

Ramsey (1969) considers the expanded model:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \delta_2 \hat{y}_i^2 + \delta_3 \hat{y}_i^3 + u_i,$$

where $\hat{y}_i$ denote the OLS fitted values of the base model, and uses $\hat{y}_i^2$ and $\hat{y}_i^3$ as proxies for the quadratic and cubic functions and cross products of regressors.

The RESET (REgression Specification Error Test) of Ramsey (1969) is an F test on the joint hypotheses $\delta_2 = \delta_3 = 0$. It has been shown that, under certain conditions, the RESET has the null distribution $F(2, n-k-3)$, where the second degrees of freedom is $n-k-1-2$ and different from other F tests . Rejecting these hypotheses suggests that the functional form of the base model is misspecified (e.g., missing some nonlinearity); otherwise, we will maintain the base model.

Regressions with Dummy Variables

Let D denote a binary variable taking values one or zero. When D is included as a regressor, it is also referred to as a dummy variable and may be used to classify observations into two groups. For example, let y_i denote the wage of the i th individual and x_i his/her working experience (in years). Also let D_i be the binary indicator: $D_i = 1$ if the i th individual has an MBA degree and $D_i = 0$ otherwise. Consider the following specification:

$$y_i = \alpha_0 + \alpha_1 D_i + \beta_0 x_i + u_i,$$

which puts together two regressions: MBA regression ($D_i = 1$) with intercept $\alpha_0 + \alpha_1$, and non-MBA regression ($D_i = 0$) with intercept α_0. Checking if an MBA degree makes a difference in the starting salary amounts to testing the null hypothesis: $\alpha_1 = 0$ (a t test will do).

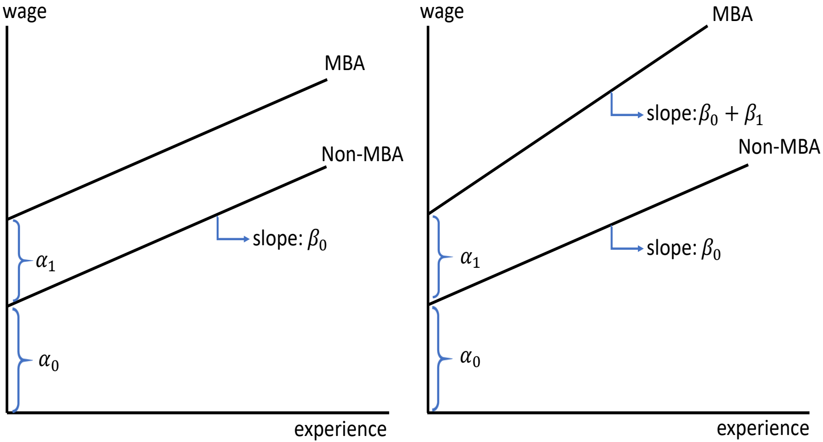

Figure: Regressions with a dummy variable

Consider the specification with the dummy variable D and its interaction with a regressor (say, experience):

$$y_i = \alpha_0 + \alpha_1 D_i + \beta_0 x_i + \beta_1 (x_i D_i) + u_i.$$

In this case, the MBA and non-MBA regressions may have different intercepts and different slopes ($\beta_0 + \beta_1$ and β_0). Checking if an MBA degree also makes experience a more important factor for salary amounts to testing the null hypothesis: $\beta_1 = 0$. Clearly, these two regressions would coincide if $\alpha_1 = 0$ and $\beta_1 = 0$. Thus, to check if an MBA degree has any effect on salary, we may apply the F test to this joint hypothesis .

Consider two dummy variables:

$D_{i1} = 1$ if i has only high school degree and $D_{i1} = 0$ otherwise;
$D_{i2} = 1$ if i has a college or graduate degree and $D_{i2} = 0$ otherwise. These two dummy variables in effect classify the data into 3 non-overlapping categories. Thus, the specification below puts together 3 regressions:

$$y_i = \alpha_0 + \alpha_1 D_{i1} + \alpha_2 D_{i2} + \beta x_i + u_i,$$

where the below-high-school regression (base model) has intercept α_0, the high-school regression has intercept $\alpha_0 + \alpha_1$, the college regression has intercept $\alpha_0 + \alpha_2$. Interesting hypotheses include: $\alpha_1 = 0$ (high school group is the same as the base model), $\alpha_2 = 0$ (college group is the same as the base model), $\alpha_1 = \alpha_2 = 0$ (high school and college groups are the same as the base model), and $\alpha_1 = \alpha_2$ (high school and college groups are the same).

Similar to the previous example, we may also consider a more general specification in which x interacts with D_1 and D_2:

$$y_t = \alpha_0 + \alpha_1 D_{i1} + \alpha_2 D_{i2} + \beta_0 x_i + \beta_1 (x_i D_{i1}) + \beta_2 (x_i D_{i2}) + u_i.$$

This results in regressions with possibly different intercepts and slopes. In addition to the hypotheses about α_1 and α_2 discussed earlier, there are now more interesting hypotheses that can be tested, such as $\beta_1 = 0$, $\beta_2 = 0$, $\beta_1 = \beta_2 = 0$, or $\beta_1 = \beta_2$.

Dummy variable trap: To avoid exact multicollinearity, the number of dummy variables in a regression with the constant term should be one less than the number of groups.

Digression: The Causal Analysis

A leading topic in empirical studies is to evaluate whether a government policy (business program, medical process) has any impact on the variable of interest. Such policy (program, process) is referred to as a treatment, and its effect is known as the causal effect or treatment effect.

Evaluating such effects may not be as straightforward as one would like. For example, suppose we are interested in knowing whether hospital care helps improve health condition, and we have the following survey results:

Group	Sample	Mean Status	Std. dev.
Hospital	7,774	3.21	0.014
No Hospital	90,049	3.93	0.003

where health status taking values 1 (poor) to 5 (excellent). Does the difference between these two sample averages provide an answer?

Let D be the binary variable such that $D = 1$ if the treatment is assigned and $D = 0$ if not. Also let Y_1 and Y_0 denote the potential outcomes (responses) when $D = 1$ and $D = 0$, respectively. As we can observe only one of Y_1 and Y_0, the observed outcome variable Y can be written as:

$$Y = Y_1 D + Y_0(1 - D) = Y_0 + D(Y_1 - Y_0),$$

where $Y_1 - Y_0$ represents the treatment effect of D. The individual i is in the treatment group if $D_i = 1$, and the observed outcome is $y_i = y_{1i}$; when i is in the control group if $D_i = 0$, and the observed outcome is $y_i = y_{0i}$. For each individual, only one outcome can be observed so that the individual treatment effect $(y_{1i} - y_{0i})$ is not identified. Yet, it is possible to identify the average treatment effect (ATE):

$$\mathbb{E}(Y_1 - Y_0).$$

The difference between the mean outcomes of the treatment and control groups may not correctly reveal the treatment effect, because

$$\begin{aligned}
&\mathbb{E}(Y|D=1) - \mathbb{E}(Y|D=0) \\
&\quad = \mathbb{E}(Y_1|D=1) - \mathbb{E}(Y_0|D=0) \\
&\quad = \underbrace{\mathbb{E}(Y_1|D=1) - \mathbb{E}(Y_0|D=1)}_{\text{treatment effect on the treated}} + \underbrace{\mathbb{E}(Y_0|D=1) - \mathbb{E}(Y_0|D=0)}_{\text{selection bias}},
\end{aligned}$$

where $\mathbb{E}(Y_0|D=1)$ is the mean outcome of those in the treatment group, had them not received the treatment. As we observe only Y_1 when $D=1$, $\mathbb{E}(Y_0|D=1)$ is counterfactual, and $\mathbb{E}(Y_1|D=1) - \mathbb{E}(Y_0|D=1)$ is the treatment effect based on the treatment group. Yet, $\mathbb{E}(Y_0|D=1) - \mathbb{E}(Y_0|D=0)$ is a bias due to selection into different groups. For example, those who actually receive hospital care $(D=1)$ are more likely to have worse health condition those who do not $(D=0)$, so that the selection bias is negative.

Randomized Treatment

It turns out that randomization eliminates the selection bias. When the treatment is randomly assigned to individuals, the outcomes (Y_1, Y_0) are independent of the treatment assignment D, denoted as $(Y_1, Y_0) \perp\!\!\!\perp D$. That is, conditioning on $D = 1$ and $D = 0$ does not affect the outcomes, so that $\mathbb{E}(Y_0|D = 1) = \mathbb{E}(Y_0|D = 0)$. Randomization ensures that the treatment and control groups are similar and interchangeable. We can then use $\mathbb{E}(Y_0|D = 0)$ to impute $\mathbb{E}(Y_0|D = 1)$. It follows that the selection bias vanishes and

$$\begin{aligned}\mathbb{E}(Y|D = 1) - \mathbb{E}(Y|D = 0) &= \mathbb{E}(Y_1|D = 1) - \mathbb{E}(Y_0|D = 0) \\ &= \mathbb{E}(Y_1) - \mathbb{E}(Y_0) = \text{ATE},\end{aligned}$$

where the 2nd equality follows from $(Y_1, Y_0) \perp\!\!\!\perp D$.

The mean group difference $\mathbb{E}(Y|D=1)-\mathbb{E}(Y|D=0)$ is readily estimated by the corresponding sample averages:

$$\widehat{\text{ATE}} = \frac{1}{n_1}\sum_{i=1}^{n} y_i D_i - \frac{1}{n-n_1}\sum_{i=1}^{n} y_i(1-D_i),$$

where $n_1 = \sum_{i=1}^{n} D_i$ is the number of observations in the treatment group, and $n-n_1$ is the number of observations in the control group. This is the estimated ATE under randomization.

Conditionally Independent Treatment

In many studies, treatment is not randomized. For example, it is unethical or illegal to randomly assign a new medicine to individuals. Instead, treatment assignment and the outcome variable may depend on some exogenous variables (a.k.a. confounders) $\boldsymbol{X}$. For example, individuals with higher education are more likely to participate a job training program and usually have higher salary. A treatment may be "as good as a randomized treatment" and independent of the outcomes when the confounders $\boldsymbol{X}$ are controlled for. This is the unconfoundedness (conditional independence) condition, denoted as $(Y_1, Y_0) \perp\!\!\!\perp D|\boldsymbol{X}$. Under this condition,

$$\begin{aligned}\mathbb{E}(Y|D=1,\boldsymbol{X}) - \mathbb{E}(Y|D=0,\boldsymbol{X}) & \\ &= \mathbb{E}(Y_1|D=1,\boldsymbol{X}) - \mathbb{E}(Y_0|D=0,\boldsymbol{X}) \\ &= \mathbb{E}(Y_1|\boldsymbol{X}) - \mathbb{E}(Y_0|\boldsymbol{X}) = \mathbb{E}(Y_1 - Y_0|\boldsymbol{X}).\end{aligned}$$

Integrating the conditional mean function above we obtain the ATE:

$$\mathbb{E}\big[\mathbb{E}(Y_1 - Y_0|\boldsymbol{X})\big] = \mathbb{E}(Y_1 - Y_0) = \text{ATE}.$$

Regression is now readily used to estimate the ATE. Let $\mu_1(\boldsymbol{X})$ and $\mu_0(\boldsymbol{X})$ be regression models for $\mathbb{E}(Y|D=1,\boldsymbol{X})$ and $\mathbb{E}(Y|D=0,\boldsymbol{X})$, respectively. With the data $(y_i, \boldsymbol{x}_i')'$, we may postulate μ_0 as a linear model:

$$y_{0i} = \mu_0(\boldsymbol{x}_i) + u_i = \alpha + \boldsymbol{x}_i'\boldsymbol{\beta} + u_i,$$

and assume that the individual treatment effect is a constant across the sample, i.e., $y_{1i} - y_{0i} = \gamma$. We then obtain the dummy variable model:

$$y_i = y_{0i} + D_i(y_{1i} - y_{0i}) = \alpha + \boldsymbol{x}_i'\boldsymbol{\beta} + D_i\gamma + u_i.$$

Estimating this model yields: $\hat{\mu}_1(\boldsymbol{x}_i) = \hat{\alpha} + \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + \hat{\gamma}$ and $\hat{\mu}_0(\boldsymbol{x}_i) = \hat{\alpha} + \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$.

Then, the ATE can now be estimated by the sample counterpart:

$$\widehat{\mathrm{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\big[\hat{\mu}_1(\boldsymbol{x}_i) - \hat{\mu}_0(\boldsymbol{x}_i)\big] = \hat{\gamma},$$

the estimated coefficient of γ.

Remark: The validity of the regression approach hinges on the following assumptions: (i) unconfoundedness, (ii) linearity, and (iii) constant individual effect. Note that the unconfoundedness condition would break down if D is endogenous; this may happen when D and (Y_1, Y_0) are both affected by some unobserved factors (hence cannot be included in the model). In this case, the OLS estimator $\hat{\gamma}$ is inconsistent and not reliable; see next lecture for more discussion of endogeneity and inconsistency.

Difference in Differences Design

In the return-to-schooling problem, there may be unobserved factors (such as individual's ability or IQ) that influence both schooling and earning. It is possible to control for such factors when data in the before- and after-treatment periods (e.g., panel data or repeated cross-section data) are available. This can be done by comparing the outcomes in these periods. Yet, such comparison may be problematic when data extend a long period of time, during which some events may arise and impact the outcomes. That is, the difference between the outcomes in the before- and after-treatment periods involves not only the treatment effect but also the time effect. The difference in differences (DID) design controls for the unobserved factor and the time effect and permits identification of the ATE. .

Let D denote the binary group indicator for group 0 ($D = 0$) and group 1 ($D = 1$), and t the binary time indicator for the before-treatment period ($t = 0$) and after-treatment period ($t = 1$). In the DID design, the group 1 is treated only in the after-treatment period, while the group 0 receives no treatment in both periods. Thus, there is a treatment when $(D, t) = (1, 1)$ and no treatment when $(D, t) = (1, 0), (0, 1), (0, 0)$.

The ATE in the DID design is identified by:

$$\begin{aligned}\text{ATE} = &\left[\mathbb{E}(Y|D = 1, t = 1) - \mathbb{E}(Y|D = 1, t = 0)\right] \\ &- \left[\mathbb{E}(Y|D = 0, t = 1) - \mathbb{E}(Y|D = 0, t = 0)\right],\end{aligned}$$

where the terms in square brackets are within-group comparisons. The first square brackets thus include the treatment and time effects, and the second square brackets contain the time effect only.

Let Y^A and Y^B denote the potential outcomes when $Dt = 1$ and $Dt = 0$, respectively; the observed outcome variable is:

$$Y = Y^A(Dt) + Y^B(1 - Dt) = Y^B + (Dt)(Y^A - Y^B).$$

For individual i at period t, we have: $y_{it} = y_{it}^B + (D_i t)(y_{it}^A - y_{it}^B)$. We may postulate a linear model for y_{it}^B:

$$y_{it}^B = \alpha + \beta t + \gamma D_i + \varphi_i + u_{it},$$

which involves the time trend t and the unobserved individual (fixed) effects φ_i that are time invariant. Assuming also constant treatment effect across the sample: $y_{it}^A - y_{it}^B = \delta$, we have the dummy variable model for the observed variable:

$$y_{it} = \alpha + \beta t + \gamma D_i + \varphi_i + \delta(D_i t) + u_{it}.$$

For i in group 0, $D_i = 0$ and $D_i t = 0$. The within-group comparison annihilates the individual effect φ_i, such that

$$y_{i1} - y_{i0} = \beta + (u_{i1} - u_{i0}).$$

For i in group 1, $D_i = 1$, $D_i t = t$, and the within-group comparison also removes the individual effect and yields:

$$y_{i1} - y_{i0} = \beta + \delta + (u_{i1} - u_{i0}).$$

Here, β characterizes the time effect in group 0, and $\beta + \delta$ is the total effect (time effect and treatment effect) in group 1. Their difference is the desired treatment effect δ. Estimating the dummy variable model with the time dummy (t), group dummy (D), and interaction of two dummies (Dt), we obtain the estimated coefficient $\hat{\delta}$ as the estimated ATE. The dummy variable model is readily extended to include confounders $\boldsymbol{X}$.

Kuan & Chen (2013): ATE of the NHI on Saving Rate

It has been found that Taiwan's household saving rate has declined since 1993. We are interested in knowing whether the National Health Insurance (NHI) launched in 1995 has any impact on household saving rate. Intuitively, households with the NHI face less risk of medical expenses and hence may reduce precautionary saving.

In this study, the treatment is NHI, and the outcome variable is saving rate $\ln Y - \ln C$, where Y is income and C is consumption. The data for the outcome variable and confounders are taken from the annual Survey of Family Income and Expenditure. Note that the NHI is expanded from Labor Insurance (1950), Government Employees' Insurance (GEI, 1958), and Farmers' Health Insurance (1985), and the coverage of NHI is the same as that of GEI. This makes the households with a member in the public sector a natural control group.

We adopt the DID design with 1990–1994 as the before-treatment period and 1996–2000 as the after-treatment period. We consider 3 designs.

- Control group: At least one of head and spouse in the public sector.
  Treatment group: neither head nor spouse in the public sector.
  The estimated ATE is -3.2%.
- Strict control group: Both head and spouse in the public sector.
  Treatment group: both head and spouse in the private sector.
  The estimated ATE is -4.3%.
- Strict control group: Both head and spouse in the public sector.
  Treatment group: one of head and spouse in the private sector (the other unemployed).
  The estimated ATE is -3.5%.
- In our sample, the average saving rate during 1990–1994 is 22.9% and that during 1996–2000 is 11.4%. Our results suggest that about one third of the decline may be attributed to the NHI.

Comparison

- Chou et al. (2003, 2004) take log saving $\ln(Y - C)$ as the outcome variable. As $\ln(Y - C)$ cannot be computed when there is negative saving, their study has to exclude 18.9% of sample data with negative saving. This results in a very small (or zero) estimated ATE in their study.
- The households with negative saving are typically those with low income or senior citizens. Our study found that the NHI has the largest impact on the oldest group (age 60–69) with the estimated ATE -11.9% and on the lowest 20% income group with the estimated ATE -5.0%.