Lecture 3
Multiple Linear Regression: Inference

CHUNG-MING KUAN

Department of Finance & CRETA

National Taiwan University

February 21, 2020

Lecture Outline

1. Mulple Linear Regression: Inference
   - Distribution of the OLS Estimators
   - Testing A Single Hypothesis
   - Testing Multiple Hypotheses
   - Regression with Dummy Variables

Distribution of the OLS Estimators

Classical Assumption III

The random variables y_i, $i = 1, \ldots, n$, follow the population model:

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

for some $b_0, b_1, \ldots, b_k$ (parameters of interest), where (i) $x_{i1}, \ldots, x_{ik}$ are non-random, and (ii) y_i are independently normally distributed with $\mathbb{E}(y_i) = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}$ and $\operatorname{var}(y_i) = \sigma_o^2$.

Remark: Assumption III(ii) is equivalent to:

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

where $\boldsymbol{b}_o = (b_0 \ b_1 \ \ldots \ b_k)'$; that is, y_i, $i = 1, \ldots, n$, are jointly normally distributed with a constant variance σ_o^2 and zero covariance.

We have learned that the OLS estimators $\hat{\beta}_j$ are linear in y. The following result 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:

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

Note that $\hat{\beta}_j$, $j = 0, 1, \ldots, k$, are correlated in general.

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

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

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 $\mathrm{var}(\hat{\boldsymbol{\beta}})$ is $\mathrm{var}(\hat{\beta}_{j-1}) = \sigma_o^2 \times m^{jj}$, and its square root is the standard deviation:

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

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

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

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

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

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

More specifically,

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

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): First observe that

$$\hat{\boldsymbol{u}} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y} = (\boldsymbol{I}_n - \boldsymbol{P})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_o).$$

Then,

$$(n-k-1)\hat{\sigma}^2/\sigma_o^2 = \hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}/\sigma_o^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_o)'(\boldsymbol{I}_n - \boldsymbol{P})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_o)/\sigma_o^2.$$

Letting $\boldsymbol{y}^* = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_o)/\sigma_o$, we have $\boldsymbol{y}^* \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, and

$$(n-k-1)\hat{\sigma}^2/\sigma_o^2 = \boldsymbol{y}^{*\prime}(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y}^*.$$

As a symmetric matrix, $\boldsymbol{I}_n - \boldsymbol{P}$ 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 a diagonal matrix with the eigenvalues of $\boldsymbol{I}_n - \boldsymbol{P}$ on the diagonal.

Proof (Cont'd): As $\boldsymbol{I}_n - \boldsymbol{P}$ is idempotent, its eigenvalues are either one or zero. Then, trace($\boldsymbol{\Lambda}$) is the number of non-zero eigenvalues and also its rank. Noting that trace($\boldsymbol{\Lambda}$) is

$$\operatorname{trace}(\boldsymbol{C}'(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{C}) = \operatorname{trace}(\boldsymbol{C}\boldsymbol{C}'(\boldsymbol{I}_n - \boldsymbol{P})) = \operatorname{trace}(\boldsymbol{I}_n - \boldsymbol{P}),$$

and for $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$,

$$\operatorname{trace}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = \operatorname{trace}(\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) = \operatorname{trace}(\boldsymbol{I}_{k+1}) = k+1,$$

we have $\operatorname{trace}(\boldsymbol{\Lambda}) = \operatorname{trace}(\boldsymbol{I}_n) - \operatorname{trace}(\boldsymbol{P}) = n - k - 1$. It follows that

$$\boldsymbol{y}^{*\prime}(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y}^* = \boldsymbol{y}^{*\prime}\boldsymbol{C}[\boldsymbol{C}'(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{C}]\boldsymbol{C}'\boldsymbol{y}^* = \boldsymbol{\eta}' \begin{bmatrix} \boldsymbol{I}_{n-k-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{\eta}.$$

where $\boldsymbol{\eta} = \boldsymbol{C}'\boldsymbol{y}^* \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. It follows that

$$\boldsymbol{y}^{*\prime}(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y}^* = \sum_{i=1}^{n-k-1} \eta_i^2 \sim \chi^2(n-k-1).$$

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 $\mathrm{var}(\hat{\beta}_j)$ 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(i) and (ii), $\hat{\beta}_j$, $j = 0, 1, \ldots, k$, 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 $\hat{\beta}_j$.

As $\hat{\beta}_j \sim \mathcal{N}(b_j, \mathrm{var}(\hat{\beta}_j))$, we have 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 σ_o which is unknown.

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

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

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

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

so that the denominator above is 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 their ratio has a t distribution; the statistic at the left-hand side is known as the t statistic.

Distribution of the t Statistic

Given Classical Assumption III, the t statistic is:

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

under the hypothesis $b_j = c$.

When the hypothesis is $b_j = 0$, the t statistic $\hat{\beta}_j / \mathrm{se}(\hat{\beta}_j)$ is also 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).

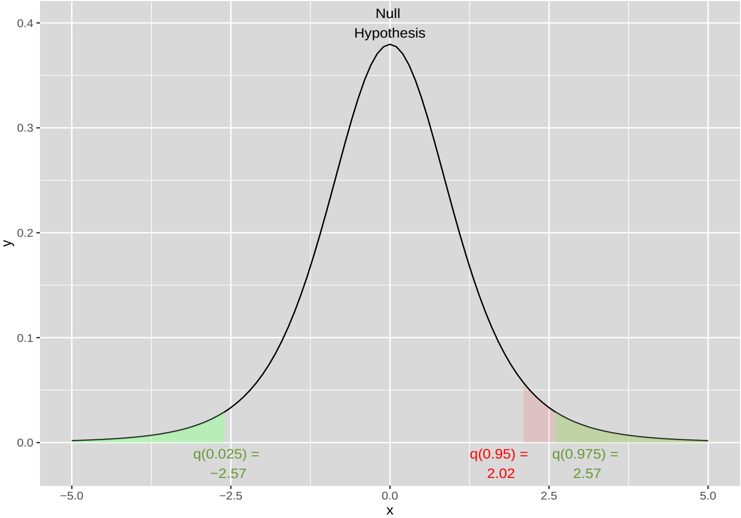

Figure: The null distribution $t(5)$ with critical values at 5% level

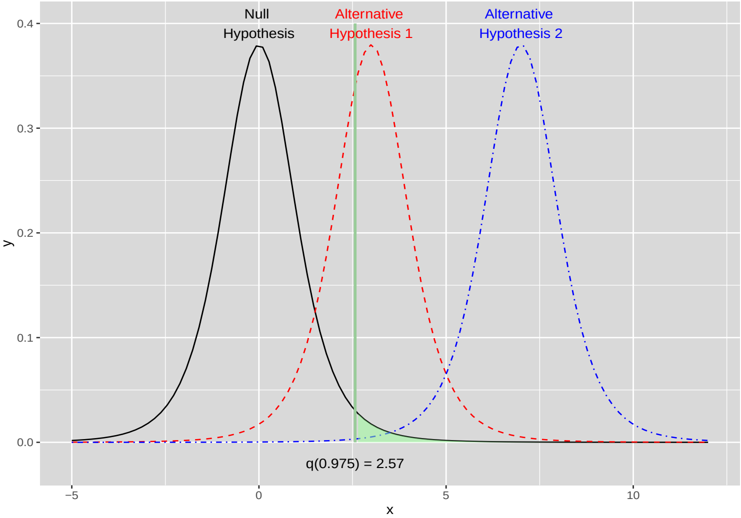

Figure: The null and two alternative distributions

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,

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

where

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

It follows that

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

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

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

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 $\text{var}(2\hat{\beta}_2-\hat{\beta}_3)=\sigma_o^2(4m^{33}+m^{44}-4m^{34})$. It follows that

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

and under the null hypothesis,

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

t Statistics in Matrix Notations

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

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

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 $\boldsymbol{R}\hat{\boldsymbol{\beta}}$ with $\boldsymbol{c}$. Note that

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

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

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

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

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

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

Replacing σ_o^2 with $\hat{\sigma}^2$ we have the following statistic:

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

where the numerator is $\chi^2(q)$ divided by its degrees of freedom, the denominator is $\chi^2(n-k-1)$ divided by its degrees of freedom, and these two terms are independent. The left-hand side thus 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:

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

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

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

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

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.

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; that is, 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.

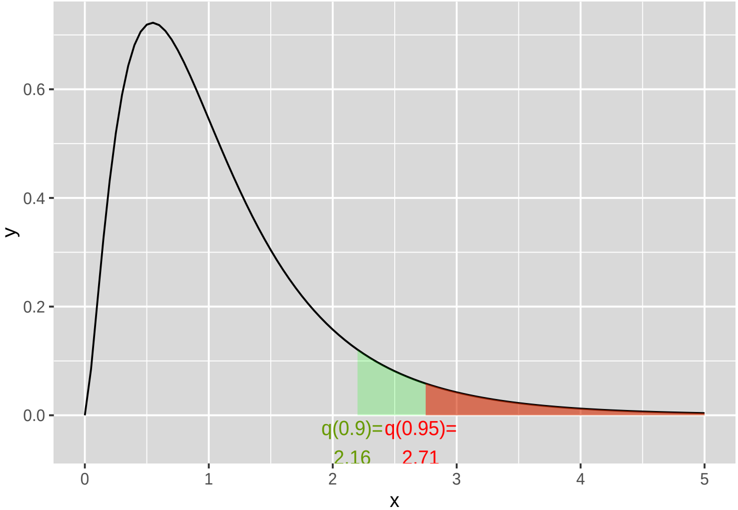

Figure: The null distribution ($F(5, 20)$) with critical values at 5% & 10% level

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 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 $1, x_4, x_5, \ldots, x_k$.

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. Equivalently, this F statistic is

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

where R_r^2 and R_{ur}^2 are the coefficients of determination of the restricted and unrestricted regressions, respectively. 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):

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

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 nonliearity 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.

Ramsey's RESET (REgression Specification Error Test) 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 (missing some nonlinearity); otherwise, we will maintain the base model.

Regression 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 different groups. For example, let y_i denote the wage of the i th individual and x_i the working experience (in years). Consider the following specification:

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

where $D_i = 1$ if i has an MBA degree and $D_i = 0$ otherwise. This specification 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).

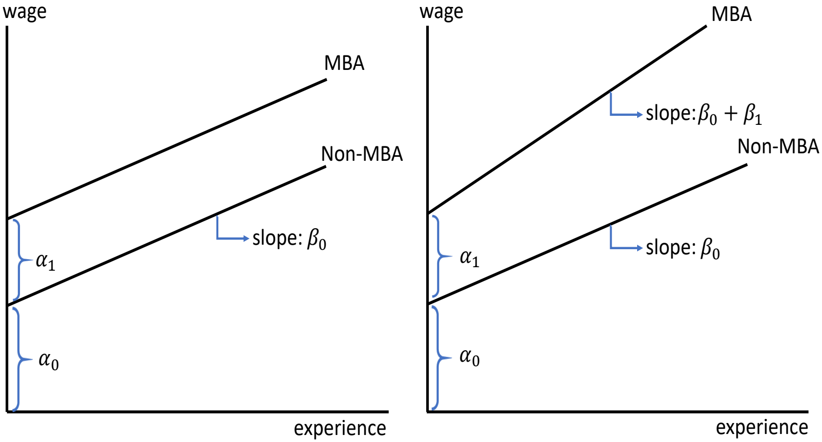

Figure: Regressions with a dummy variable

Consider the specification with a dummy variable 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 makes experience a more important factor in determining 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, we may check if an MBA degree affects salary by testing the joint hypotheses: $\alpha_1 = 0$ and $\beta_1 = 0$ (an F test will do).

These examples show that t and F tests may be applied to check interesting economic hypotheses.

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:

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

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:

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

The resulting regressions thus may have different intercept 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 $\beta_1 = 0$, $\beta_2 = 0$, $\beta_1 = \beta_2 = 0$, and $\beta_1 = \beta_2$.

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.

Including a dummy variable in a regression also allows us to evaluate whether a program (treatment) is effective. For example, suppose we are interested in knowing if participating a job training program can help improve the salary. Let y_i denote the log salary of individual i. We set $D_i = 1$ when i participates this training program, and $D_i = 0$ otherwise. The observations with $D_i = 1$ are the treatment group, and those with $D_i = 0$ are the control group. To evaluate the difference between these two groups, we may estimate the following specification:

$$y_i = \alpha_0 + \alpha_1 D_i + \boldsymbol{x}_i' \boldsymbol{\beta} + u_i,$$

where $\boldsymbol{x}_i$ is a vector of regressor observations. The coefficient α_1 characterizes the group difference and hence is understood as the average treatment effect (ATE) of this training program.