Advanced Econometrics

Lecture 8: Asymptotic Theory for Least Square (Hansen Chapters 7 and 9)

Instructor: Ma, Jun

Renmin University of China

November 29, 2021

Introduction

- The model is

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + e_i,\ i = 1, ..., n$$
$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{X}_iY_i\right).$$

Assumption

1. The obervations $(Y_i, \boldsymbol{X}_i), i = 1, \ldots n$, *are independent and identically distributed.*

2.$\mathbb{E}\left(Y^2\right) < \infty$.

3.$\mathbb{E}\left\|\boldsymbol{X}^2\right\| < \infty$.

4.$\boldsymbol{Q}_{\boldsymbol{XX}} = \mathbb{E}\left(\boldsymbol{XX}'\right)$ *is positive definite.*

Consistency of Least-Squares Estimator

- "$(Y_i, \boldsymbol{X}_i), i = 1, \ldots n$ are iid" implies that any function of $(Y_i, \boldsymbol{X}_i)$ is iid, including $\boldsymbol{X}_i \boldsymbol{X}_i'$ and $\boldsymbol{X}_i Y_i$.
- The LS estimator:

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{X}_i \boldsymbol{X}_i' \right) \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{X}_i Y_i \right) \right) = \hat{\boldsymbol{Q}}_{\boldsymbol{XX}}^{-1} \hat{\boldsymbol{Q}}_{\boldsymbol{X}Y}$$

$$\hat{\boldsymbol{Q}}_{\boldsymbol{XX}} = \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{X}_i \boldsymbol{X}_i' \right) \rightarrow_p \mathbb{E} \left(\boldsymbol{X}_i \boldsymbol{X}_i' \right) = \boldsymbol{Q}_{\boldsymbol{XX}}$$

$$\hat{\boldsymbol{Q}}_{\boldsymbol{X}Y} = \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{X}_i Y_i' \right) \rightarrow_p \mathbb{E} \left(\boldsymbol{X}_i Y_i \right) = \boldsymbol{Q}_{\boldsymbol{X}Y}.$$

- By Continuous Mapping Theorem,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \hat{\boldsymbol{Q}}_{\boldsymbol{XX}}^{-1} \hat{\boldsymbol{Q}}_{\boldsymbol{X}Y} \\ &\rightarrow_p \boldsymbol{Q}_{\boldsymbol{XX}}^{-1} \boldsymbol{Q}_{\boldsymbol{X}Y} \\ &= \boldsymbol{\beta}. \end{aligned}$$

- A different approach:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \hat{\boldsymbol{Q}}_{\boldsymbol{X}\boldsymbol{X}}^{-1} \hat{\boldsymbol{Q}}_{\boldsymbol{X}e}$$

$$\hat{\boldsymbol{Q}}_{\boldsymbol{X}e} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{X}_i e_i) .$$

- The WLLN:

$$\hat{\boldsymbol{Q}}_{\boldsymbol{X}e} \to_p \mathbb{E}(\boldsymbol{X}_i e_i) = 0.$$

- Therefore,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \hat{\boldsymbol{Q}}_{\boldsymbol{X}\boldsymbol{X}}^{-1} \hat{\boldsymbol{Q}}_{\boldsymbol{X}e} \to_p \boldsymbol{Q}_{\boldsymbol{X}\boldsymbol{X}}^{-1} \mathbf{0} = \mathbf{0}.$$

Theorem

Consistency of Least-Squares

$\hat{\boldsymbol{Q}}_{\boldsymbol{X}\boldsymbol{X}} \to_p \boldsymbol{Q}_{\boldsymbol{X}\boldsymbol{X}}$, $\hat{\boldsymbol{Q}}_{\boldsymbol{X}Y} \to_p \boldsymbol{Q}_{\boldsymbol{X}Y}$, $\hat{\boldsymbol{Q}}_{\boldsymbol{X}\boldsymbol{X}}^{-1} \to_p \boldsymbol{Q}_{\boldsymbol{X}\boldsymbol{X}}^{-1}$, $\hat{\boldsymbol{Q}}_{\boldsymbol{X}e} \to_p 0$, *and* $\hat{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}$.

Asymptotic Normality

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{X}_i\boldsymbol{X}_i'\right)\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\boldsymbol{X}_i e_i\right)\right)$$

- $\boldsymbol{X}_i e_i = \boldsymbol{X}_i\left(Y_i - \boldsymbol{X}_i'\boldsymbol{\beta}\right)$, $i = 1, \ldots, n$ are iid and mean zero ($\mathbb{E}\boldsymbol{X}_i e_i = \mathbf{0}$).
- The covariance matrix: $\boldsymbol{\Omega} = \mathbb{E}\left(e_i^2\boldsymbol{X}_i\boldsymbol{X}_i'\right)$:

$$\begin{aligned}\|\boldsymbol{\Omega}\| &\leq \mathbb{E}\left\|\boldsymbol{X}_i\boldsymbol{X}_i'e_i^2\right\| = \mathbb{E}\left(\|\boldsymbol{X}_i\|^2 e_i^2\right) \leq \mathbb{E}\left(\|\boldsymbol{X}_i\|^4\right)^{1/2}\left(\mathbb{E}\left(e_i^4\right)\right)^{1/2} \\ &< \infty.\end{aligned}$$

Theorem

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\boldsymbol{X}_i e_i) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Omega}\right).$$

Slutsky's theorem:

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) &\xrightarrow{d} \boldsymbol{Q}_{\boldsymbol{X}\boldsymbol{X}}^{-1} \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Omega}\right) \\ &= \mathrm{N}\left(\mathbf{0}, \boldsymbol{Q}_{\boldsymbol{X}\boldsymbol{X}}^{-1} \boldsymbol{\Omega} \boldsymbol{Q}_{\boldsymbol{X}\boldsymbol{X}}^{-1}\right).\end{aligned}$$

Theorem

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right)$$

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{Q}_{\boldsymbol{X}\boldsymbol{X}}^{-1} \boldsymbol{\Omega} \boldsymbol{Q}_{\boldsymbol{X}\boldsymbol{X}}^{-1},$$

$\boldsymbol{Q}_{\boldsymbol{X}\boldsymbol{X}} = \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i'\right)$, *and* $\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i' e_i^2\right)$.

- $\boldsymbol{V}_{\boldsymbol{\beta}}$ is often referred to as the **asymptotic covariance matrix** of $\hat{\boldsymbol{\beta}}$.
- Distributional approximation: when n is large,

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} \mathrm{N}\left(\boldsymbol{\beta}, \frac{\boldsymbol{V}_{\boldsymbol{\beta}}}{n}\right).$$

- The finite-sample conditional variance

$$\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} = \mathrm{Var}\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

  $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$ is the exact conditional variance of $\hat{\boldsymbol{\beta}}$.
- We should expect $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} \approx \frac{\boldsymbol{V}_{\boldsymbol{\beta}}}{n}$.

$$n\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} = \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

  and $n\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} \to_p \boldsymbol{V}_{\boldsymbol{\beta}}$.

Asymptotic Normality

- Under homoskedasticity, $\mathbb{E}\left(e_i^2 \mid \boldsymbol{X}_i\right) = \sigma^2 = \text{constant}$,

$$\boldsymbol{\Omega} = \mathbb{E}\mathbb{E}\left(e_i^2 \boldsymbol{X}_i \boldsymbol{X}_i' \mid \boldsymbol{X}_i\right) = \boldsymbol{Q}_{\boldsymbol{XX}}\sigma^2$$
$$\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{Q}_{\boldsymbol{XX}}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{\boldsymbol{XX}}^{-1} = \boldsymbol{Q}_{\boldsymbol{XX}}^{-1}\sigma^2.$$

- We define $\boldsymbol{V}_{\boldsymbol{\beta}}^0 = \boldsymbol{Q}_{\boldsymbol{XX}}^{-1}\sigma^2$ no matter $\mathbb{E}\left(e_i^2 \mid \boldsymbol{X}_i\right) = \sigma^2$ is true or false. When it is true, $\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{V}_{\boldsymbol{\beta}}^0$. $\boldsymbol{V}_{\boldsymbol{\beta}}^0$ is called the homoskedastic asymptotic covariance matrix.

Consistency of Error Variance Estimators

- Write the residual $\hat{e}_i$ as the error e_i plus a deviation term:

$$\begin{aligned}\hat{e}_i &= Y_i - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}} \\ &= e_i + \boldsymbol{X}_i'\boldsymbol{\beta} - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}} \\ &= e_i - \boldsymbol{X}_i'\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right).\end{aligned}$$

- Thus

$$\hat{e}_i^2 = e_i^2 - 2e_i\boldsymbol{X}_i'\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\boldsymbol{X}_i'\boldsymbol{X}_i\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right).$$

- The estimator $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n \hat{e}_i^2$ of $\sigma^2 = \mathbb{E}e_i^2$:

$$\begin{aligned}\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n e_i^2 &- 2\left(\frac{1}{n}\sum_{i=1}^n e_i\boldsymbol{X}_i'\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\ &+ \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i'\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right).\end{aligned}$$

- WLLN:

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 \rightarrow_p \sigma^2$$

$$\frac{1}{n}\sum_{i=1}^{n} e_i \boldsymbol{X}_i' \rightarrow_p \mathbb{E}\left(e_i^2 \boldsymbol{X}_i'\right) = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \rightarrow_p \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i'\right) = \boldsymbol{Q}_{\boldsymbol{XX}}.$$

- Another estimator $s^2 = (n-k)^{-1} \sum_{i=1}^{n} \hat{e}_i^2$. Since $n/(n-k) \rightarrow 1$ as $n \rightarrow \infty$,

$$s^2 = \left(\frac{n}{n-k}\right) \hat{\sigma}^2 \rightarrow_p \sigma^2.$$

Theorem
$\hat{\sigma}^2 \rightarrow_p \sigma^2$ and $s^2 \rightarrow_p \sigma^2$.

Homoskedastic Covariance Matrix Estimation

- For inference (confidence intervals and tests), we need a consistent estimate of $\boldsymbol{V}_{\boldsymbol{\beta}}$.
- Under homoskedasticity, $\boldsymbol{V}_{\boldsymbol{\beta}}$ simplifies to $\boldsymbol{V}_{\boldsymbol{\beta}}^0 = \boldsymbol{Q}_{\boldsymbol{XX}}^{-1}\sigma^2$.
- A natural estimator of $\boldsymbol{V}_{\boldsymbol{\beta}}^0 = \boldsymbol{Q}_{\boldsymbol{XX}}^{-1}\sigma^2$ is $\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^0 = \hat{\boldsymbol{Q}}_{\boldsymbol{XX}}^{-1}s^2$.
- By CMT,
$$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^0 = \hat{\boldsymbol{Q}}_{\boldsymbol{XX}}^{-1}s^2 \rightarrow_p \boldsymbol{Q}_{\boldsymbol{XX}}^{-1}\sigma^2 = \boldsymbol{V}_{\boldsymbol{\beta}}^0.$$
- $\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^0$ is consistent for $\boldsymbol{V}_{\boldsymbol{\beta}}^0$ regardless if the regression is homoskedastic or heteroskedastic.
- However, $\boldsymbol{V}_{\boldsymbol{\beta}}^0 = \boldsymbol{V}_{\boldsymbol{\beta}}$, the asymptotic covariance matrix, only under homoskedasticity.

Heteroskedastic Covariance Matrix Estimation

- A method of moments estimator for $\boldsymbol{\Omega}$:

$$\hat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \hat{e}_i^2.$$

- The White covariance matrix estimator

$$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{W} = \hat{\boldsymbol{Q}}_{\boldsymbol{X}\boldsymbol{X}}^{-1} \hat{\boldsymbol{\Omega}} \hat{\boldsymbol{Q}}_{\boldsymbol{X}\boldsymbol{X}}^{-1}.$$

- Observe

$$\begin{aligned}\hat{\boldsymbol{\Omega}} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \hat{e}_i^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' e_i^2 + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \left(\hat{e}_i^2 - e_i^2\right).\end{aligned}$$

- By WLLN,

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' e_i^2 \rightarrow_p \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i' e_i^2\right) = \boldsymbol{\Omega}.$$

- It remains to show

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \left(\hat{e}_i^2 - e_i^2\right) \to_p 0.$$

- Recall matrix norm: $\|\boldsymbol{A}\| = \operatorname{tr}\left(\boldsymbol{A}'\boldsymbol{A}\right)^{1/2}$ and therefore,

$$\left\|\boldsymbol{X}_i \boldsymbol{X}_i'\right\| = \operatorname{tr}\left(\boldsymbol{X}_i \boldsymbol{X}_i'\right)^{1/2} = \operatorname{tr}\left(\boldsymbol{X}_i' \boldsymbol{X}_i\right)^{1/2} = \left\|\boldsymbol{X}_i\right\|.$$

- Thus,

$$\begin{aligned} \left\|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \left(\hat{e}_i^2 - e_i^2\right)\right\| &\leq \frac{1}{n}\sum_{i=1}^{n} \left\|\boldsymbol{X}_i \boldsymbol{X}_i' \left(\hat{e}_i^2 - e_i^2\right)\right\| \\ &= \frac{1}{n}\sum_{i=1}^{n} \left\|\boldsymbol{X}_i\right\|^2 \left|\hat{e}_i^2 - e_i^2\right|. \end{aligned}$$

- By the triangle inequality and Cauchy-Schwarz inequality,

$$
\begin{aligned}
\left|\hat{e}_i^2 - e_i^2\right| &\leq 2\left|e_i \boldsymbol{X}_i' \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right| + \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \boldsymbol{X}_i' \boldsymbol{X}_i \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\
&= 2\left|e_i\right| \left|\boldsymbol{X}_i' \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right| + \left|\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \boldsymbol{X}_i\right|^2 \\
&\leq 2\left|e_i\right| \left\|\boldsymbol{X}_i\right\| \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\| + \left\|\boldsymbol{X}_i\right\|^2 \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|^2 .
\end{aligned}
$$

- Thus,

$$
\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \left(\hat{e}_i^2 - e_i^2\right)\right\| \leq{}& 2\left(\frac{1}{n}\sum_{i=1}^{n} \left\|\boldsymbol{X}_i\right\|^3 \left|e_i\right|\right) \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\| \\
&+ \left(\frac{1}{n}\sum_{i=1}^{n} \left\|\boldsymbol{X}_i\right\|^4\right) \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|^2 .
\end{aligned}
$$

Theorem
$\hat{\mathbf{\Omega}} \to_p \mathbf{\Omega}$ *and* $\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{W} \to_p \boldsymbol{V}_{\boldsymbol{\beta}}$.

Functions of Parameters

- The parameter of interest $\boldsymbol{\theta}$ is a function of the coefficients, $\boldsymbol{\theta} = \boldsymbol{r}(\boldsymbol{\beta})$ for some function $\boldsymbol{r} : \mathbb{R}^k \to \mathbb{R}^q$. The estimate of $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{r}\left(\hat{\boldsymbol{\beta}}\right).$$

Theorem
If $\boldsymbol{r}(\cdot)$ is continuous at the true value of $\boldsymbol{\beta}$, then $\hat{\boldsymbol{\theta}} \to_p \boldsymbol{\theta}$.

- By the Delta Method, $\hat{\boldsymbol{\theta}}$ is asymptotically normal.

Assumption
$\boldsymbol{r} : \mathbb{R}^k \to \mathbb{R}^q$ is continuously differentiable at the true value of $\boldsymbol{\beta}$ and $\boldsymbol{R} = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{r}(\boldsymbol{\beta})'$ has rank q.

Theorem

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\theta}}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}$$

- $\boldsymbol{r}$ can be linear: $\boldsymbol{r}\left(\boldsymbol{\beta}\right) = \boldsymbol{R}'\boldsymbol{\beta}$, for some $k \times q$ matrix $\boldsymbol{R}$.
- An even simpler case is when $\boldsymbol{R}$ is of the form $\boldsymbol{R} = \begin{pmatrix} \boldsymbol{I} \\ \mathbf{0} \end{pmatrix}$.
- Then we can partition $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2'\right)'$ so that $\boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{\beta}_1$. Then

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \begin{pmatrix} \boldsymbol{I} & \mathbf{0} \end{pmatrix} \boldsymbol{V}_{\boldsymbol{\beta}} \begin{pmatrix} \boldsymbol{I} \\ \mathbf{0} \end{pmatrix} = \boldsymbol{V}_{11},$$

where $\boldsymbol{V}_{\boldsymbol{\beta}}$ is partitioned: $\boldsymbol{V}_{\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{bmatrix}$.

- Take the example $\theta = \beta_j / \beta_l$ for $j \neq l$. Then

$$\boldsymbol{R} = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{r}(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial}{\partial \beta_1} (\beta_j/\beta_l) \\ \vdots \\ \frac{\partial}{\partial \beta_j} (\beta_j/\beta_l) \\ \vdots \\ \frac{\partial}{\partial \beta_l} (\beta_j/\beta_l) \\ \vdots \\ \frac{\partial}{\partial \beta_k} (\beta_j/\beta_l) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 1/\beta_l \\ \vdots \\ -\beta_j/\beta_l^2 \\ \vdots \\ 0 \end{pmatrix}.$$

- So

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \boldsymbol{V}_{jj}/\beta_l^2 + \boldsymbol{V}_{ll}\beta_j^2/\beta_l^4 - 2\boldsymbol{V}_{jl}\beta_j/\beta_l^3.$$

- For inference, we need an estimate of $\boldsymbol{V}_{\boldsymbol{\theta}} = \boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}$. The natural estimator of $\boldsymbol{R}$ is

$$\hat{\boldsymbol{R}} = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{r}\left(\hat{\boldsymbol{\beta}}\right)'.$$

- The estimate of $\boldsymbol{V}_{\boldsymbol{\theta}}$ is

$$\hat{\boldsymbol{V}}_{\boldsymbol{\theta}} = \hat{\boldsymbol{R}}'\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\hat{\boldsymbol{R}}.$$

Asymptotic Standard Errors

- A standard error is an estimate of the standard deviation of the distribution of an estimator.
- Since $\hat{\boldsymbol{\beta}} \overset{a}{\sim} \mathrm{N}\left(\boldsymbol{\beta}, \frac{\boldsymbol{V}_{\boldsymbol{\beta}}}{n}\right)$ and $\hat{\beta}_j \overset{a}{\sim} \mathrm{N}\left(\beta_j, \frac{[\boldsymbol{V}_{\boldsymbol{\beta}}]_{jj}}{n}\right)$, the standard error takes the form
$$s\left(\hat{\beta}_j\right) = \sqrt{\frac{\left[\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{W}\right]_{jj}}{n}}.$$
- Suppose the parameter of interest is $\theta = r(\boldsymbol{\beta})$ ($r : \mathbb{R}^k \to \mathbb{R}$, $q = 1$), the standard error for $\hat{\theta} = r\left(\hat{\boldsymbol{\beta}}\right)$ is
$$s\left(\hat{\theta}\right) = \sqrt{\frac{\hat{\boldsymbol{R}}'\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\hat{\boldsymbol{R}}}{n}}.$$

t-statistic

- $\theta = r(\boldsymbol{\beta})$ is the parameter of interest. Consider

$$T(\theta) = \frac{\hat{\theta} - \theta}{s(\hat{\theta})}.$$

- Since $\sqrt{n}\left(\hat{\theta} - \theta\right) \rightarrow_d \mathrm{N}(0, V_\theta)$ and $\hat{V}_\theta \rightarrow_p V_\theta$,

$$\begin{aligned} T(\theta) &= \frac{\hat{\theta} - \theta}{s(\hat{\theta})} \\ &= \frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\hat{V}_\theta}} \\ &\rightarrow_d \frac{N(0, V_\theta)}{\sqrt{V_\theta}} \\ &= Z \sim N(0, 1). \end{aligned}$$

- Since $T(\theta) \to_d Z$, CMT yields $|T(\theta)| \to_d |Z|$.
-

$$
\begin{aligned}
\Pr(|Z| \le u) &= \Pr(-u \le Z \le u) \\
&= \Pr(Z \le u) - \Pr(Z < -u) \\
&= \Phi(u) - \Phi(-u) \\
&= 2\Phi(u) - 1.
\end{aligned}
$$

Theorem

$T(\theta) \to_d Z \sim N(0,1)$ and $|T(\theta)| \to_d |Z|$.

Confidence Intervals

- A conventional confidence interval takes the form

$$\hat{C} = \left[\ \hat{\theta} - c \cdot s(\hat{\theta}), \quad \hat{\theta} + c \cdot s(\hat{\theta})\ \right],$$

where $c = F_{|Z|}^{-1}(1-\alpha)$ or $2\Phi(c) - 1 = 1 - \alpha$.

- Equivalently,

$$\hat{C} = \{\theta : \ | T(\theta) | \leq c\} = \left\{\theta : \ -c \leq \frac{\hat{\theta} - \theta}{s(\hat{\theta})} \leq c\right\}.$$

- The coverage probability:

$$\Pr\left(\theta \in \hat{C}\right) = \Pr\left(| T(\theta) | \leq c\right) \to \Pr\left(| Z | \leq c\right) = 1 - \alpha.$$

> **Theorem**
> *With* $c = \Phi^{-1}(1 - \alpha/2)$, $\Pr\left(\theta \in \hat{C}\right) \to 1 - \alpha$. *For* $c = 1.96$, $\Pr\left(\theta \in \hat{C}\right) \to 0.95$.

- Under homoskedasticity,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right) \to_d N\left(0, \sigma^2\left(\mathbb{E}\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)\right)^{-1}\right).$$

- We estimate the asymptotic variance by $s^2\left(n^{-1}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i'\right)^{-1}$.
- The confidence interval for β_j is given by

$$\left[\widehat{\beta}_j \pm z_{1-\alpha/2}\sqrt{\left[s^2\left(n^{-1}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i'\right)^{-1}\right]_{jj}/n}\right] = \left[\widehat{\beta}_j \pm z_{1-\alpha/2}\sqrt{\left[s^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{jj}}\right]$$

which is the same as the finite sample confidence interval.

Wald Statistic

- The parameter of interest is $\boldsymbol{\theta} = \boldsymbol{r}(\boldsymbol{\beta})$. $\boldsymbol{r} : \mathbb{R}^k \to \mathbb{R}^q$. Consider the Wald statistic

$$W(\boldsymbol{\theta}) = n\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \hat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right).$$

- Since

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \to_d \boldsymbol{Z} \sim N\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\theta}}\right)$$

and $\hat{\boldsymbol{V}}_{\boldsymbol{\theta}} \to_p \boldsymbol{V}_{\boldsymbol{\theta}}$,

$$W(\boldsymbol{\theta}) = n\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \hat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \to_d \boldsymbol{Z}'\boldsymbol{V}_{\boldsymbol{\theta}}^{-1}\boldsymbol{Z} \sim \chi_q^2.$$

Theorem

$$W(\boldsymbol{\theta}) \to_d \chi_q^2.$$

Confidence Regions

- A confidence region $\hat{C}$ is a set estimator for $\boldsymbol{\theta} \in \mathbb{R}^q$ when $q > 1$. Ideally, we hope $\Pr\left(\boldsymbol{\theta} \in \hat{C}\right) = 1 - \alpha$.
- A natural confidence region is

$$\hat{C} = \{\boldsymbol{\theta} : W(\boldsymbol{\theta}) \leq c_{1-\alpha}\},$$

  with $c_{1-\alpha}$ being the $1 - \alpha$ quantile of the χ_q^2 distribution: $F_{\chi_q^2}(c_{1-\alpha}) = 1 - \alpha$.
- Thus,

$$\Pr\left(\boldsymbol{\theta} \in \hat{C}\right) \to \Pr\left(\chi_q^2 \leq c_{1-\alpha}\right) = 1 - \alpha.$$

- Hypothesis tests attempt to assess whether there is evidence to contradict a proposed parametric restriction.
- Let $\boldsymbol{\theta} = \boldsymbol{r}(\boldsymbol{\beta})$ be a $q \times 1$ parameter of interest where $\boldsymbol{r} : \mathbb{R}^k \to \Theta \subset \mathbb{R}^q$ is some transformation.
- A point hypothesis concerning $\boldsymbol{\theta}$ is a proposed restriction such as $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ is a hypothesized (known) value.
- A hypothesis is a restriction $\boldsymbol{\beta} \in \boldsymbol{B}_0$. In the case of the hypothesis $\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$, $\boldsymbol{B}_0 = \{\boldsymbol{\beta} : \boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0\}$.

Definition
The null hypothesis, written $\mathbb{H}_0$, is the restriction $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ or $\boldsymbol{\beta} \in \boldsymbol{B}_0$.

- We often write the null hypothesis as $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ or $\mathbb{H}_0 : \boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$.

Definition
The alternative hypothesis, written $\mathbb{H}_1$, is the set $\{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0\}$ or $\{\boldsymbol{\beta} : \boldsymbol{\beta} \notin \boldsymbol{B}_0\}$

- We often write the alternative hypothesis as $\mathbb{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ or $\mathbb{H}_1 : \boldsymbol{r}(\boldsymbol{\beta}) \neq \boldsymbol{\theta}_0$.
- The goal of hypothesis testing is to assess whether or not $\mathbb{H}_0$ is true, by asking if $\mathbb{H}_0$ is consistent with the observed data.

Acceptance and Rejection

- The decision is based on a function of the data. It is convenient to express this function as a real-valued function called a test statistic
$$T = T\left((Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)\right).$$
- The hypothesis test then consists of the decision rule:
$$\text{Accept } \mathbb{H}_0 \text{ if } T \le c$$
$$\text{Reject } \mathbb{H}_0 \text{ if } T > c.$$
- Small values of T are likely when $\mathbb{H}_0$ is true and large values are likely when $\mathbb{H}_1$ is true.

Acceptance and Rejection

- The most commonly used test statistic is the absolute value of the t-statistic $T = |T(\theta_0)|$ where

$$T(\theta) = \frac{\widehat{\theta} - \theta}{s\left(\widehat{\theta}\right)}.$$

$\widehat{\theta}$ is a point estimate and $s\left(\widehat{\theta}\right)$ is its standard error.

Type I Error

- A false rejection of $\mathbb{H}_0$ (rejecting $\mathbb{H}_0$ when $\mathbb{H}_0$ is true) is called a Type-I error. The probability of a Type I error is

$$\Pr\left(\text{Reject } \mathbb{H}_0 \mid \mathbb{H}_0 \text{ is true}\right) = \Pr\left(T > c \mid \mathbb{H}_0 \text{ is true}\right).$$

- The first goal is to control the type-I error: it should not be large.
- In typical econometric models the exact sampling distributions of estimators and test statistics are unknown.

- Suppose that when $\mathbb{H}_0$ is true,

$$T \rightarrow_d \xi.$$

Let $G(u) = \Pr(\xi \leq u)$ be the distribution of ξ. We call G the asymptotic null distribution. In simple cases, G is known and does not depend on unknown parameters.

- We define the asymptotic size of the test as the asymptotic probability of a Type I error:

$$\begin{aligned} \lim_{n\rightarrow\infty} \Pr(T > c \mid \mathbb{H}_0 \text{ is true}) &= \Pr(\xi > c) \\ &= 1 - G(c). \end{aligned}$$

- In the dominant approach to hypothesis testing, the researcher pre-selects a significance level $\alpha \in (0, 1)$ and then selects c so that the asymptotic size is no larger than α.

t tests

- The most common test of "scalar" hypothesis: $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta \neq \theta_0$.

Theorem

Under $\mathbb{H}_0 : \theta = \theta_0$,

$$T(\theta_0) \to_d Z.$$

For c *satisfying* $\alpha = 2(1 - \Phi(c))$,

$$\Pr(|T(\theta_0)| > c \mid \mathbb{H}_0 \textit{ is true}) \to \alpha,$$

and the test "Reject $\mathbb{H}_0$ *if* $|T(\theta_0)| > c$*" has asymptotic size* α.

- The alternative $\theta \neq \theta_0$ is called a two-sided alternative.

- One-sided alternative could be $\mathbb{H}_1 : \theta > \theta_0$.
- Tests of $\theta = \theta_0$ against $\theta > \theta_0$ are based on the signed t-statistic $T = T(\theta_0)$.
- We reject $\mathbb{H}_0$ if $T > c$ where c satisfies $\alpha = 1 - \Phi(c)$. Negative values of are not taken as evidence against $\mathbb{H}_0$.
- We should use one-sided tests and critical values only when the parameter space is known to satisfy a one-sided restriction such as $\theta \geq \theta_0$.

Type II Error and Power

- A false acceptance of the null hypothesis $\mathbb{H}_0$ (accepting $\mathbb{H}_0$ when $\mathbb{H}_1$ is true) is called a Type II error.
- The rejection probability under the alternative hypothesis is called the power of the test.
- Power = 1 - the probability of a Type II error:

$$\pi(\boldsymbol{\theta}) = \Pr\left(\text{Reject } \mathbb{H}_0 \mid \mathbb{H}_1 \text{ is true}\right) = \Pr\left(T > c \mid \mathbb{H}_1 \text{ is true}\right)$$

  $\pi(\boldsymbol{\theta})$ is called power function. The power depends on the true value of the parameter $\boldsymbol{\theta}$.
- A well behaved test the power is increasing both as $\boldsymbol{\theta}$ moves away from $\boldsymbol{\theta}_0$ and as the sample size n increases.

- Four possibilities:

		Truth	
		H_0	H_1
Decision	H_0	✓	Type II error
	H_1	Type I error	✓

- When $T \leq c$, we accept H_0 (and risk making a Type II error).
- When $T > c$, we reject H_0 (and risk making a Type I error).

- Unfortunately, the probabilities of Type I and II errors are inversely related.
- By decreasing the probability of Type I error, one makes c larger, which increases the probability of the Type II error. Thus it is impossible to make both errors arbitrary small.
- We want the probability of a type-II error to be as small as possible for a given probability of a type-I error.

p-Values

- p-value is a measure of the strength of information against the null hypothesis:
$$p = 1 - G(T).$$
G is the (asymptotic) distribution of T under $\mathbb{H}_0$.
- p-value is the marginal significant level: the largest value of α for which the test rejects $\mathbb{H}_0$.
- $T \to_d \xi$ under $\mathbb{H}_0$, then $p = 1 - G(T) \to_d 1 - G(\xi)$:
$$\begin{aligned}
\Pr(1 - G(\xi) \le u) &= \Pr(1 - u \le G(\xi)) \\
&= 1 - \Pr\left(\xi \le G^{-1}(1-u)\right) \\
&= 1 - G\left(G^{-1}(1-u)\right) \\
&= 1 - (1-u) \\
&= u.
\end{aligned}$$

Wald Tests

- The parameter of interest is $\boldsymbol{\theta} = \boldsymbol{r}(\boldsymbol{\beta})$. Estimator: $\widehat{\boldsymbol{\theta}} = \boldsymbol{r}\left(\widehat{\boldsymbol{\beta}}\right)$. To test $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $\mathbb{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, one approach is to measure the discrepancy $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$:

$$W = n\left(\boldsymbol{r}\left(\widehat{\boldsymbol{\beta}}\right) - \boldsymbol{\theta}_0\right)'\left(\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}\widehat{\boldsymbol{R}}\right)^{-1}\left(\boldsymbol{r}\left(\widehat{\boldsymbol{\beta}}\right) - \boldsymbol{\theta}_0\right).$$

- When $\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{R}'\boldsymbol{\beta}$,

$$W = \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)'\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right).$$

Theorem

Under $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$*,*
then

$$W \to_d \chi^2_q,$$

and for c satisfying $\alpha = 1 - G_q(c)$*,*

$$\Pr\left(W > c \mid \mathbb{H}_0 \textit{ is true}\right) \to \alpha$$

so the test "Reject $\mathbb{H}_0$ *if* $W > c$*" has asymptotic size* α*.*

Homoskedastic Wald Tests

- If the error is known to be homoskedastic,

$$W^0 = \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \left(\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\theta}}}\right)^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$$
$$= \left(\boldsymbol{r}\left(\widehat{\boldsymbol{\beta}}\right) - \boldsymbol{\theta}_0\right)' \left(\widehat{\boldsymbol{R}}' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \widehat{\boldsymbol{R}}\right)^{-1} \left(\boldsymbol{r}\left(\widehat{\boldsymbol{\beta}}\right) - \boldsymbol{\theta}_0\right) / s^2.$$

- In the case of linear hypotheses $\mathbb{H}_0 : \boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$,

$$W^0 = \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)' \left(\boldsymbol{R}' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{R}\right)^{-1} \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right) / s^2.$$

- In this case, the F testing statistic: $F = W^0/q$ and $F \rightarrow_d \chi^2_q/q$.

Power and Test Consistency

- The power of a test is the probability of rejecting $\mathbb{H}_0$ when $\mathbb{H}_1$ is true.
- Random sample from $N\left(\theta, \sigma^2\right)$, with σ^2 known: $\{Y_1, ..., Y_n\}$. For testing $\mathbb{H}_0 : \theta = 0$ against $\mathbb{H}_1 : \theta > 0$,

$$T = \frac{\sqrt{n}\overline{Y}}{\sigma}.$$

  We reject $\mathbb{H}_0$ if $T > c$.
- Note $T = \frac{\sqrt{n}\left(\overline{Y}-\theta\right)}{\sigma} + \frac{\sqrt{n}\theta}{\sigma}$. The power of the test is

$$\Pr(T > c) = \Pr\left(Z + \sqrt{n}\theta/\sigma > c\right) = 1 - \Phi\left(c - \sqrt{n}\theta/\sigma\right).$$

- This power function is monotonically increasing in θ and n.
- If $\theta > 0$, the power increases to 1 as $n \to \infty$. This means whenever $\mathbb{H}_1$ is true, the test will reject $\mathbb{H}_0$ with a high probability if n is sufficiently large.

Definition

A test of $\mathbb{H}_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ is consistent against fixed alternatives if for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$, $\Pr\left(\text{Reject } \mathbb{H}_0 \mid \boldsymbol{\theta} \text{ is the true parameter}\right) \to 1$ as $n \to \infty$.

- In general, t test and Wald test are consistent. Take a t statistic for testing $\mathbb{H}_0 : \theta = \theta_0$,

$$T = \frac{\widehat{\theta} - \theta_0}{s\left(\widehat{\theta}\right)} = \frac{\widehat{\theta} - \theta}{s\left(\widehat{\theta}\right)} + \frac{\sqrt{n}\left(\theta - \theta_0\right)}{\sqrt{\widehat{V}_\theta}}.$$

- $\frac{\widehat{\theta}-\theta}{s\left(\widehat{\theta}\right)}$ converges in distribution to $N\left(0, 1\right)$ but $\frac{\sqrt{n}(\theta-\theta_0)}{\sqrt{\widehat{V}_\theta}}$ tends to be large if n is large, since $\sqrt{\widehat{V}_\theta}$ converges in probability to a positive constant.

Effects of covariates

- In practical applications, we often have a long list of potential explanatory variables.
- In addition, to capture the nonlinear effects and interaction effects, we may expand the linear model by incorporating higher order polynomials and interaction terms.
- While only few of the potential covariates may have non-zero coefficients in the true model, unfortunately we do not know which ones.
- Covariates with zero coefficients are called irrelevant.
- To avoid the omitted variables bias, the researcher may attempt to include all potential covariates. Unfortunately, that results in large variances and standard errors on the main parameters of interest.
- Two wrong practices: (1) include only significant regressors; (2) data snooping/p-hacking.
- Right way: consistent model selection.

- If a subset of the coefficients in the linear model

$$Y_i = \beta_1 X_{i,1} + \ldots + \beta_k X_{i,k} + U_i$$

  are exactly zero, we wish to find the smallest sub-model consisting of only explanatory variables with non-zero coefficients.
- Estimate the full model with all variables. Let T_j denote the t-statistic for testing $\mathbb{H}_0 : \beta_j = 0$ versus $\mathbb{H}_1 : \beta_j \neq 0$.
- What if we run a second regression with only statistically significant coefficients in the first stage?
- Such a practice would typically result in exclusion of relevant covariates and the omitted variables bias.
  - Hypothesis testing controls for the probability of Type I error but leaves the probability of Type II error uncontrolled.
  - You find a coefficient to be non-significant, possibly due to a high probability of Type II error.
  - Failure to reject $\mathbb{H}_0 : \beta_j = 0$ cannot be used as a reliable evidence that the true coefficient is zero.

Data snooping

- Data snooping or p-hacking occurs when the researcher uses the same data in order to produce statistically significant estimates with large t-statistics or small p-values.
- Data snooping destroys the validity of t-statistics and p-values and makes the empirical results less convincing.
- You may try dropping different combinations of potential explanatory variables from the regression to get a statistically significant estimate for the main variable of interest.
- Suppose that the researcher can construct J independent estimators for θ such that $\widehat{\theta}_j \sim \mathrm{N}\left(\theta, \sigma_j^2\right)$, $j = 1, 2, ..., J$, where σ_j^2 is known.
- The researcher conducts J tests with significance level 5% of $\mathbb{H}_0 : \theta = 0$ against $\mathbb{H}_1 : \theta \neq 0$.

- The researcher concludes that $\theta \neq 0$ if one of the J tests rejects $\theta = 0$.
- Suppose that in fact $\theta = 0$. The probability of concluding that $\theta \neq 0$ (known as false discovery) is given by

$$
\begin{aligned}
\Pr\left(\max_{1 \leq j \leq J} \left|\frac{\widehat{\theta}_j}{\sigma_j}\right| > 1.96\right) &= 1 - \Pr\left(\max_{1 \leq j \leq J} \left|\frac{\widehat{\theta}_j}{\sigma_j}\right| \leq 1.96\right) \\
&= 1 - \prod_{i=1}^{J} \Pr\left(\left|\frac{\widehat{\theta}_j}{\sigma_j}\right| \leq 1.96\right) \\
&= 1 - (0.95)^J .
\end{aligned}
$$

- The false discovery probability quickly grows as $J \uparrow \infty$. E.g., $1 - (0.95)^{10} \approx 40\%$.
- When the researcher performs many of tests, the Type I error probability is not controlled and may be much larger than the nominal significance level.

- In practice, estimators are rarely independent, the same relationship holds qualitatively.
- If the researcher searchers long enough, with a high probability they would find a significant estimate.
- A procedure that automatically detects the smallest sub-model consisting of only relevant explanatory variables guards against data snooping and makes the empirical results more convincing to readers.

Consistent model selection

- Order $T_1, ..., T_k$ in absolute value:

$$\left|T_{(1)}\right| \geq \left|T_{(2)}\right| \geq \cdots \geq \left|T_{(k)}\right|.$$

- Let $\hat{j}$ denote the value of j that minimizes $RSS(j) + js^2 \log(n)$, where $RSS(j)$ is the residual sum of squares from the model with j variables corresponding to the j largest absolute t-statistics and $s^2 = (n-k)^{-1} \sum_{i=1}^{n} \widehat{U}_i^2$.
- The selected model is the model with $\hat{j}$ variables corresponding to the $\hat{j}$ largest absolute t-statistics.
- When n is large, with high probability, this selected model is the same as the smallest sub-model with only nonzero coefficients.

Bonferroni Corrections

- Under the joint hypothesis that a set of k hypotheses are all true, what is the probability that the smallest p-value is smaller than α?
- Suppose our null hypothesis $\mathbb{H}_0$ is a joint hypothesis: "$\mathbb{H}_0^1$ is true, $\mathbb{H}_0^2$ is true, ..., and $\mathbb{H}_0^k$ is true" and for each hypothesis we have a test (a testing statistic with an asymptotic p-value p_j).
- Consider the following rule: reject $\mathbb{H}_0$ if any of the hypotheses is rejected, or the smallest p-value is smaller than α.

- But the test may not have "correct size" (the type-I error could be very large):

$$\Pr\left(\min_{1\leq j\leq k} p_j < \alpha\right) \leq \sum_{j=1}^{k} \Pr\left(p_j < \alpha\right) \to k\alpha.$$

- Bonferroni correction: use the adjusted significance level α/k,

$$\Pr\left(\min_{1\leq j\leq k} p_j < \frac{\alpha}{k}\right) \leq \sum_{j=1}^{k} \Pr\left(p_j < \frac{\alpha}{k}\right) \to \alpha.$$

  So the type-I error associated with the decision rule should not be much larger than α.