Topic 1 – Continued…….

Finite-Sample Properties of the LS Estimator

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad ; \qquad \boldsymbol{\varepsilon} \sim N[0\,,\sigma^2 I_n]$$

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = f(\boldsymbol{y})$$

$\boldsymbol{\varepsilon}$ is random ➡ $\boldsymbol{y}$ is random ➡ $\boldsymbol{b}$ is random

- ***b*** is an *estimator* of ***β*.** It is a function of the *random* sample data.
- ***b*** is a “statistic”.
- ***b*** has a probability distribution – called its *Sampling Distribution*.
- Interpretation of *sampling distribution* –

  Repeatedly draw all possible samples of size *n*.

  Calculate values of ***b*** each time.

  Construct relative frequency distribution for the ***b*** values and probability of occurrence.

  It is a *hypothetical* construct. Why?

- Sampling distribution offers *one* basis for answering the question:

“How good is *b* as an estimator of *β* ?”

Note:

Quality of estimator is being assessed in terms of performance in *repeated samples*. Tells us nothing about quality of estimator for *one particular sample*.

- Let’s explore some of the properties of the LS estimator, ***b***, and build up its sampling distribution.
- Introduce some general results, and apply them to our problem.

Definition: An estimator, $\widehat{\boldsymbol{\theta}}$ is an *unbiased* estimator of the parameter vector, $\boldsymbol{\theta}$, if $E[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$.

That is, $E[\widehat{\boldsymbol{\theta}}(\boldsymbol{y})] = \boldsymbol{\theta}$.

That is, $\int \hat{\theta}(\boldsymbol{y}) p(\boldsymbol{y} \mid \boldsymbol{\theta}) d\boldsymbol{y} = \boldsymbol{\theta}$.

The quantity, $\boldsymbol{B}(\boldsymbol{\theta}, \boldsymbol{y}) = E[\widehat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}]$, is called the "Bias" of $\widehat{\boldsymbol{\theta}}$.

Example: $\{y_1, y_2, \dots \dots, y_n\}$ is a random sample from population with a finite mean, μ, and a finite variance, σ^2 .

Consider the *statistic* $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Then, $E[\bar{y}] = E\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \frac{1}{n}\sum_{i=1}^{n} E(y_i)$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu = \left(\frac{1}{n} n\mu\right) = \mu \ .$$

So, $\bar{y}$ is an *unbiased estimator* of the parameter, μ.

- Here, there are lots of possible unbiased estimators of μ.
- So, need to consider additional characteristics of estimators to help choose.

Return to our LS problem –

$$\boldsymbol{b} = (X'X)^{-1}X'\boldsymbol{y}$$

- Recall – either assume that X is *non-random*, or condition on X.
- We'll assume X is non-random – get same result if we condition on X.

Then: $E(\boldsymbol{b}) = E[(X'X)^{-1}X'\boldsymbol{y}] = (X'X)^{-1}X'E(\boldsymbol{y})$

So,

$$E(\boldsymbol{b}) = (X'X)^{-1}X'E[X\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = (X'X)^{-1}X'[X\boldsymbol{\beta} + E(\boldsymbol{\varepsilon})]$$

$$= (X'X)^{-1}X'[X\boldsymbol{\beta} + \boldsymbol{0}] = (X'X)^{-1}X'X\boldsymbol{\beta}$$

$$= \boldsymbol{\beta} \ .$$

The LS estimator of β is Unbiased

Definition: Any estimator that is a *linear function* of the random sample data is called a *Linear Estimator*.

Example: $\{y_1, y_2, \dots \dots, y_n\}$ is a random sample from population with a finite mean, μ, and a finite variance, σ^2 .

Consider the *statistic* $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}[y_1 + y_2 + \cdots + y_n]$.

This statistic is a *linear estimator* of μ.

(Note that the "weights" are non-random.)

Return to our LS problem –

$$\boldsymbol{b} = (X'X)^{-1}X'\boldsymbol{y} = A\boldsymbol{y}$$

$(k\times1)$ $(k\times n)(n\times1)$

Note that, under our assumptions, A is a *non-random* matrix.

So,

$$\begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \cdots & \vdots \\ a_{k1} & \cdots & a_{kn} \end{bmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} .$$

For example, $b_1 = [a_{11}y_1 + a_{12}y_2 + \cdots + a_{1n}y_n]$; *etc*.

The LS estimator, *b*, is a linear (& unbiased) estimator of *β*

Now let's consider the dispersion (variability) of $\boldsymbol{b}$, as an estimator of $\boldsymbol{\beta}$.

Definition: Suppose we have an (n×1) random vector, $\boldsymbol{x}$. Then the *Covariance Matrix* of $\boldsymbol{x}$ is defined as the (n×n) matrix:

$$V(\boldsymbol{x}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))'].$$

- Diagonal elements of $V(\boldsymbol{x})$ are $var.(x_1)$, ……., $var.(x_n)$.
- Off-diagonal elements are $covar.(x_i, x_j)$; $i, j = 1, \ldots, n$; $i \neq j$.

Return to our LS problem –

We have a (k×1) random vector, $\boldsymbol{b}$, and we know that $E(\boldsymbol{b}) = \boldsymbol{\beta}$.

$$V(\boldsymbol{b}) = E[(\boldsymbol{b} - E(\boldsymbol{b}))(\boldsymbol{b} - E(\boldsymbol{b}))']$$

Now,

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

$$= (X'X)^{-1}(X'X)\boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{\varepsilon}$$

$$= I\boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{\varepsilon}.$$

So,

$$(\boldsymbol{b} - \boldsymbol{\beta}) = (X'X)^{-1}X'\boldsymbol{\varepsilon} \,. \qquad \textbf{[*]}$$

Using the result, **[*]**, in $V(\boldsymbol{b})$, we have:

$$V(\boldsymbol{b}) = E\{[(X'X)^{-1}X'\boldsymbol{\varepsilon}][(X'X)^{-1}X'\boldsymbol{\varepsilon}]'\}$$

$$= (X'X)^{-1}X'E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']X(X'X)^{-1} \,.$$

We showed, earlier, that because $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $V(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 I_n$.

(*What other assumptions did we use to get this result?*)

So, we have:

$$V(\boldsymbol{b}) = (X'X)^{-1}X'E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']X(X'X)^{-1} = (X'X)^{-1}X'\sigma^2 IX(X'X)^{-1} = \sigma^2(X'X)^{-1}(X'X)(X'X)^{-1}$$

$$= \sigma^2(X'X)^{-1}.$$

$$V(\boldsymbol{b}) = \sigma^2(X'X)^{-1}$$

$(k \times k)$

Interpret diagonal and off-diagonal elements of this matrix.

Finally, because the error term, $\boldsymbol{\varepsilon}$ is assumed to be Normally distributed,

1. $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$: this implies that $\boldsymbol{y}$ is also Normally distributed. (Why?)
2. $\boldsymbol{b} = (X'X)^{-1}X'\boldsymbol{y} = A\boldsymbol{y}$: this implies that $\boldsymbol{b}$ is also Normally distributed.

So, we now have the full **Sampling Distribution** of the LS estimator, $\boldsymbol{b}$:

$$\boldsymbol{b} \sim N[\boldsymbol{\beta}, \sigma^2(X'X)^{-1}]$$

Note:

- This result depends on our various, *rigid*, assumptions about the various components of the regression model.
- The Normal distribution here is a "*multivariate* Normal" distribution.
  (*See handout on "Spherical Distributions"*.)
- As with estimation of population mean, $\boldsymbol{\mu}$, in previous example, there are lots of other *unbiased* estimators of $\boldsymbol{\beta}$ in the model $= X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.
- How might we choose between these possibilities? Is *linearity* desirable?

- We need to consider other *desirable* properties that these unbiased estimators may have.
- *One option* is to take account of estimators' *precisions*.

Definition: Suppose we have two *unbiased* estimators, $\widehat{\theta_1}$ and $\widehat{\theta_2}$, of the (scalar) parameter, θ. Then we say that $\widehat{\theta_1}$ is **at least as efficient** as $\widehat{\theta_2}$ if $var.(\widehat{\theta_1}) \leq var.(\widehat{\theta_2})$.

Note:

1. The variance of an estimator is just the variance of its sampling distribution.
2. "Efficiency" is a *relative* concept.
3. What if there are 3 or more unbiased estimators being compared?

- What if one or more of the estimators being compared is *biased* ?
- In this case we can take account of both variance, and any bias, at the same time by using "*mean squared error*" (MSE) of the estimators.

Definition: Suppose that $\hat{\theta}$ is an estimator of the (*scalar*) parameter, θ. Then the MSE of $\hat{\theta}$ is defined as:

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right].$$

Note that:

$$MSE(\hat{\theta}) = var.(\hat{\theta}) + [Bias(\hat{\theta})]^2$$

To prove this, write:

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] = E\{[((\hat{\theta}) - E(\hat{\theta})) + (E(\hat{\theta}) - \theta)]^2\},$$

expand out, and note that

$$E[E(\hat{\theta})] = E(\hat{\theta}) ;$$

and

$E[\hat{\theta} - E(\hat{\theta})] = 0.$

Definition: Suppose we have two (possibly) *biased* estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, of the (scalar) parameter, θ. Then we say $\hat{\theta}_1$ is **at least as efficient** as $\hat{\theta}_2$ if $MSE(\hat{\theta}_1) \leq MSE(\hat{\theta}_2)$.

If we extend all of this to the case where we have a vector of parameters, , then we have the following definitions:

Definition: Suppose we have two *unbiased* estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, of the parameter vector, $\boldsymbol{\theta}$. Then we say that $\hat{\theta}_1$ is **at least as efficient** as $\hat{\theta}_2$ if $\Delta = V(\hat{\theta}_2) - V(\hat{\theta}_1)$ is *at least positive semi-definite*.

Definition: Suppose we have two (possibly) *biased* estimators, $\hat{\theta}_1$and $\hat{\theta}_2$, of the parameter vector, $\boldsymbol{\theta}$. Then we say that $\hat{\theta}_1$ is **at least as efficient** as $\hat{\theta}_2$ if $\Delta = MMSE(\hat{\theta}_2) - MMSE(\hat{\theta}_1)$ is *at least positive semi-definite*.

Note: $MMSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)(\hat{\theta} - \theta)'\right] = V[\hat{\theta}] + Bias(\hat{\theta})Bias(\hat{\theta})'$.

Taking account of its *linearity*, *unbiasedness*, and its *precision*, in what sense is the LS estimator, ***b***, of β *optimal*?

Theorem (Gauss-Markhov):

In the "standard" linear regression model, $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the LS estimator, ***b***, of $\boldsymbol{\beta}$ is **Best Linear Unbiased** (BLU). That is, it is **Efficient** in the class of all linear and unbiased estimators of β.

1. Is this an *interesting* result?
2. What *assumptions* about the "standard" model are we going to exploit?

Proof

Let $\boldsymbol{b_0}$ be any other *linear* estimator of $\boldsymbol{\beta}$:

$$\boldsymbol{b_0} = C\boldsymbol{y} \qquad ; \qquad \text{for } \textit{some} \text{ non-random } C \, .$$

$(k\times1) \quad (k\times n)(n\times1)$

Now, $\quad V(\boldsymbol{b_0}) = CV(\boldsymbol{y})C' = C(\sigma^2 I_n)C' = \sigma^2 CC'$

$(k\times k)$

Define: $\quad D = C - (X'X)^{-1}X'$

so that $\quad D\boldsymbol{y} = C\boldsymbol{y} - (X'X)^{-1}X'\boldsymbol{y} = \boldsymbol{b_0} - \boldsymbol{b} \quad .$

Now restrict $\boldsymbol{b_0}$ to be *unbiased*, so that $E(\boldsymbol{b_0}) = E(C\boldsymbol{y}) = CX\boldsymbol{\beta} = \boldsymbol{\beta}$.

This requires that $CX = I$, which in turn implies that

$$DX = [C - (X'X)^{-1}X']X = CX - I = 0 \qquad (and\ D'X' = 0)$$

(*What assumptions have we used so far?*)

Now, focus on covariance matrix of $\boldsymbol{b_0}$:

$$V(\boldsymbol{b_0}) = \sigma^2[D + (X'X)^{-1}X'][D + (X'X)^{-1}X']'$$

$$= \sigma^2[DD' + (X'X)^{-1}X'X(X'X)^{-1}] \qquad ; \qquad DX = 0$$

$$= \sigma^2 DD' + \sigma^2(X'X)^{-1}$$

$$= \sigma^2 DD' + V(\boldsymbol{b}),$$

or, $\quad [V(\boldsymbol{b_0}) - V(\boldsymbol{b})] = \sigma^2 DD' \qquad ; \qquad \sigma^2 > 0$

Now we just have to "sign" this (matrix) difference:

$$\boldsymbol{\eta}'(DD')\boldsymbol{\eta} = (D'\boldsymbol{\eta})'(D'\boldsymbol{\eta}) = v'v = \sum_{i=1}^{n} v_i^2 \geq 0 \, .$$

So, $\varDelta = [V(\boldsymbol{b_0}) - V(\boldsymbol{b})]$ is a p.s.d. matrix, implying that $\boldsymbol{b_0}$ is *relatively less efficient* than $\boldsymbol{b}$.

Result:

The LS estimator is the Best Linear Unbiased estimator of $\boldsymbol{\beta}$.

- What assumptions did we use, and where?
- Were there any standard assumptions that we *didn't* use?
- What does this suggest?

Estimating σ^2

- We now know a lot about estimating $\boldsymbol{\beta}$.
- There's another parameter in the regression model - σ^2 – the variance of each ε_i .
- Note that $\sigma^2 = var.(\varepsilon_i) = E[(\varepsilon_i - E(\varepsilon_i))^2] = E(\varepsilon_i^2)$.
- The *sample* counterpart to this *population* parameter is the *sample* average of the "residuals": $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\boldsymbol{e'e}$.
- However, there is a *distortion* in this estimator of σ^2 .
- Although mean of e_i's is zero (if intercept in model), not all of e_i's are independent of each other – only (*n* – *k*) of them are.
- Why does this distort our potential estimator, $\hat{\sigma}^2$?

Note that: $e_i = (y_i - \hat{y}_i) = (y_i - x_i'\boldsymbol{b})$

$$= (x_i'\boldsymbol{\beta} + \varepsilon_i) - x_i'\boldsymbol{b}$$

$$= \varepsilon_{\boldsymbol{i}} + \boldsymbol{x_i}'(\boldsymbol{\beta} - \boldsymbol{b})$$

Let's see what properties $\hat{\sigma}^2$ has as an estimator of σ^2 :

$$\boldsymbol{e} = (\boldsymbol{y} - \hat{\boldsymbol{y}}) = (\boldsymbol{y} - X\boldsymbol{b}) = \boldsymbol{y} - X(X'X)^{-1}X'\boldsymbol{y} = M\boldsymbol{y} ,$$

where

$$M = I_n - X(X'X)^{-1}X' \quad ; \quad \textit{idempotent}\text{, and } MX = 0 \ .$$

So, $$\boldsymbol{e} = M\boldsymbol{y} = M(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = M\boldsymbol{\varepsilon} \ ,$$

and $$\boldsymbol{e}'\boldsymbol{e} = (M\boldsymbol{\varepsilon})'(M\boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon}'M\boldsymbol{\varepsilon} \quad ; \quad \textit{scalar}$$

From this, we see that:

$$E(\boldsymbol{e}'\boldsymbol{e}) = E[\boldsymbol{\varepsilon}'M\boldsymbol{\varepsilon}] = E[tr.(\boldsymbol{\varepsilon}'M\boldsymbol{\varepsilon})] = E[tr.(M\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')]$$

$$= tr.[ME(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')] = tr.[M\sigma^2 I_n] = \sigma^2 tr.(M)$$

$$= \sigma^2(n-k)$$

So:

$$E(\hat{\sigma}^2) = E(\tfrac{1}{n}\boldsymbol{e}'\boldsymbol{e}) = \tfrac{1}{n}(n-k)\sigma^2 < \sigma^2 \quad ; \quad \textbf{BIASED}$$

Easy to convert this to an ***Unbiased estimator* –**

$$s^2 = \frac{1}{(n-k)}\boldsymbol{e}'\boldsymbol{e}$$

- "$(n-k)$" is the "*degrees of freedom*" – number of independent sources of information in the "n" residuals (e_i's).
- We can use "s" as an estimator of , but it is a *biased estimator*.
- Call "s" the "*standard error of the regression*", or the "*standard error of estimate*".
- s^2 is a *statistic* – has its own sampling distribution, *etc*. More on this to come.
- Let's see one immediate *application* of s^2 and s.
- Recall sampling distribution for LS estimator, $\boldsymbol{b}$:

$$\boldsymbol{b} \sim N[\boldsymbol{\beta} \ , \ \sigma^2(X'X)^{-1}]$$

- So, $var.(b_i) = \sigma^2[(X'X)^{-1}]_{ii}$; σ^2 is *unobservable*.

- If we want to report variability associated with b_i as an estimator of β_i, we need to use estimator of σ^2 .
- $est.var.(b_i) = s^2[(X'X)^{-1}]_{ii}$.
- $\sqrt{est.var.(b_i)} = \widehat{s.d.}(b_i) = s\{[(X'X)^{-1}]_{ii}\}^{1/2}$.
- We call this the "*standard error*" of b_i.
- This quantity will be very important when it comes to constructing *interval estimates* of our regression coefficients; and when we construct *tests of hypotheses* about these coefficients.

Confidence Intervals & Hypothesis Testing

- So far, we've concentrated on "*point*" estimation.
- Need to move on – to do this we'll need the full sampling distributions of **both** $\boldsymbol{b}$ and s^2.
- We will make use of the assumption of *Normally distributed* errors.
- Recall that:

$$\boldsymbol{b} \sim N[\boldsymbol{\beta}, \sigma^2(X'X)^{-1}]$$

$$b_i \sim N[\beta_i, \sigma^2((X'X)^{-1})_{ii}] \quad ; \quad \text{why still } Normal?$$

- So, we can *standardize*:

$$z_i = (b_i - \beta_i)/\sqrt{\sigma^2[(X'X)^{-1}]_{ii}}$$

- But σ^2 is *unknown*, so we can't use z_i directly to draw inferences about b_i.

Need some preliminary results in order to proceed from here –

Definition: Let $z \sim N[0, 1]$. Then z^2 has a *"Chi-Square" distribution* with one "degree of freedom".

Definition: Let $z_2, z_2, z_3, \ldots, z_m$ be *independent* N[0 , 1] variates. Then the quantity $\sum_{i=1}^{m}(z_i^2)$ has a Chi-Square distribution with "m" d.o.f.

Theorem: Let $\boldsymbol{x} \sim N[\boldsymbol{0}, V]$, and let A be a fixed matrix. Then the *quadratic form*, $'A\boldsymbol{x}$, follows a Chi-Square distribution with r ($= rank(A)$) degrees of freedom, iff AV is an *idempotent matrix*.

Definition: Let $z \sim N[0\,,1]$, and let $x \sim \chi^2_{(v)}$, where z and x are *independent*. Then the statistic, $t = z/\sqrt{x/v}$ follows ***Student's t distribution***, with "v" degrees of freedom.

Now let's consider the sampling distribution of s^2:

We have $$s^2 = \frac{1}{(n-k)}\boldsymbol{e}'\boldsymbol{e}\ .$$

So,

$$(n-k)s^2 = (\boldsymbol{e}'\boldsymbol{e}) = (\boldsymbol{\varepsilon}'M\boldsymbol{\varepsilon})\ \ .$$

Define the random variable

$$C = \frac{(n-k)s^2}{\sigma^2} = \left(\frac{\varepsilon}{\sigma}\right)'M\left(\frac{\varepsilon}{\sigma}\right)\ \ ,$$

where $\boldsymbol{\varepsilon} \sim N[\mathbf{0}\,,\sigma^2 I_n]$; and so $\left(\frac{\varepsilon}{\sigma}\right) \sim N[\mathbf{0}\,,I_n]$.

Using the Theorem from last slide, we get the following result for C:

$$C = \left(\frac{\varepsilon}{\sigma}\right)' M\left(\frac{\varepsilon}{\sigma}\right) \sim \chi^2_{(n-k)}\ \ ,$$

because $AV = MI = M$, is *idempotent*, and $r = d.o.f. = rank(A) = rank(M) = tr.(M) = (n-k)$. (Why?)

So, we have the result:

$$\frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{(n-k)}$$

Next, we need to show that b and s^2 are *statistically independent*.

Theorem: Let $\boldsymbol{x}$ be a *normally distributed* random vector, and L and A are *non-random* matrices. Then, the "Linear Form", $L\boldsymbol{x}$, and the "Quadratic Form", $'A\boldsymbol{x}$, are independent if $LA = \mathbf{0}$.

How does this result help us?

- We have $C = \frac{(n-k)s^2}{\sigma^2} = \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)' M \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)$.
- Also, $\boldsymbol{b} = (X'X)^{-1}X'\boldsymbol{y} = (X'X)^{-1}X'(X\boldsymbol{\beta} + \boldsymbol{\varepsilon})$
  $= \boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{\varepsilon}$.
- So, $\left[\frac{\boldsymbol{b}-\boldsymbol{\beta}}{\sigma}\right] = (X'X)^{-1}X'\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)$.
- Let $L = (X'X)^{-1}X'$; $A = M$; $\boldsymbol{x} = \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)$
- So, $LA = (X'X)^{-1}X'M = \mathbf{0}$
- This implies that $C = \frac{(n-k)s^2}{\sigma^2}$ and $\left[\frac{\boldsymbol{b}-\boldsymbol{\beta}}{\sigma}\right]$ are *independent*, and so $\boldsymbol{b}$ and s^2 are also *statistically independent*.
- C is $\chi^2_{(n-k)}$, and $\left[\frac{b-\beta}{\sigma}\right] \sim N[\mathbf{0}\,, (X'X)^{-1}]$, so we immediately get:

Theorem: $t_i = (b_i - \beta_i)/\, s.e.(b_i)$

has a Student's t distribution with (n - k) d.o.f.

Proof: $\left[\frac{b-\beta}{\sigma}\right] \sim N[\mathbf{0}\,, (X'X)^{-1}]$, $\left[\frac{b_i-\beta_i}{\sigma}\right] \sim N[\mathbf{0}\,, ((X'X)^{-1})_{ii}]$

so, $\left[\frac{b_i-\beta_i}{\sigma\sqrt{((X'X)^{-1})_{ii}}}\right] \sim N[0\,, 1]$.

Also, $C = \frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{(n-k)}$; and we have *independence*.

So, $t_v = N[0\,, 1] / \sqrt{\chi^2_{(v)}/v}$

$$= \left[\frac{b_i-\beta_i}{\sigma\sqrt{((X'X)^{-1})_{ii}}}\right] / \left[\frac{(n-k)s^2}{\sigma^2}/(n-k)\right]^{1/2}$$

$$= \left[\frac{b_i - \beta_i}{s\sqrt{((X'X)^{-1})_{ii}}}\right] = \left[\frac{b_i - \beta_i}{s.e.(b_i)}\right] .$$

In this case, $v = (n - k)$, and so:

$$\left[\frac{b_i - \beta_i}{s.e.(b_i)}\right] \sim t_{(n-k)}$$

We can use this to construct *confidence intervals* and *test hypotheses* about β_i .

Note: This last result used all of our assumptions about the linear regression model – including the assumption of *Normality for the errors*.

Example 1:

$$\hat{y} = 1.4 + 0.2x_2 + 0.6x_3$$
$$(0.7) \quad (0.05) \quad (1.4)$$

$$H_0: \beta_2 = 0 \quad vs. \quad H_A: \beta_2 > 0$$

$$t = \left[\frac{b_2 - \beta_2}{s.e.(b_2)}\right] = \left[\frac{0.2 - 0}{0.05}\right] = 4$$; suppose $n = 20$

$t_c(5\%) = 1.74$; $t_c(1\%) = 2.567$; d.o.f. = 17

$t > t_c \Rightarrow$ *Reject* H_0 .

Degrees of Freedom	90th Percentile	95th Percentile	97.5th Percentile	99th Percentile	99.5th Percentile
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
:	:	:	:	:	:
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	**1.333**	**1.740**	**2.110**	**2.567**	**2.898**

Example 2:

$$\hat{y} = 1.4 + 0.2x_2 + 0.6x_3$$
$$(0.7)\ \ (0.05)\ \ \ (1.4)$$

$H_0\colon\ \beta_1 = 1.5 \quad vs. \quad H_A\colon\ \beta_1 \neq 1.5$

$t = \left[\frac{b_1 - \beta_1}{s.e.(b_1)}\right] = \left[\frac{1.4 - 1.5}{0.7}\right] = -0.1429$; d.o.f. = 17

$t_c(5\%) = \pm 2.11$

$|t| < t_c \quad \Rightarrow \quad Do\ Not\ Reject\ H_0$

(Against H_A , at the 5% significance level.)

Example 3:

$$\hat{y} = 1.4 + 0.2x_2 + 0.6x_3$$
$$(0.7)\ \ (0.05)\ \ \ (1.4)$$

$H_0\colon\ \beta_1 = 1.5 \quad vs. \quad H_A\colon\ \beta_1 < 1.5$

$t = \left[\frac{b_1 - \beta_1}{s.e.(b_1)}\right] = \left[\frac{1.4 - 1.5}{0.7}\right] = -0.1429$; d.o.f. = 17

$p - value = Pr.[t < -0.1429\ |H_0\ is\ True]$

in R: `pt(-0.1429,17)`

***p* = 0.444**

What do you conclude?

Some Properties of Tests:

Null Hypothesis (H_0) Alternative Hypothesis (H_A)

Classical hypothesis testing –

- Assume that H_0 is *TRUE*
- Compute value of test statistic using random sample of data
- Determine *distribution* of the test statistic (*when* H_0 *is true*)
- Check of observed value of test statistic is likely to occur, *if* H_0 *is true*
- If this event is sufficiently *unlikely*, then **REJECT H_0** (in favour of H_A)

Note:

1. Can never **accept** H_0. Why not?
2. What constitutes "*unlikely*" – subjective?
3. Two types of errors we might incur with this process

Type I Error: **Reject** H_0 when in fact it is **True**

Type II Error: **Do Not Reject** H_0 when in fact it is **False**

- Pr.[I] = α = Significance level of test = "size" of test
- Pr.[II] = β ; say
- Value of β will depend on <u>*how*</u> H_0 is **False**. Usually, many ways.
- In classical testing, decide in advance on max. acceptable value of α and then try and design test so as to *minimize* β.
- As β can take different values, may be difficult to design test optimally.
- Why not minimize both? A trade-off for fixed value of n.
- Consider some desirable properties for a test.

Definition:

The "**Power**" of a test is Pr.[**Reject** H_0 when it is **False**].

So, Power = 1 – Pr.[**Do Not Reject** H_0 | H_0 is **False**] = $1 - \beta$.

- As β typically changes, depending on the *way* that H_0 is false, we usually have a **Power Curve**.
- For a fixed value of α, this curve plots Power against parameter value(s).
- We want our tests to have *high power*.
- We want the power of our tests to *increase* as H_0 becomes *increasingly false*.

Property 1

Consider a fixed sample size, n, and a fixed significance level, α.

Then, a test is "Uniformly Most Powerful" if its power exceeds (or is no less than) that of *any other test*, for all possible ways that H_0 could be False.

Property 2

Consider a fixed significance level, α.

Then, a test is "Consistent" if its power → 1, as $n \to \infty$, for all possible ways that H_0 is false.

Property 3

Consider a fixed sample size, n, and a fixed significance level, α.

Then, a test is said to be "Unbiased" its power *never* falls below the significance level.

Property 4

Consider a fixed sample size, n, and a fixed significance level, α.

Then, a test is said to be "Locally Most Powerful" if the *slope* of its power curve is greater than the slope of the power curves of all other size – α tests, in a neighbourhood of H_0.

Note:

- For many testing problems, no UMP test exists. This is why LMP tests are important.
- Why do we use our "t-test" in the regression model –
    1. It is UMP, against 1 –sided alternatives.
    2. It is Unbiased.
    3. It is Consistent.
    4. It is LMP, against both 1-sided and 2-sided alternatives.

Confidence Intervals

We can also use our t-statistic to construct a confidence interval for β_i.

$$Pr.[-t_c \leq t \leq t_c] = (1-\alpha)$$

$$\Rightarrow \quad Pr.\left[-t_c \leq \left[\frac{b_i-\beta_i}{s.e.(b_i)}\right] \leq t_c\right] = (1-\alpha)$$

$$\Rightarrow \quad Pr.[-t_c\, s.e.(b_i) \leq (b_i-\beta_i) \leq t_c\, s.e.(b_i)] = (1-\alpha)$$

$$\Rightarrow \quad Pr.[-b_i - t_c\, s.e.(b_i) \leq (-\beta_i) \leq -b_i + t_c\, s.e.(b_i)]$$

$$= (1-\alpha)$$

$$\Rightarrow \quad Pr.[b_i + t_c\, s.e.(b_i) \geq \beta_i \geq b_i - t_c\, s.e.(b_i)] = (1-\alpha)$$

$$\Rightarrow \quad Pr.[b_i - t_c\, s.e.(b_i) \leq \beta_i \leq b_i + t_c\, s.e.(b_i)] = (1-\alpha)$$

Interpretation –

The interval, $[b_i - t_c\, s.e.(b_i)\ ,\ b_i + t_c\, s.e.(b_i)]$ is *random*.

The parameter, β_i, is *fixed* (but unknown).

If we were to take a sample of *n* observations, and construct such an interval, and then repeat this exercise many, many, times, then $100(1-\alpha)\%$ of such intervals would cover the true value of β_i.

If we just construct an interval, for our *given* sample of data, we'll never know if *this particular* interval covers β_i, or not.

Example 1

$$\hat{y} = 0.3 - 1.4x_2 + 0.7x_3$$

(0.1) (1.1) (0.2)

Construct a 95% confidence interval for β_1 when n = 30.

d.o.f. = (n – k) = 27 ; (α/2) = 0.025

$t_c = \pm 2.052$; $b_1 = 0.3$; $s.e.(b_1) = 0.1$

The 95% Confidence Interval is:

$$[b_1 - t_c\, s.e.(b_1)\ , \quad b_1 + t_c\ s.e.(b_1)]$$

⇒ [0.3 – (2.052)(0.1) , 0.3 + (2.052)(0.1)]

⇒ [0.0948 , 0.5052]

Don't forget the units of measurement!

Example 2

$$\hat{y} = 0.3 - 1.4x_2 + 0.7x_3$$

(0.1) (1.1) (0.2)

Construct a 90% confidence interval for β_2 when n = 16.

d.o.f. = (n – k) = 13 ; (α/2) = 0.05

$t_c = \pm 1.771$; $b_2 = -1.4$; $s.e.(b_2) = 1.1$

The 95% Confidence Interval is:

$$[b_2 - t_c\, s.e.(b_2)\ , \quad b_2 + t_c\ s.e.(b_2)]$$

⇒ [-1.4 – (1.771)(1.1) , -1.4 + (1.771)(1.1)]

⇒ [-3.3481 , 0.5481]

Don't forget the units of measurement!

Questions:

- Why do we construct the interval *symmetrically* about point estimate, b_i?
- How can we use a Confidence Interval to test hypotheses?
- For instance, in the last Example, can we reject H_0: $\beta_2 = 0$, against a 2-sided alternative hypothesis?