# CONFIDENCE INTERVAL FOR REGRESSION COEFFICIENTS $(\hat{a}, \hat{b})$

$\hat{a} \sim N(a, \sigma_{\hat{a}}^2) \Rightarrow Z = \frac{\hat{a}-a}{\sigma_{\hat{a}}} \sim N(0,1)$

$\hat{b} \sim N(b, \sigma_{\hat{b}}^2) \Rightarrow Z = \frac{\hat{b}-b}{\sigma_{\hat{b}}} \sim N(0,1)$

NOTE: $(n-2)\frac{\hat{\sigma}_u^2}{\sigma_u^2} \sim \chi^2_{n-2}$ (CHI-SQUARE DISTRIBUTION)

WHERE $\hat{\sigma}_u^2$ IS UNBIASED ESTIMATOR FOR $\sigma_u^2$

AND $\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$. (FROM LAST MEETING)

$\frac{\hat{b}-b}{\sigma_{\hat{b}}} = Z \sim N(0,1)$ WHERE $\sigma_{\hat{b}} = \sqrt{\frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}}$

$= \frac{\sigma_u}{\sqrt{\sum_{i=1}^{n} x_i^2}}$

BUT! $\sigma_u$ IS UNKNOWN.

IF $\sigma_u^2$ IS KNOWN, WE CAN USE THE STANDARDIZED NORMAL DISTRIBUTION TO FIND THE CONFIDENCE INTERVAL FOR PARAMETER $b$.

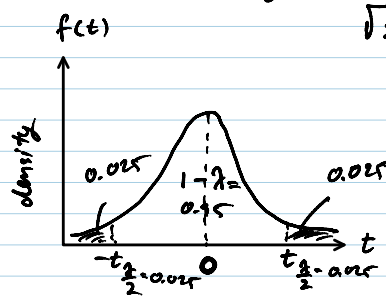
IN PRATICE, WE USE THE UNBIASED ESTIMATOR FOR ESTIMATING $\sigma_u^2$:

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

IF WE USE $\hat{\sigma}_{\hat{b}} = \frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^{n} x_i^2}}$, THEN $\frac{\hat{b}-b}{\hat{\sigma}_{\hat{b}}} = t$ WITH $(n-2)$ df.

THEREFORE $\frac{\hat{b}-b}{S_{\hat{b}}} \sim t_{n-2}$

WHERE $S_{\hat{b}} = \frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^{n} x_i^2}}$




WE USE THE t-DISTRIBUTION TO "CONSTRUCT" A CONFIDENCE INTERVAL FOR $b$.

IF $\lambda = 0.05$ (WITH 95% CONFIDENCE)

**T-DISTRIBUTION:** $Y_i = a + bX_i + u_i$

- $n = 22$, $k = 2$, $n-k = 22-2 = 20$
- LET $\lambda = 0.05$ ← LEVEL OF SIGNIFICANCE

$$Pr\left\{-t_{\frac{\lambda}{2}, n-k} \leq \frac{\hat{b}-b}{S_{\hat{b}}} \leq +t_{\frac{\lambda}{2}, n-k}\right\} = (1-\lambda)$$

(0.95)

$$Pr\left\{-t_{\frac{\lambda}{2},n-k} \leq \frac{\hat{b}-b}{s_{\hat{b}}} \leq +t_{\frac{\lambda}{2},n-k}\right\} = (1-\lambda)$$

$$= Pr\left\{-t_{0.025,20} \leq \frac{\hat{b}-b}{s_{\hat{b}}} \leq t_{0.025,20}\right\} = 0.95$$

$$= Pr\left\{-2.086 \leq \frac{\hat{b}-b}{s_{\hat{b}}} \leq 2.086\right\} = 0.95$$

IS 95% CONFIDENCE INTERVAL FOR $b$, WHICH IS

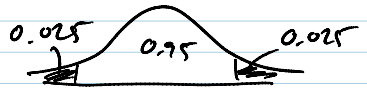
$$\hat{b} \pm s_{\hat{b}}(2.086)$$

Q: HOW TO INTERPRET THIS ?

A: WHEN THE PROCEDURE WE USED IS APPLIED TO MANY RANDOM SAMPLES OF DATA FROM THE SAME POPULATION, THEN 95% OF ALL INTERVAL ESTIMATES CONSTRUCTED USING THIS PROCEDURE WILL CONTAIN THE TRUE VALUE OF THE PARAMETER.

NOTE $Pr$ 0.025 $\Rightarrow$ FOR ONE-TAILED TEST


$Pr$ 0.5 $\Rightarrow$ FOR TWO-TAILED TEST




(IN GUJARATI T-TABLE)

# HYPOTHESIS TESTING

EX: IF MODEL IS $Y_i = a + b X_i + u_i$

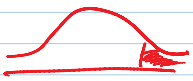
AND $\hat{b} = 5.0$, $s_{\hat{b}} = 2.0$

NULL HYPOTHESIS / ALTERNATIVE HYPOTHESIS

| | | |
|---|---|---|
| $H_0$: | $b = 0$ | (NO RELATIONSHIP BETWEEN Y AND X) |
| $H_1$: | $b \neq 0$ | (2-SIDED TEST) |

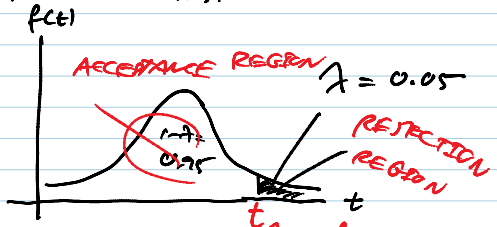
OR

| | | |
|---|---|---|
| $H_0$: | $b \leq 0$ | |
| $H_1$: | $b > 0$ | (1-SIDED TEST) |



SUPPOSE WE LOOK AT ONE-SIDED TEST

FROM $\frac{\hat{b}-b}{s_{\hat{b}}} = t$




CRITICAL $t$

$\Rightarrow Pr(t > t_{\lambda, n-k})$

$= 0.05$

IF $n - k = 20$, $\alpha = 0.05$

$t_{0.05, 20} = 1.725$

THEN $Pr\left( \hat{t} = \frac{\hat{b} - b}{S_{\hat{b}}} > 1.725 \right) = 0.05$

EX: $Pr\left\{ \frac{\hat{b} - 0}{S_{\hat{b}}} > 1.725 \right\} = 0.05$

BY SETTING THE HYPOTHESES AS $H_0: b = 0$

$H_1: b > 0$

- IF NULL HYPOTHESIS CANNOT BE REJECTED, THEN X HAS NO INFLUENCE ON Y.
- IF ALTERNATIVE HYPOTHESIS IS ACCEPTED, THEN X HAS AN INFLUENCE ON Y.

$\therefore Pr\left\{ \frac{2.5}{2} = \hat{t} > 1.725 \right\} = 0.05$

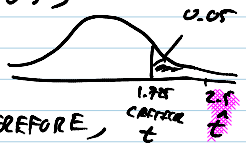
SINCE $\underbrace{\hat{t} = 2.5}_{\text{observed } t}$ BUT $\underbrace{t_{0.05, 20} = 1.725}_{\text{CRITICAL } t}$

$\therefore$ WE FOUND THAT $\hat{t} > t_{0.05, 20}$, WE REJECT THE NULL HYPOTHESIS.

REASON ?

① BECAUSE THE PROBABILITY OF OBSERVING $\hat{t} > 1.725$ IS SO LOW (0.05)

[WE FOUND THAT $\hat{t} = 2.5$]




BUT WE STILL OBSERVE IT. THEREFORE, THE NULL HYPOTHESIS SHOULD NOT BE TRUE (AND THEN WE REJECT IT.)

NEXT, ANALYSIS OF VARIANCE (ANOVA)

8 MAR 2012

CONSIDER $\boxed{\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2}$

WHERE $\sum y_i^2 = \sum (Y_i - \bar{Y})^2$ CALLED "TOTAL SUM OF SQUARE (TSS)

$\sum \hat{y}_i^2 = \sum (\hat{Y}_i - \bar{Y})^2$ CALLED "EXPLAINED SUM OF SQUARE (ESS)

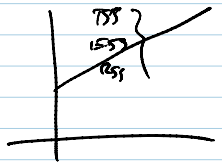
$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$ CALLED "UNEXPLAINED SUM OF SQUARE OR RESIDUAL SUM OF SQUARE (RSS)

SO, $\boxed{TSS = ESS + RSS}$

| SOURCE OF VARIATION | SUM OF SQUARE | MEAN SUM OF SQUARE |
|---|---|---|
| REGRESSION (X) | $\sum_{i=1}^{n} \hat{y}_i^2$ | $\sum \hat{y}_i^2 / (k-1) \rightarrow df$ |
| ERROR | $\sum_{i=1}^{n} e_i^2$ | $\sum e_i^2 / (n-k) \rightarrow df$ |
| TOTAL | $\sum_{i=1}^{n} y_i^2$ | $\sum y_i^2 / (n-1) \rightarrow df$ |

TAKE THE RATIO.

$$\frac{\sum_{i=1}^{n} \hat{y}_i^2 / (k-1)}{\sum e_i^2 / (n-k)} \sim F_{k-1,\, n-k}$$




NOTE THAT $\sum \hat{y}_i^2 / k-1$ AND $\sum e_i^2 / (n-k)$ have $\chi^2$ DISTRIBUTION

IF $\hat{F} = \dfrac{\sum_{i=1}^{n} \hat{y}_i^2 / (k-1)}{\sum_{i=1}^{n} e_i^2 / (n-k)} > F_{k-1,\, n-k,\, \alpha}$

THEN WE REJECT $H_0$ : X CANNOT EXPLAIN VARIATION IN Y.

NOTE THAT

$H_0$ : X CANNOT EXPLAIN VARIATION IN Y

$H_1$ : OTHERWISE.

FROM $\hat{F} = \dfrac{\sum_{i=1}^{n} \hat{y}_i^2 / (k-1)}{\sum_{i=1}^{n} e_i^2 / (n-k)} \div \dfrac{\sum_{i=1}^{n} y_i^2}{\sum_{i=1}^{n} y_i^2}$

$$= \frac{\sum_{i=1}^{n} \hat{y}_i^2 / \sum_{i=1}^{n} y_i^2}{\sum_{i=1}^{n} e_i^2 / \sum_{i=1}^{n} y_i^2} \cdot \frac{(n-k)}{(k-1)}$$

($\sum \hat{y}_i^2 / \sum y_i^2 = R^2$, $\sum e_i^2 / \sum y_i^2 = 1 - R^2$)

SINCE $\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2$

$$\frac{\sum_{i=1}^{n} y_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} y_i^2} + \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2}$$

$$1 = R^2 + \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2}$$

$\therefore$ $\hat{F} = \dfrac{R^2}{\quad^2} \cdot \dfrac{(n-k)}{(k-1)}$

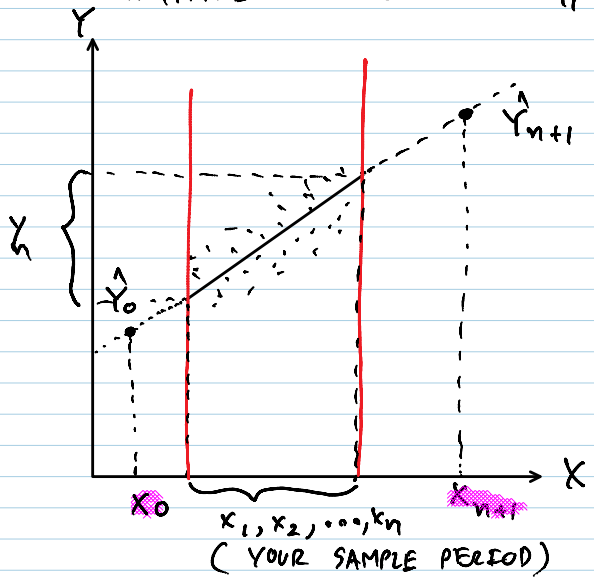
$$\therefore \quad F = \frac{R^2}{1 - R^2} \cdot \frac{(n-k)}{(k-1)}$$

MEANS THAT IF $R^2$ (GOODNESS OF FIT) IS HIGH, THE MORE LIKELIHOOD THAT $\hat{F}$ IS HIGH AND THAT WE REJECT $H_0$.

## PREDICTION

MODEL : $Y_i = a + b X_i + u_i$

ESTIMATED MODEL : $\hat{Y}_i = \hat{a} + \hat{b} X_i$




IF WE WOULD LIKE TO CONSIDER THE VALUE OF $X_0$ OR $X_{n+1}$, OUTSIDE THE SAMPLE PERIOD, WE WANT TO PREDICT/ESTIMATE THE VALUE OF $\hat{Y}_0$ AND $\hat{Y}_{n+1}$, RESPECTIVELY.

EXTENSION = (IMPLICIT ASSUMPTION) THAT THE RELATIONSHIP IS STABLE.

MODEL : $Y_i = a + b X_i + u_i \Rightarrow$ OLS ESTIMATION : $\hat{a}, \hat{b}$

GIVEN $X_{n+1}$, THE VALUE OF X WHICH LIES OUTSIDE THE SAMPLE PERIOD, WE WOULD LIKE TO PREDICT THE MEAN VALUE OF Y ASSOCIATED WITH $X_{n+1}$, WHICH IS $\hat{Y}_{n+1}$

THEREFORE, $E(Y_{n+1} \mid X_{n+1}) = a + b X_{n+1}$

$\because E(u_{n+1} \mid X_{n+1}) = 0$

WE WANT TO FIND THE BEST LINEAR UNBIASED PREDICTOR FOR $E(Y_{n+1} \mid X_{n+1})$ ← mean value of Y

**PROPOSITION** THE BEST LINEAR UNBIASED PREDICTOR FOR THE MEAN VALUE OF $Y_{n+1}$ GIVEN $X_{n+1}$, i.e., $E(Y_{n+1} \mid X_{n+1})$ IS ???

LET'S START TO DISCOVER A RESULT TO FILL IN THE STATEMENT ABOVE.

FIRST, DEFINE $\hat{Y}_{n+1}$ TO BE OUR PREDICTOR AS A LINEAR FUNCTION OF $Y_i$, THAT IS

$$\hat{Y}_{n+1} = \sum_{i=1}^{n} c_i Y_i$$

($n$ → TO $n$ ONLY (SAMPLE PERIOD))

WHERE $c_i$ = WEIGHTS TO BE CHOSEN SO AS TO "MAKE" $\hat{Y}_{n+1}$ A BEST LINEAR UNBIASED PREDICTOR.

START:

FROM THE MODEL $Y_i = a + b X_i + u_i$ AND GIVEN $E(Y_{n+1} \mid X_{n+1}) = a + b X_{n+1}$,

$$\hat{Y}_{n+1} = \sum_{i=1}^{n} c_i Y_i = \sum_{i=1}^{n} c_i (a + b X_i + u_i)$$

$$\hat{Y}_{n+1} = a \sum_{i=1}^{n} c_i + b \sum_{i=1}^{n} c_i X_i + \sum_{i=1}^{n} c_i u_i$$

$E\left(\sum_{i=1}^{n} c_i u_i\right) = 0$

$$E(\hat{Y}_{n+1}) = a + b X_{n+1}$$

IFF $\sum_{i=1}^{n} c_i = 1$

AND $\sum_{i=1}^{n} c_i X_i = X_{n+1}$

UNBIASED CONDITIONS

SO FAR, .... $\hat{Y}_{n+1} = \sum_{i=1}^{n} c_i Y_i$ ⇒ LINEAR

$\sum_{i=1}^{n} c_i = 1$

$\sum_{i=1}^{n} c_i X_i = X_{n+1}$

} UNBIASED CONDITIONS

LAST THING TO ENSURE THAT $\hat{Y}_{n+1}$ IS THE BEST IS

$\hat{Y}_{n+1}$ MUST GIVE THE LOWEST VARIANCE, i.e.,

$VAR(\hat{Y}_{n+1})$

Q: WHAT IS $VAR(\hat{Y}_{n+1})$ THEN ?

A: $$VAR(\hat{Y}_{n+1}) = E\left[\underbrace{\hat{Y}_{n+1}}_{\text{I}} - \underbrace{E(\hat{Y}_{n+1})}_{\text{II}}\right]^2$$

AND $$\text{I}\quad \hat{Y}_{n+1} = \sum_{i=1}^{n} c_i Y_i = a\sum_{i=1}^{n} c_i + b\sum_{i=1}^{n} c_i x_i + \sum_{i=1}^{n} c_i u_i$$

$$\text{II}\quad E(\hat{Y}_{n+1}) = a\sum_{i=1}^{n} c_i + b\sum_{i=1}^{n} c_i x_i$$

$$VAR(\hat{Y}_{n+1}) = E\left[\sum_{i=1}^{n} c_i u_i\right]^2 .$$

$$VAR(\hat{Y}_{n+1}) = E\left[c_1u_1 + c_2u_2 + \dots + c_nu_n\right]\left[c_1u_1 + c_2u_2 + \dots + c_nu_n\right]$$

$$= E\left[c_1^2u_1^2 + c_2^2u_2^2 + \dots + c_n^2u_n^2 + \underbrace{2c_iu_j\,c_iu_j + \dots}_{2E[c_iu_i\,c_ju_j]=0}\right]$$

$$= c_1^2E(u_1^2) + c_2^2E(u_2^2) + \dots + c_n^2E(u_n^2)$$

$$= c_1^2\sigma_u^2 + c_2^2\sigma_u^2 + \dots + c_n^2\sigma_u^2$$

$$= \sigma_u^2\sum_{i=1}^{n} c_i^2 .$$

MIN $\sigma_u^2\sum_{i=1}^{n} c_i^2$ S.T. $\sum_{i=1}^{n} c_i = 1$ AND $\sum_{i=1}^{n} c_i x_i = x_{n+1}$