ECONOMETRICS-I Lecture Notes

Chapters 4 & 5

Confidence interval & Hypothesis Testing:

Basic Concepts;

Remember, if $X \sim Normal(\mu; \sigma^2)$

then $\bar{X} \sim Normal\left(\mu; \frac{\sigma^2}{n}\right)$

where $\bar{X}$ is the sample mean, which is an estimator for population mean: μ. (in fact, $\bar{X}$ is BLUE)

	Population Parameters		Sample statistics
MEAN	μ	⟸ INFERENCE (i) Confidence Interval	$\bar{X}$
VARIANCE	σ^2	(ii) Hypothesis Testing	s^2

* $(1-\alpha).100\%$ confidence interval for μ is;
(For small samples ($n<30$) and σ^2 unknown)

$$\bar{X} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

* Hypothesis testing steps;

(i) H_0, H_A and α → State what to test.
(ii) Test statistics → which table? what is the formula?
(iii) Decision Criteria → When to "Reject H_0"
(iv) Calculation → Calculate Test statistics
(v) Decision and Conclusion → "Reject H_0" OR "Do NOT Reject H_0"

Alp Giray Özen | 0533 549 91 08 | alp@lecturemania.com | www.lecturemania.com

Question_1: Most water treatment facilities monitor the quality of their drinking water on hourly basis. One variable monitored it is pH, which measures the degree of alkalinity or acidity in the water. A pH below 7.0 is acidic, one above 7.0 is alkaline, and a pH of 7.0 is neutral. One water treatment plant has a target pH of 8.5 (most try to maintain a slightly alkaline level). The mean and Standard deviation of 1 hour's test results, based on 17 water samples at this plant are: $\bar{x} = 8.24$ and $s = 0.16$

a. Does this sample provide sufficient evidence that the mean pH level in the water differs from 8.5? (Use 5 % significance level.)

b. Find a 95% confidence interval for the mean pH level. What conclusion can you draw for the relationship between a two sided test and confidence interval?

Question_2: A major car manufacturer wants to test a new engine to determine whether it meets new air-pollution standards. The mean emission μ for all engines of this type must be less than 20 parts per million of carbon. 10 engines are manufactured for testing purposes, and the emission level for each is determined.

The mean and standard deviation for the tests are: $\bar{x} = 17.17$ and $s = 2.98$ Do the data supply enough evidence to allow the manufacturer to conclude that this type of engine meets the pollution standard? Assume the manufacturer is willing to risk a Type I error with probability $\alpha = 0.01$.

Answer-1// $n = 17$; $\bar{X} = 8,24$; $S = 0,16$

a) Hypothesis Testing ⇒ YES / NO Questions

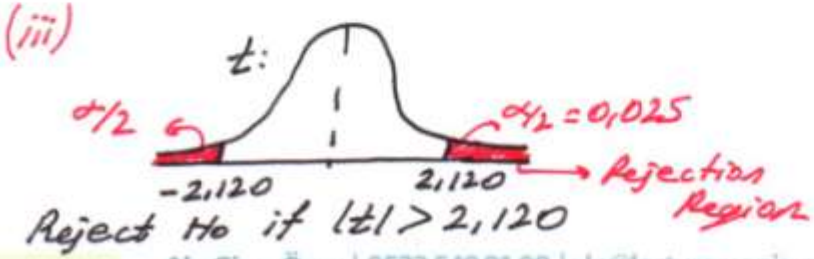
(i) $H_0 = 8,5$ → H_0 and H_A are complementary

$H_A \neq 8,5$ → Inequality is always at H_A

$\alpha = 0,05$: Significance level / Type I Error

(ii) $t = \dfrac{\bar{X} - \mu}{S/\sqrt{n}}$: $df = n - 1 = 17 - 1 = 16$

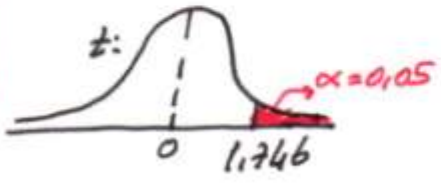
(iii)



Reject H_0 if $|t| > 2,120$

* Note that, Rejection Region has the same side with H_A. Consider the following tests; with df = 16

$H_0: \mu \leq 100$
$H_A: \mu > 100$
$\alpha = 0,05$

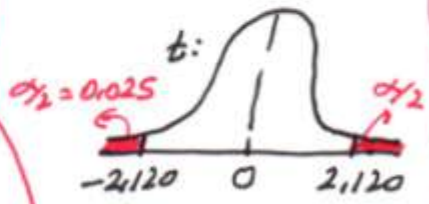


Reject H_0 if $t > 1,746$

$H_0: \mu \geq 70$
$H_A: \mu < 70$
$\alpha = 0,05$

Reject H_0 if $t < -1,746$

(This is called a two sided test)
$H_0: \mu = 18$
$H_A: \mu \neq 18$
$\alpha = 0,05$



Reject H_0 if $|t| > 2,120$

(iv) $$t = \frac{8,24 - 8,5}{0,16/\sqrt{17}} = -6,7$$

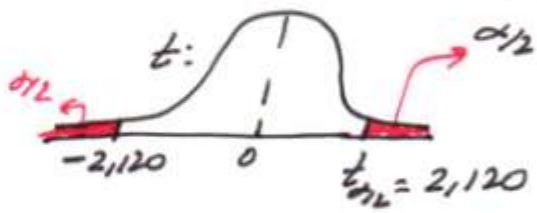
(v) $|-6,7| > 2,120$ so we Reject H_0. Mean pH level in the water significantly differs from 8,5 at $\alpha = 0,05$.

b) Confidence Interval:

95% C.I. for μ is;

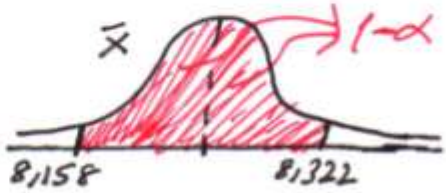
$$\bar{X} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

$1-\alpha = 0,95$
$\alpha = 0,05$
$\alpha/2 = 0,025$



$$8,24 \pm 2,120 \cdot \frac{0,16}{\sqrt{17}}$$

(8,158 ; 8,322)



Note that; $H_0: \mu = 8,5$ is NOT in the 95% C.I. So, we reject H_0 at 5% significance level.

Answer 2/f $n=10;\ \bar{X}=17,17;\ s=2,98;\ s_{\bar{x}}=\frac{s}{\sqrt{n}}=\frac{2.98}{\sqrt{10}}=0,942$

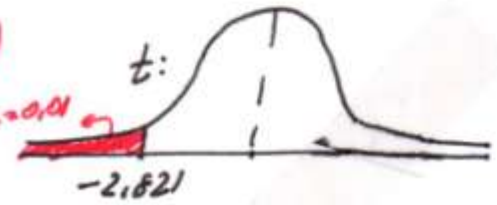
(i) $H_0: \mu \geq 20$

$H_A: \mu < 20$

$\alpha = 0,01$

(ii) $t=\frac{\bar{X}-\mu}{s_{\bar{x}}}$; $df=10-1=9$

(iii)



Reject H_0 if $t < -2,821$

(iv) $t=\frac{17,17-20}{0,962}=-3,004$

(v) Reject H_0. The engine significantly meets the standards at $\alpha=0,01$

NORMALITY ASSUMPTION for u_i:

$$Y_i = \beta_0 + \beta_1 . X_i + u_i$$

Normality of u_i is important because if u_i does NOT follow a Normal Distribution, we can NOT construct confidence intervals or make Hypothesis Testing using t or F tables. Under CLRM, we assume;

$$u_i \sim \underbrace{NID}_{\text{Normally and independently Distributed.}}\left(\underbrace{0}_{=E(u_i)};\ \underbrace{\sigma^2}_{=Var(u_i)}\right)$$

* we assume the Residuals Have;

(i) Normal Distribution (iii) Constant Variance $\sigma^2 = Var(u_i)$

(ii) Zero Mean: $E(u_i)=0$ (iv) Independence $\Rightarrow$ Uncorrelated $E(u_i u_j)=0$ for $i \neq j$

* Linear combinations of • Normal Random Variables has also a Normal Distribution.

• Since $Y_i = \beta_1 + \beta_2 X_i + u_i$

$$E(Y_i) = E(\beta_1 + \beta_2 X_i + u_i) = \beta_1 + \beta_2 X_i + E(u_i) = \beta_1 + \beta_2 X_i$$

$$Var(Y_i) = Var(\beta_1 + \beta_2 X_i + u_i) = Var(u_i) = \sigma^2$$

$$Y_i \sim Normal(\beta_1 + \beta_2 X_i \; ; \sigma^2)$$

• Since $\hat{\beta}_1$ and $\hat{\beta}_2$ are linear in Y_i;

$$E(\hat{\beta}_1) = \beta_1 \quad \text{and} \quad Var(\hat{\beta}_1) = \sigma_{\hat{\beta}_1}^2 = \frac{\sum X_i^2}{n \cdot \sum x_i^2} \cdot \sigma^2$$

$$\hat{\beta}_1 \sim Normal(\beta_1 \; ; \sigma_{\hat{\beta}_1}^2)$$

$$E(\hat{\beta}_2) = \beta_2 \quad \text{and} \quad Var(\hat{\beta}_2) = \sigma_{\hat{\beta}_2}^2 = \frac{\sigma^2}{\sum x_i^2}$$

$$\hat{\beta}_2 \sim Normal(\beta_2 \; ; \sigma_{\hat{\beta}_2}^2)$$

Maximum Likelihood Estimation

So far, we have obtained least Square Estimators of β_1 and β_2, which are $\hat{\beta}_1$ and $\hat{\beta}_2$. Maximum Likelihood Estimation (MLE) is another estimation technique, which is based on distribution of sample data.

Note that, we didn't make any distribution assumption for OLS estimators. We'll see how to find MLE's and obtain $\tilde{\beta}_1$ and $\tilde{\beta}_2$

* To obtain Maximum Likelihood Estimator(s), follow:

(i) Obtain Likelihood Function:

$$LF(\theta) = f(x_1).f(x_2).\ldots.f(x_n)$$

(ii) Take ln of Likelihood function

(iii) Take partial derivative(s) and equate to zero:

$$\frac{\partial \ln LF(\theta)}{\partial \theta} = 0$$

and the solution is: $\hat{\theta}_{MLE}\ (= \tilde{\theta})$

4.3. A random variable X follows the **exponential distribution** if it has the following probability density function (PDF):

$$f(X) = (1/\theta)e^{-X/\theta} \quad \text{for } X > 0$$
$$= 0 \quad \text{elsewhere}$$

where $\theta > 0$ is the parameter of the distribution. Using the ML method, show that the ML estimator of θ is $\hat{\theta} = \sum X_i/n$, where n is the sample size. That is, show that the ML estimator of θ is the sample mean $\bar{X}$.

4.3) $LF(\theta) = f(x_1).f(x_2).\ldots.f(x_n)$

$$= \left(\frac{1}{\theta}\right).e^{-x_1/\theta}.\frac{1}{\theta}.e^{-x_2/\theta}.\ldots.\frac{1}{\theta}.e^{-x_n/\theta}$$

$$= \left(\frac{1}{\theta}\right)^n.e^{-\sum x_i/\theta}$$

$$\ln LF(\theta) = \ln\left[\left(\frac{1}{\theta}\right)^n.e^{-\frac{1}{\theta}.\sum x_i}\right] = n.\ln\left(\frac{1}{\theta}\right) - \frac{1}{\theta}.\sum x_i$$

$$\ln LF(\theta) = -n.\ln(\theta) - \frac{\sum x_i}{\theta}$$

$$\frac{\partial \ln LF(\theta)}{\partial \theta} = -n.\frac{1}{\theta} + \frac{\sum x_i}{\theta^2} = 0 \Rightarrow \frac{\sum x_i}{\theta^2} = \frac{n}{\theta}$$

$$\hat{\theta}_{MLE} = \frac{\sum x_i}{n} = \bar{X}$$

Maximum Likelihood Estimators of β_1 and β_2:

Normal Distribution: $W \sim Normal(\mu; \sigma^2)$

$$f(w) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left\{-\frac{1}{2} \cdot \frac{(w-\mu)^2}{\sigma^2}\right\}$$

Remember; $Y_i \sim Normal(\beta_1 + \beta_2 X_i ; \sigma^2)$

$$f(Y_i) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \exp\left\{-\frac{1}{2} \cdot \frac{(Y_i - \beta_1 - \beta_2 \cdot X_i)^2}{\sigma^2}\right\}$$

$$LF(\beta_1, \beta_2, \sigma^2) = f(y_1) \cdot f(y_2) \cdot \ldots \cdot f(y_n)$$

$$= \frac{1}{\sigma^n \cdot (\sqrt{2\pi})^n} \cdot \exp\left\{-\frac{1}{2\sigma^2} \sum (y_i - \beta_1 - \beta_2 X_i)^2\right\}$$

$$\ln LF(\beta_1, \beta_2, \sigma^2) = -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2} \sum \underbrace{(y_i - \beta_1 - \beta_2 X_i)^2}_{= u_i}$$

We have;

① $$\frac{\partial \ln LF}{\partial \beta_1} = -\frac{1}{\sigma^2} \sum (y_i - \beta_1 - \beta_2 \cdot X_i) \cdot (-1) = 0$$

② $$\frac{\partial \ln LF}{\partial \beta_2} = -\frac{1}{\sigma^2} \cdot \sum (y_i - \beta_1 - \beta_2 \cdot X_i) \cdot (-X_i) = 0$$

③ $$\frac{\partial \ln LF}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \cdot \sum u_i^2 = 0$$

Since ① and ② are the same with that of Normal Equations when we cancel out $\frac{1}{\sigma^2}$ terms, MLE's are the same with OLS estimators:

$$\tilde{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} \quad \text{and} \quad \tilde{\beta}_1 = \bar{y} - \tilde{\beta}_2 \cdot \bar{X}$$

$\left(\tilde{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n}\right.$ However, $\left.\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2}\right)$

Formula Sheet

(i) $\Sigma X_i \quad \Sigma y_i \quad \Sigma X_i^2 \quad \Sigma y_i^2 \quad \Sigma X_i y_i$

(ii) $x_i = X_i - \bar{X} \quad y_i = y_i - \bar{y} \quad \bar{X} = \frac{\Sigma X_i}{n} \quad \bar{y} = \frac{\Sigma y_i}{n}$

$$\Sigma x_i^2 = \Sigma (X_i - \bar{X})^2 = \Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}$$

$$\Sigma y_i^2 = \Sigma (y_i - \bar{y})^2 = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}$$

$$\Sigma x_i y_i = \Sigma (X_i - \bar{X}).(y_i - \bar{y}) = \Sigma X_i y_i - \frac{\Sigma X_i \Sigma y_i}{n}$$

(iii) $\hat{\beta}_2 = \frac{\Sigma x_i y_i}{\Sigma x_i^2} \quad \hat{\beta}_1 = \bar{y} - \hat{\beta}_2 . \bar{X} \quad \hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 . X_i$

(iv) $TSS = \Sigma y_i^2 = \Sigma (y_i - \bar{y})^2$

$ESS = \hat{\beta}_2^2 . \Sigma x_i^2 = \Sigma (\hat{y}_i - \bar{y})^2$

$RSS = TSS - ESS = \Sigma (y_i - \hat{y}_i)^2 = \Sigma \hat{u}_i^2$

(v) $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$

$\hat{\sigma}^2 = \frac{RSS}{n-2} = \frac{\Sigma \hat{u}_i^2}{(n-2)}$ ← df

Test stat: $\chi^2 = \frac{(n-2).\hat{\sigma}^2}{\sigma^2}$ ⟺

(vi) $\hat{\sigma}_{\hat{\beta}_2}^2 = \frac{\hat{\sigma}^2}{\Sigma x_i^2} \quad \hat{\sigma}_{\hat{\beta}_1}^2 = \frac{\Sigma X_i^2}{n . \Sigma x_i^2} . \hat{\sigma}^2$

C.I. for σ^2: $\left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{\alpha/2}} ; \frac{(n-2)\hat{\sigma}^2}{\chi^2_{1-\alpha/2}} \right)$

$$\hat{\sigma}_{\hat{y}_0}^2 = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x_i^2} \right] : \text{Mean } y_0$$

$$\hat{\sigma}_{y_0 - \hat{y}_0}^2 = \hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x_i^2} \right] : \text{Individual } y_0$$

(vii) $t = \frac{\hat{\beta}_i - \beta_i}{\sigma_{\hat{\beta}_i}}$ → $SE(\hat{\beta}_i)$

CI: $\hat{\beta}_i \pm t_{\alpha/2, n-2} . \hat{\sigma}_{\hat{\beta}_i}$ { $\hat{y}_0 \pm t_{\alpha/2, n-2} . \hat{\sigma}_{\hat{y}_0}$ ind $\hat{y}_0 - \hat{y}_0$

"Give a man three weapons – correlation, regression and a pen – and he will use all three (Anon, 1978)"

Base Question: In a study of the age and cholesterol level, 18 samples are selected and their age and cholesterol level is recorded. The following paired data were reported on scaled time (x) versus proportion surviving (y)

ID	X:Age	Y:Chol	X*X	Y*Y	X*Y
1	46	3,5	2116	12,25	161
2	20	1,9	400	3,61	38
3	52	4	2704	16	208
4	30	2,6	900	6,76	78
5	57	4,5	3249	20,25	256,5
6	25	3	625	9	75
7	28	2,9	784	8,41	81,2
8	36	3,8	1296	14,44	136,8
9	22	2,1	484	4,41	46,2
10	43	3,8	1849	14,44	163,4
11	57	4,1	3249	16,81	233,7
12	33	3	1089	9	99
13	22	2,5	484	6,25	55
14	63	4,6	3969	21,16	289,8
15	40	3,2	1600	10,24	128
16	48	4,2	2304	17,64	201,6
17	28	2,3	784	5,29	64,4
18	49	4	2401	16	196
TOTAL	**699**	**60**	**30287**	**211,96**	**2511,6**
MEAN	**38,83**	**3,33**			

a. Plot the data.

b. Assuming a straight line regression compute the least squares estimates of the slope and the intercept. Interpret the slope.

c. Estimate the cholesterol of a person at age of 30.

d. What percentage of the total variation in cholesterol can be explained by age?

e. Estimate the (true) effect of cholesterol change by getting one year older at 95% confidence.

f. Compute the residual sum of squares and give an unbiased estimate of $\sigma^2 = Var(Y)$ also find a 80% confidence interval for error variance.

g. Does cholesterol changes significantly with respect to age?

h. Test the validity of the model using an F test.

i. Test if cholesterol increases more than 0.03 units when a person gets one year older. Use 1% significance level.

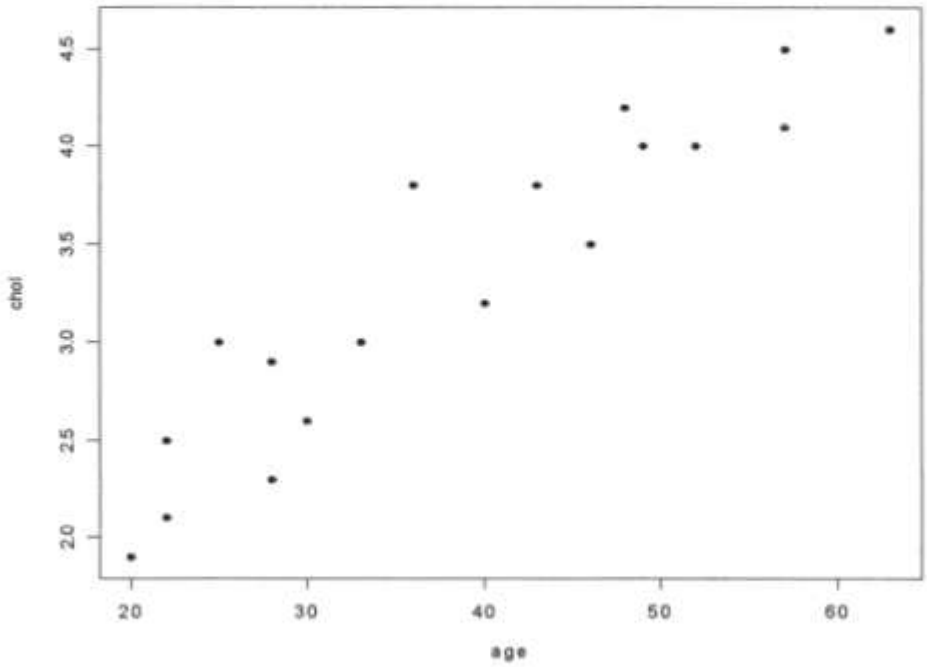
j. Can we conclude that the error variance of the model is less than 0.15?

k. Find a 95% confidence interval for the mean cholesterol of all people at age 30. (prediction)

l. Sarah is a woman whose age is 30. Find 95% confidence interval for her cholesterol.

Answer:

a. X: Age versus Y: Cholesterol data is plotted in the following figure. We can follow the linear pattern.



b) (i) $\Sigma X_i = 699$ $\Sigma y_i = 60$ $\Sigma X_i y_i = 2511,6$

$\Sigma X_i^2 = 30287$ $\Sigma y_i^2 = 211,96$ $n = 18$

(ii) $\bar{X} = \frac{\Sigma X_i}{n} = \frac{699}{18} = 38,83$ $\bar{y} = \frac{\Sigma y_i}{n} = \frac{60}{18} = 3,33$

$$\Sigma x_i^2 = \Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n} = 30287 - \frac{699^2}{18} = 3142,5$$

$$\Sigma y_i^2 = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n} = 211,96 - \frac{60^2}{18} = 12$$

$$\Sigma x_i y_i = \Sigma X_i y_i - \frac{\Sigma X_i \Sigma y_i}{n} = 2511,6 - \frac{699 \cdot 60}{18} = 181,6$$

(iii) $\hat{\beta}_2 = \frac{\Sigma x_i y_i}{\Sigma x_i^2} = \frac{181,6}{3142,5} = 0,0578$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \cdot \bar{X} = 3,33 - 0,0578 \cdot 38,33 = 1,11$$

$$\hat{y}_i = 1,11 + 0,0578 \cdot X_i$$

c) $(\hat{y}_i | X_i = 30) = 1,11 + 0,0578 \cdot 30 = 2,85$

d) (iv) $TSS = \Sigma y_i^2 = 12$

$$ESS = \hat{\beta}_2^2 \cdot \Sigma x_i^2 = 0,0578^2 \cdot 30287 = 10,5$$

$$RSS = TSS - ESS = 12 - 10,5 = 1,5$$

(v) R^2: Coefficient of Determination

$$R^2 = \frac{ESS}{TSS} = \frac{10,5}{12} = 0,875$$

87,5% can be explained

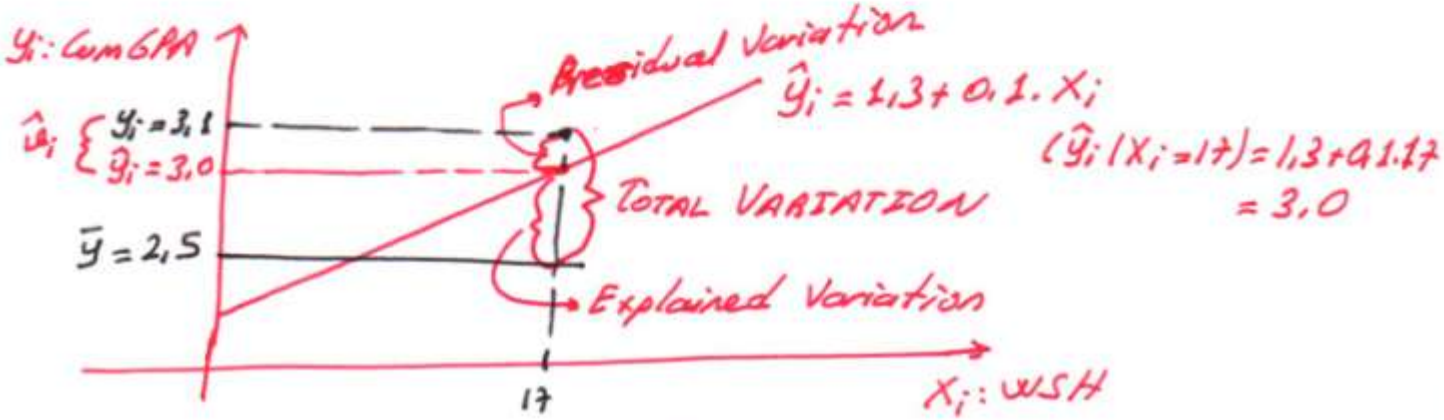
$$\hat{\sigma}^2 = \frac{RSS}{n-2} = \frac{1,5}{16} = 0,094$$

Coefficient of Determination

Example: X_i : Weekly Studying Hours (WSH)

Y_i : Cummulative GPA (Cum. GPA

X_i	8	9	9	11	---	17	---	25	
Y_i	1,9	2,3	2,1	2,4	---	3,1	---	3,8	$\bar{Y}=2,5$



$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

$$\Sigma (y_i - \bar{y})^2 = \Sigma (\hat{y}_i - \bar{y})^2 + \Sigma (y_i - \hat{y}_i)^2$$

$$TSS = ESS + RSS$$

$\Rightarrow$ R^2: Percentage of the explained part

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

INTERVAL ESTIMATION

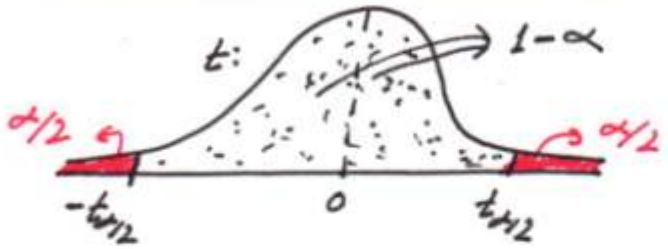
Let W be a statistics (Like $\hat{\beta}_i$)

Also let $W \sim Normal(\mu; \sigma_w^2)$

* If σ_w^2 is known; $Z = \frac{W-\mu}{\sigma_w} \sim Normal(0;1)$

* If σ_w^2 is unknown and estimated from the data; (which is our case); $t = \frac{W-\mu}{\hat{\sigma}_w} \sim t(df)$ → $df = n-2$ for Simple Reg.

$\hat{\sigma}_w$ → OR $\widehat{SE}(W)$ ← Standard Error



$$P(-t_{\alpha/2} < t < t_{\alpha/2}) = 1-\alpha$$

$$-t_{\alpha/2} < \frac{W-\mu}{\sigma_W} < t_{\alpha/2}$$

$$-t_{\alpha/2}\cdot\sigma_W < W-\mu < \sigma_W \cdot t_{\alpha/2}$$

$$P(W - t_{\alpha/2}\cdot\sigma_W < \mu < W + t_{\alpha/2}\cdot\sigma_W) = 1-\alpha$$

The interval $(W - t_{\alpha/2}\cdot\sigma_W ; W + t_{\alpha/2}\cdot\sigma_W)$ is called $(1-\alpha).100\%$ Confidence Interval for μ

We have;

$$\hat{\beta}_1 \sim Normal\left(\beta_1 ; \underbrace{\frac{\sum X_i^2}{n\cdot\sum x_i^2}\cdot\sigma^2}_{\sigma_{\hat{\beta}_1}^2}\right) \text{ and } \hat{\beta}_2 \sim Normal\left(\beta_2 ; \underbrace{\frac{\sigma^2}{\sum x_i^2}}_{\sigma_{\hat{\beta}_2}^2}\right)$$

So, $(1-\alpha).100\%$ Confidence intervals for β_1 and β_2 are;

$$\hat{\beta}_1 \pm t_{\alpha/2}\cdot\hat{\sigma}_{\hat{\beta}_1} \quad \text{and} \quad \hat{\beta}_2 \pm t_{\alpha/2}\cdot\hat{\sigma}_{\hat{\beta}_2}$$

respectively. (df = degrees of freedom $= n-2$)

Note that, general degrees of freedom is $n-k$ where k: Number of β's in the model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki}$$

e) Remember; β_2 = Slope: The effect of a unit increase in X to Y.

Then, we want 95% C.I. for β_2.

$\hat{y}_i = \underbrace{1,11}_{\hat{\beta}_1} + \underbrace{0,0578}_{\hat{\beta}_2} . X_i$

(vi) $\hat{\sigma}^2_{\hat{\beta}_2} = \frac{\hat{\sigma}^2}{\sum x_i^2} = \frac{0,094}{3142,5} = 3.10^{-5}$

$\hat{\sigma}^2 = \frac{RSS}{n-2} = \frac{1,5}{16} = 0,094$

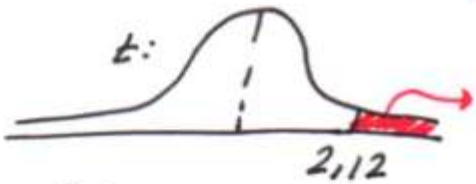
$SE(\hat{\beta}_2) = \hat{\sigma}_{\hat{\beta}_2} = \sqrt{3.10^{-5}} = 0,0055$

$df = n-2 = 18-12 = 16$

$1-\alpha = 0,95$

$\alpha = 0,05$

$\alpha/2 = 0,025$



95% C.I. for β_2 is;

$$\hat{\beta}_2 \pm t_{\alpha/2} . SE(\hat{\beta}_2)$$

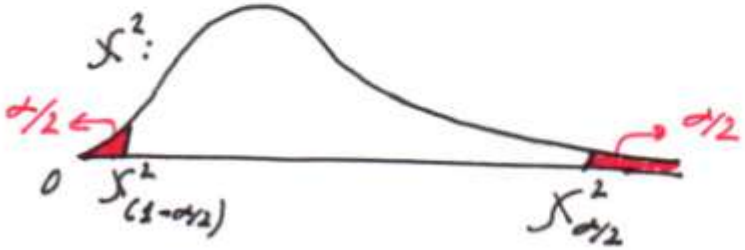
$$0,0578 \pm 2,12 . 0,0055$$

$$(0,566 ; 0,590)$$

Confidence interval for σ^2: Error Variance

C.I. for σ^2 is found by using χ^2 distribution

$$W = \frac{(n-2).\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-2)}$$



$$P\left(\chi^2_{1-\alpha/2} < W < \chi^2_{\alpha/2}\right) = 1-\alpha$$

$$\chi^2_{1-\alpha/2} < \frac{(n-2).\hat{\sigma}^2}{\sigma^2} < \chi^2_{\alpha/2}$$

$$\frac{1}{\chi^2_{\alpha/2}} < \frac{\sigma^2}{(n-2)\hat{\sigma}^2} < \frac{1}{\chi^2_{1-\alpha/2}}$$

$$P\left(\frac{(n-2).\hat{\sigma}^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-2).\hat{\sigma}^2}{\chi^2_{(1-\alpha/2)}}\right) = 1-\alpha$$

So, $(1-\alpha).100\%$ C.I. for σ^2 is;

$$\left(\frac{(n-2).\hat{\sigma}^2}{\chi^2_{\alpha/2}} \; ; \; \frac{(n-2).\hat{\sigma}^2}{\chi^2_{(1-\alpha/2)}}\right) \qquad df = n-2$$

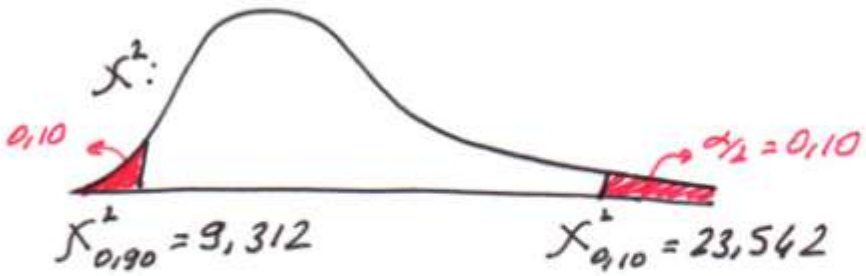
f) $RSS = 1,5$; $\hat{\sigma}^2 = \frac{RSS}{n-2} = \frac{1,5}{16} = 0,094$; $1-\alpha = 0,80$

$\alpha = 0,20$

$\alpha/2 = 0,10$

$1-\alpha/2 = 0,90$

$df = n-2 = 16$



80% C.I. for σ^2 is;

$$\left(\frac{16.\ 0,094}{23,542} \; ; \; \frac{16.0,094}{9,312}\right)$$

$$(0,0639 \; ; \; 0,162)$$

Hypothesis Testing

$$y_i = \beta_0 + \beta_2 . X_i + u_i$$

$$E(y_i) = \beta_0 + \beta_2 . X_i$$

Remember, β_2: Estimated effect of a Unit Increase in X to Y.

So, if $\beta_2 = 0$, X has NO effect on Y.

To answer the questions: "Is the model Valid?" and "Is X a significant variable for X?" or "Does a unit increase in X changes the Value of Y?" We make the hypothesis testing

(i) H_0, H_A and α

$H_0 : \beta_2 = 0$

$H_A : \beta_2 \neq 0$

$\alpha = \alpha_0$ → if NOT given, take $\alpha = 0,05$

This can be done either using a t-test or an F-test. Note that, specifically, we use

F-test for

$H_0 : \beta_2 = \beta_3 = \dots = \beta_k = 0$ (Multiple Regression)

H_A: At least one β_j is NON-zero

and t-test for

$H_0 : \beta_j = c$ OR $H_0 : \beta_j \leq c$ OR $H_0 : \beta_j \geq c$

$H_A : \beta_j \neq c$ $H_A : \beta_j > c$ $H_A : \beta_j < c$

Namely, F-test is used for overall significance and t-test is used for individual comparison

(ii) Test Statistics

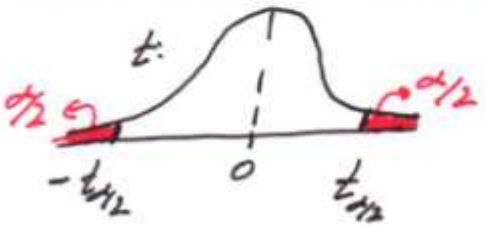
$$t = \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)}$$ → we call it c

OR

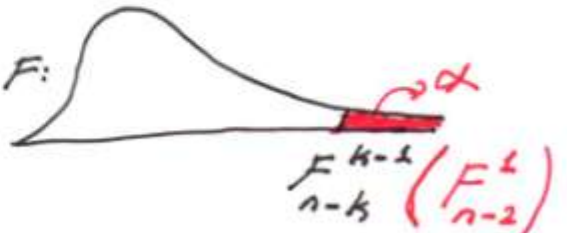
$$F = \frac{MSS \text{ of } ESS}{MSS \text{ of } RSS}$$

* We'll find value of F from ANOVA table

(iii) Decision CRITERIA



OR



df = n-k (here, n-2)

Reject Ho if $|t| > t_{\alpha/2}$

Reject Ho if $F > F^{k-1}_{n-k;\alpha}$

* Note that for one sided tests (we can only use t-statistics), one side is shaded by α

Remember the last two steps of Hypothesis testing:

(iv) Calculation

Calculate test statistics

Calculate t OR Make ANOVA table to find F

(v) Decision & Conclusion

We decide one of the following;

Reject Ho. The model is significant at α

OR

Do NOT Reject Ho. The model is NOT significant at α.

* We write α again since the test result is based on α.

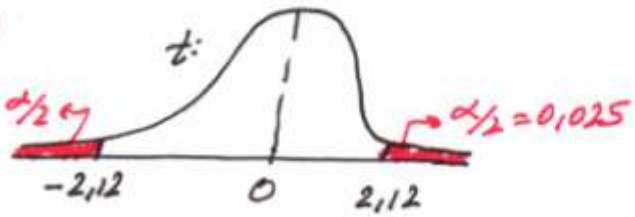
g) (i) $H_0: \beta_2 = 0$

$H_A: \beta_2 \neq 0$

$\alpha = 0,05$

(ii) $t = \dfrac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)}$; $df = n-2 = 16$

(iii)



Reject H_0 if $|t| > 2,12$

(iv) $\hat{\beta}_2 = 0,0578$

$SE(\hat{\beta}_2) = 0,0055$ (p.35)

$t = \dfrac{0,0578 - 0}{0,0055} = 10,51$

(v) $|10,51| > 2,12$, Reject H_0.

Age significantly changes cholesterol at $\alpha = 0,05$

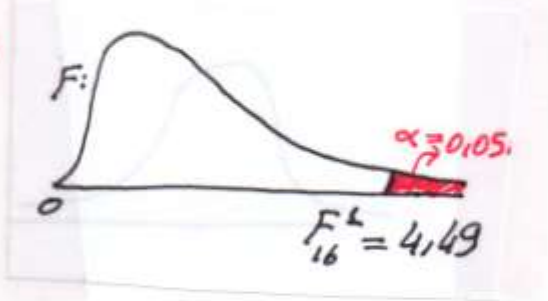
h) (i) $H_0: \beta_2 = 0$

$H_A: \beta_2 \neq 0$

$\alpha = 0,05$

(ii) $F = \dfrac{MS(ESS)}{MS(RSS)}$ → $df = k-1 = 1$; → $df = n-2 = 16$

(iii)



Reject H_0 if $F > 4,49$

(iv) $TSS = 12$ p.32

$ESS = 10,5$ $RSS = 1,5$

ANOVA TABLE:

Source	df	SS	MS	F
Regression	$k-1 = 1$	$ESS = 10,5$	$\frac{10,5}{1} = 10,5$	$\frac{10,5}{0,094} = 112$
Residuals	$n-k = 16$	$RSS = 1,5$	$\frac{1,5}{16} = 0,094$	—
TOTAL	$n-1 = 17$	$12 = TSS$	—	—

(v) $112 > 4,49$ Reject H_0.

Model is Valid at $\alpha = 0,05$

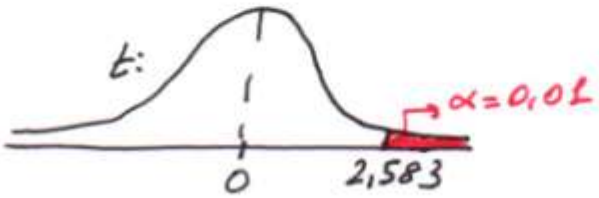
i) (i) $H_0: \beta_2 \leq 0,03$

$H_A: \beta_2 > 0,03$

$\alpha = 0,01$

(ii) $t = \dfrac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)}$; $df = n-2 = 16$

(iii)



Reject H_0 if ~~possible~~ $t > 2,583$

(iv) $t = \dfrac{0,0578 - 0,03}{0,0055}$

$t = 5,05$

(v) $5,05 > 2,583$

Reject H_0.

Cholesterol increase is significantly greater than 0,03 by getting one year older at $\alpha = 0,01$.

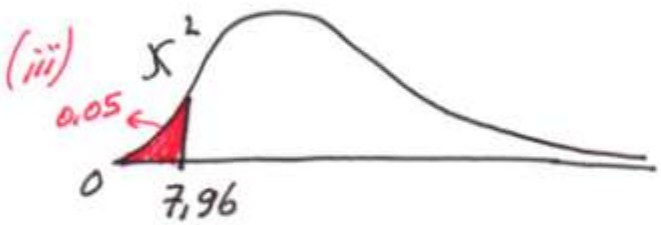
j) (i) $H_0: \sigma^2 \geq 0,15$

$H_A: \sigma^2 < 0,15$

$\alpha = 0,05$

(ii) $\chi^2 = \dfrac{(n-2).\hat{\sigma}^2}{\sigma^2}$; $df = n-2 = 16$

(iii)



Reject H_0 if $\chi^2 < 7,96$

(iv) $\chi^2 = \dfrac{16 . 0,094}{0,15} = 10,03$

(v) $10,03 > 7,96$,

Do NOT Reject H_0.

Error Variance is NOT significantly less than 0,15 at $\alpha = 0,05$

Prediction and Estimation:

Given the value of X_0, we can find a Confidence Interval for mean Y_0 (estimation) and also for individual Y_0 (prediction). We have;

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 \cdot X_0$$

$$Var(\hat{Y}_0) = \hat{\sigma}_{\hat{Y}_0}^2 = \hat{\sigma}^2 \cdot \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right] : \text{Mean } Y_0$$

$$Var(Y_0 - \hat{Y}_0) = \hat{\sigma}_{Y_0 - \hat{Y}_0}^2 = \hat{\sigma}^2 \cdot \left[(1+) \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right] : \text{Individual } Y_0$$

Then,

Mean Confidence interval for Y_0 is; $\hat{Y}_0 \pm t_{\alpha/2;n-2} \cdot \hat{\sigma}_{\hat{Y}_0}$

Individual Prediction interval for Y_0 is; $\hat{Y}_0 \pm t_{\alpha/2;n-2} \cdot \hat{\sigma}_{Y_0 - \hat{Y}_0}$

k) $1-\alpha = 95\%$

$n = 18 \quad \alpha/2 = 0,025$

$df = 16; \quad t_{\alpha/2} = 2,12$

$$\hat{Y}_0 = (\hat{Y} \mid X = 30) = 1,11 + 0,0578 \cdot 30 = 2,85$$

$\bar{X} = 38,83 \; ; \; \sum x_i^2 = 3162,5 \; ; \; \hat{\sigma}^2 = 0,094$ (p.32)

$$\hat{\sigma}_{\hat{Y}_0}^2 = 0,094 \cdot \left[\frac{1}{18} + \frac{(30 - 38,83)^2}{3162,5} \right] = 0,00755$$

95% C.I. for mean Y_0 is; $2,85 \pm 2,12 \cdot \sqrt{0,00755}$

$(1,90 \; ; \; 2,27)$

l) $\hat{\sigma}_{Y_0 - \hat{Y}_0}^2 = \left[1 + \frac{1}{18} + \frac{(30 - 38,83)^2}{3162,5} \right] = 1,08$; 95% P.I. for individual Y_0 is;

$2,85 \pm 2,12 \cdot \sqrt{1,08}$

$(0,65 \; ; \; 5,05)$ → wider

p-value of the test & Computer Output of the Regression Analysis

Question: XYZ Auto periodically has a special week-long sale. As part of the advertising campaign XYZ runs one or more television commercials during the weekend preceding the sale. The excel output of the regression analysis is given below:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,833333
R Square	0,694444
Adjusted R Square	0,592593
Standard Error	2,708013
Observations	5

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression (SSR)	1	50	50	6,818	0,079604981
Residual (SSE)	3	22	7,33333333		
Total (SST)	4	72			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	10	4,016632088	2,48964799	0,089	-2,78272794	22,7827279
ad	5	1,914854216	2,61116484	0,08	-1,09392644	11,0939264

a. Estimate the regression function.

b. What percentage of the total variation in sales can be explained by the advertisement?

c. Test the validity of the model at 5% and 10% significance using F test.

d. Does a unit increase in Advertisement increases Sales? Test at α=0.05.

e. Can we assume a model without an intercept term? Test at α=0.05.

f. What is the relationship between 95% confidence interval for the slope and validity of the model?

g. Find a 95% confidence interval for the estimation variance.

a) From Coefficients Column; SRF is;

$$\hat{y} = 10 + 5.X$$

S.E = (4,02) (1,91)

t = (2,49) (2,61)

p-value = (0,089) (0,08) $R^2 = 0,6944$

This is the usual report of the Results of Regression Analysis.

The first row S.E. shows the standard Errors; $SE(\hat{\beta}_1)$ and $SE(\hat{\beta}_2)$ respectively

The secant t row shows the calculated t-statistics of the individual significance tests

$H_0: \beta_1 = 0$ AND $H_0: \beta_2 = 0$ respectively
$H_A: \beta_1 \neq 0$ $H_A: \beta_2 \neq 0$

The third p-value row shows the p-values of the corresponding individual significance tests.

p-value of the test:

(iii) Decision Criteria: Reject H_0 if p-value $< \alpha$

Namely, p-value of the test is the minimum value of α to "Reject H_0".

b) $R^2 = 0,6944$

69,44% can be explained

c) $y_i = \beta_1 + \beta_2 X_i + u_i$

(i) $H_0: \beta_2 = 0$

$H_A: \beta_2 \neq 0$

$\alpha = \alpha_0$

(ii) $F = \dfrac{MS(Regression)}{MS(Residual)}$

(iii) Reject H_0 if p-value $< \alpha$

(iv) p-value = 0,0796 → Significance F

(v) $\alpha = 0,05$: Do NOT Reject H_0. Model is NOT valid at $\alpha = 0,05$

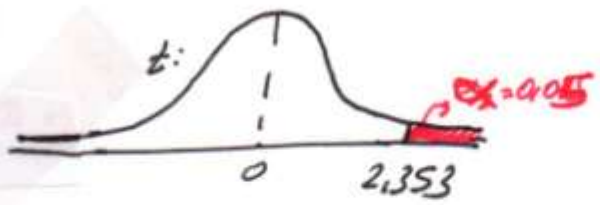
$\alpha = 0,10$: Reject H_0. Model is valid at $\alpha = 0,10$

d) (i) $H_0: \beta_2 \leq 0$

$H_A: \beta_2 > 0$

(ii) $t = \dfrac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)}$; $df = 3$

(iii)



Reject H_0 if $t > 2,353$

(iv) $t = \dfrac{5-0}{1,915} = 2,611$ → t-stat

(v) Reject H_0. A unit increase in Advertisement significantly increases Sales at $\alpha = 0,05$

*Note that this result is different from the result at two-sided test.

e) (i) $H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

$\alpha = 0,05$

(ii) $t = \dfrac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$; $df = 3$

(iii) Reject H_0 if p-value $< \alpha$

(iv) p-value = 0,089

(v) Do NOT Reject H_0. we may assume a model without an intercept term at $\alpha = 0,05$

f) 95% C.I. for β_2 is;

$(-2,78; 22,78)$

Since the 95% C.I. for β_2 contains 0, β_2 may be 0 at 95% confidence. Therefore, we do NOT Reject $H_0: \beta_2 = 0$ against a two sided alternative.

g) $\hat{\sigma}^2 = \frac{RSS}{n-2} = MS(Residual) = 7,33$; $df = 2$

95% C.I. for σ^2 is;

$$\left(\frac{(n-2).\hat{\sigma}^2}{X^2_{\alpha/2}} ; \frac{(n-2).\hat{\sigma}^2}{X^2_{1-\alpha/2}} \right)$$

$$\left(\frac{3.\ 7,33}{9,348} ; \frac{3.\ 7,33}{0,216} \right)$$

$$(2,352 ; 101,806)$$

5.3. From the data given in Table 2.6 on earnings and education, we obtained the following regression [see Eq. (3.7.3)]:

$$\widehat{\text{Meanwage}}_i = 0.7437 + 0.6416\ \text{Education}_i \qquad \Sigma x_i^2 = 182$$

$$\text{se} = (0.8355)\quad (\quad)(i)$$

$$t = (\quad)(ii)\quad (9.6536) \qquad r^2 = 0.8944 \qquad n = 13$$

a. Fill in the missing numbers.

b. How do you interpret the coefficient 0.6416?

c. Would you reject the hypothesis that education has no effect whatsoever on wages? Which test do you use? And why? What is the p value of your test statistic?

d. Set up the ANOVA table for this example and test the hypothesis that the slope coefficient is zero. Which test do you use and why?

a)(i) $t = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} \Rightarrow SE(\hat{\beta}_2) = \frac{\hat{\beta}_2}{t} = \frac{0,6416}{9,6536} = 0,0665$

(ii) $t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{0,7437}{0,8355} = 0,8901$

b) A year increase in Education is expected to increase Meanwage by 0,6416 units.

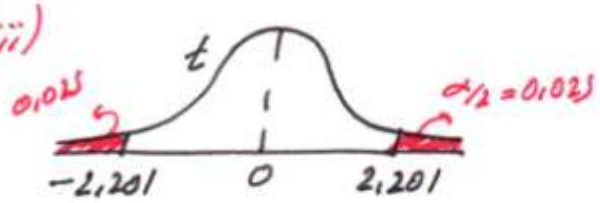
c) (i) $H_0: \beta_2 = 0$

$H_A: \beta_2 \neq 0$

$\alpha = 0,05$

(ii) $t = \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)}$; $df = 13 - 2 = 11$

(iii)

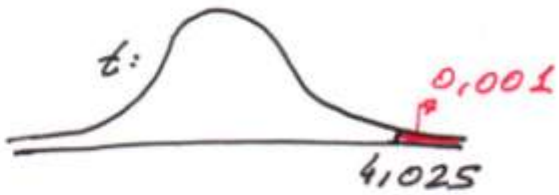


Reject H_0 if $|t| > 2,201$

(iv) $t = 9,6536$

(v) Reject H_0. Education is significant on Meanwage at $\alpha = 0,05$

*Note that;



Since $9,6536 > 4,025$

$\frac{p\text{-value}}{2} < 0,001$

$p\text{-value} < 0,002$

d) $ESS = \hat{\beta}_2^2 \cdot \Sigma x_i^2 = 182 \cdot 0,6416 = 116,77$

$r^2 = \frac{ESS}{TSS} \Rightarrow TSS = \frac{ESS}{r^2} = \frac{116,18}{0,8944} = 130,56$

$RSS = TSS - ESS = 130,56 - 116,77 = 13,79$

Source	df	SS	MS	F
Regression	1	116,77	$\frac{116,77}{1} = 116,77$	$\frac{116,77}{1,25} = 93,16$
Residual	12-1=11	13,79	$\frac{13,79}{11} = 1,25$	
TOTAL	n-1=12	130,56		

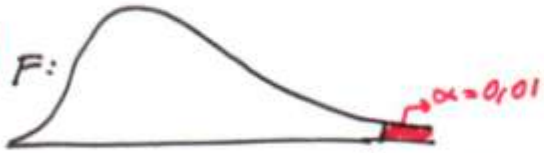
(i) $H_0: \beta_2 = 0$

$H_A: \beta_2 \neq 0$

$\alpha = 0,01$

(ii) $F = \frac{MS(\text{Regression})}{MS(\text{Error})}$ $df=1$ / $df=11$

(iii) F:



$F'_{11} = 9,65$

Reject H_0 if $F > 9,65$

(iv) $F = 93,16$

(v) $93,16 > 9,65$ Reject H_0.

Model is significant at $\alpha = 0,01$

* Note that F-test and t-test (c and d) must have same decision at same α.

5.8. Consider the following regression output†:

$$\hat{Y}_i = 0.2033 + 0.6560X_i$$

$$\text{se} = (0.0976)\ (0.1961)$$

$$r^2 = 0.397 \qquad \text{RSS} = 0.0544 \qquad \text{ESS} = 0.0358$$

where Y = labor force participation rate (LFPR) of women in 1972 and X = LFPR of women in 1968. The regression results were obtained from a sample of 19 cities in the United States.

✓ **a.** How do you interpret this regression?

✓ **b.** Test the hypothesis: $H_0: \beta_2 = 1$ against $H_1: \beta_2 > 1$. Which test do you use? And why? What are the underlying assumptions of the test(s) you use?

✓ **c.** Suppose that the LFPR in 1968 was 0.58 (or 58 percent). On the basis of the regression results given above, what is the mean LFPR in 1972? Establish a 95% confidence interval for the mean prediction.

5.8) a) Labor forces of Woman seems to be related for both years (1968 and 1972) but there's a significant change in Laborforce characteristics because only 39,7% of 72 can be explained by 68. The remaining 60,3% comes from other factors.

b) There's No need to make this test because $\hat{\beta}_2$ is already less than 1 and we do NOT Reject H_0. To see why, follow the steps:

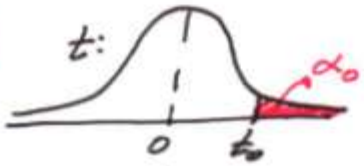
(i) $H_0: \beta_2 \leq 1$

$H_A: \beta_2 > 1$

$\alpha = \alpha_0$

(ii) $t = \dfrac{\hat{\beta}_2 - 1}{SE(\hat{\beta}_2)}$

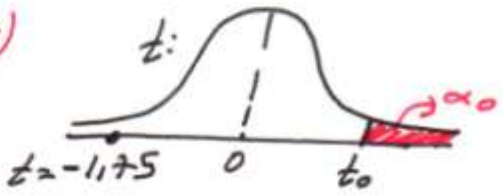
(iii)



Reject H_0 if $t > t_0$

(iv) $t = \dfrac{0,6560 - 1}{0,1961} = -1,75$

(v)



Do NOT Reject H_0.

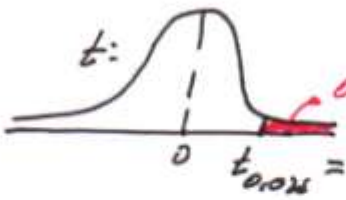
c) $\hat{y}_0 = 0,2033 + 0,6560 \cdot \underbrace{0,58}_{=X_0 = \bar{X}} = 0,5838$

$1 - \alpha = 0,95$

$\alpha/2 = 0,025$

$n = 19$

$df = 19 - 2 = 17$



$\hat{\sigma}^2 = \dfrac{RSS}{n-2} = \dfrac{0,054}{17}$

$\hat{\sigma}^2 = 0,0032$

$$\hat{\sigma}^2_{\hat{y}_0} = \hat{\sigma}^2 \cdot \left[\frac{1}{n} + \frac{\overbrace{X_0 - \bar{X}}^{=0}}{\sum x_i^2} \right] = 0,0032 \cdot \frac{1}{19} = 1,68 \cdot 10^{-4}$$

95% C.I. for Mean $\hat{y}_0$ is;

$\hat{y}_0 \pm t_{\alpha/2} \cdot \hat{\sigma}_{y_0}$

$0,5838 \pm 2,110 \cdot \sqrt{1,68 \cdot 10^{-4}}$

$(0,5564 \; ; \; 0,6112)$