ECON STAT-2 LECTURE NOTES | CHAPTERS 12 & 13

SIMPLE REGRESSION

Hypothesis Test for Correlation;

Sample Correlation: r

Population Correlation: ρ → measures linear relationship between X & Y

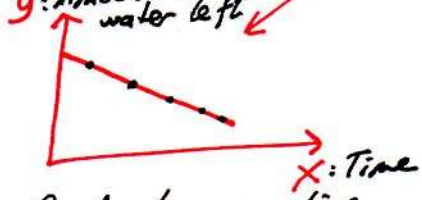
EXE X: weekly studying Hours (WSH)
y: Cummulative GPA

$$-1 \leq \rho \leq 1$$

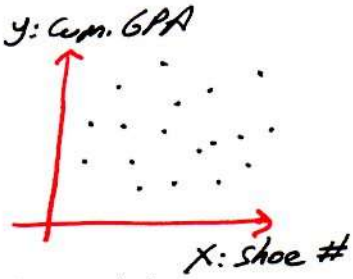
r is an estimator for ρ

$$-1 \leq r \leq 1$$

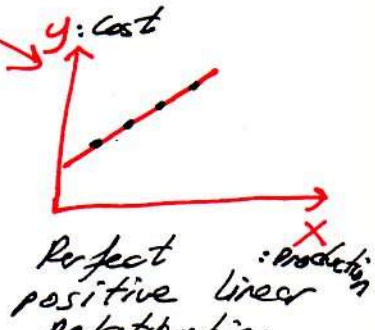
(Havuz Problemi)



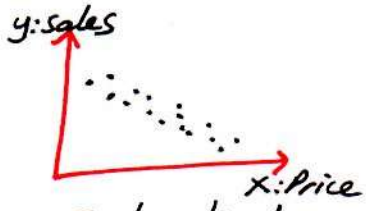
Perfect negative linear Relationship;
$r = -1$



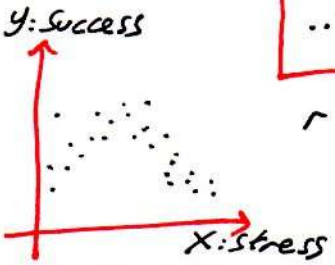
No relationship (Independent)
r close to 0



Perfect positive linear Relationship;
$r = 1$



r close to -1



Relationship, but NOT linear
r close to 0

r close to 1

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

* Through the Chapter, we have calculations chain. I'll code them as C1, C2, ---

C0) Find Σx_i ; Σy_i ; Σx_i^2 ; Σy_i^2 ; $\Sigma x_i y_i$

C1) $SS_{xx} = \sum (x_i - \bar{x})^2 = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}$; $\bar{x} = \frac{\Sigma x_i}{n}$

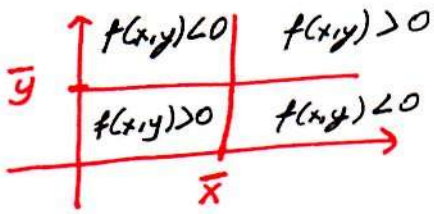
$$SS_{yy} = \Sigma (y_i - \bar{y})^2 = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n} \; ; \; \bar{y} = \frac{\Sigma y_i}{n}$$

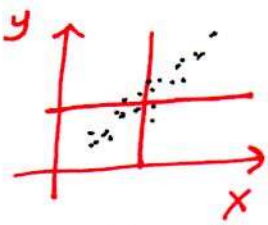
$$SS_{xy} = \Sigma (x_i - \bar{x}).(y_i - \bar{y}) = \Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}$$

C2) $$r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}}$$

Consider $SS_{xy} = \Sigma (x_i - \bar{x}).(y_i - \bar{y})$. we have;

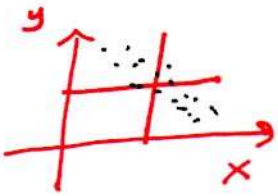
let $f(x,y) = (x_i - \bar{x}).(y_i - \bar{y})$





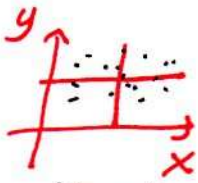
most $f(x_i, y_i) > 0$

$SS_{xy} > 0$



most $f(x_i, y_i) < 0$

$SS_{xy} < 0$



$f(x_i, y_i) > 0$ and $f(x_i, y_i) < 0$ cancel out, $SS_{xy} \to 0$

Dividing by $\sqrt{SS_{xx} \cdot SS_{yy}}$; we have $-1 \leq r \leq 1$.

That's why r measures the linear relationship.

* Test statistics to make hypothesis testing about ρ is ; $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$; d.of $f = n-2$

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

* Remember; Hypothesis testing Steps;

(i) H_0, H_A & α
(ii) Test statistics
(iii) Decision Criteria
(iv) Calculation
(v) Decision & Conclusion.

12.7 The accompanying table and the data file **Dow Jones** show percentage changes (x_i) in the Dow-Jones index over the first five trading days of each of 13 years and also the corresponding percentage changes (y_i) in the index over the whole year.

x	y	x	y
1.5	14.9	5.6	2.3
0.2	−9.2	−1.4	11.9
−0.1	19.6	1.4	27.0
2.8	20.3	1.5	−4.3
2.2	−3.7	4.7	20.3
−1.6	27.7	1.1	4.2
−1.3	22.6		

a. Calculate the sample correlation.
b. Test at the 10% significance level, against a two-sided alternative, the null hypothesis that the population correlation is 0.

12.7) a) ca)

X_i	y_i	X_i^2	y_i^2	$X_i \cdot y_i$
1,5	14,9	$1,5^2$	$14,9^2$	1,5.14,9
0,2	−9,2	$0,2^2$	$(-9,2)^2$	0,2.(−9,2)
⋮	⋮	⋮	⋮	⋮
1,1	22,6	$1,1^2$	$22,6^2$	1,1.22,6
+				

$\Sigma X_i = 16,6$ $\quad \Sigma X_i^2 = 80,06$ $\quad \Sigma X_i y_i = 121,15$

$\Sigma y_i = 153,6$ $\quad \Sigma y_i^2 = 3718,76$ $\quad \boxed{n=13}$

c1) $SS_{xx} = 80,06 - \frac{16,6^2}{13} = 58,9$ $\qquad SS_{yy} = 3718,76 - \frac{153,6^2}{13} = 1918,1$

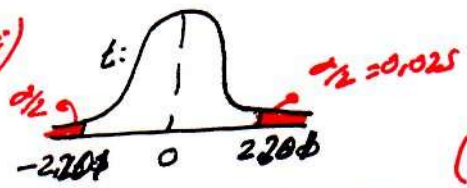
$$SS_{xy} = 121,15 - \frac{16,6 \cdot 153,6}{13} = -75,0$$

$$r = \frac{-75,0}{\sqrt{58,9 \cdot 1918,1}} = -0,223$$

b)(i) $H_0: \rho = 0$
$H_A: \rho \neq 0$
$\alpha = 0,01$

(ii) $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$; d.f. = 11

(iii)



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

(iv) $t = \frac{-0,223\sqrt{11}}{\sqrt{1-0,223^2}} = 0,76$

(v) Do NOT Reject H_0. Correlation is NOT different from 0 at $\alpha = 0,05$

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

12.6 The sample correlation for 68 pairs of annual returns on common stocks in country A and country B was found to be 0.51. Test the null hypothesis that the population correlation is 0 against the alternative that it is positive.

$n = 68$; $r = 0{,}51$

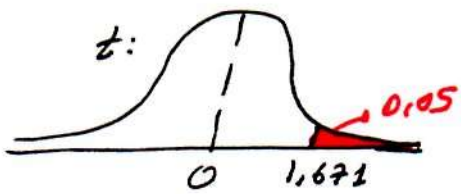
(i) $H_0: \rho \leq 0$

$H_A: \rho > 0$

$\alpha = 0{,}05$

(ii) $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$; d.of f = 66

(iii)



Reject H_0 if $t > 1{,}671$

(iv) $t = \dfrac{0{,}51 \cdot \sqrt{66}}{\sqrt{1-0{,}51^2}} = 4{,}82$

(v) Reject H_0. The correlation between Country A's common stocks and that of Country B is significantly positive at $\alpha = 0{,}05$

Simple Linear Regression Model;

Population Regression Model;

$$y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$$

Population Regression Function;

$$E(y_i) = \beta_0 + \beta_1 \cdot X_i$$

Sample Regression Model;

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_i + e_i$$

Sample Regression Function;

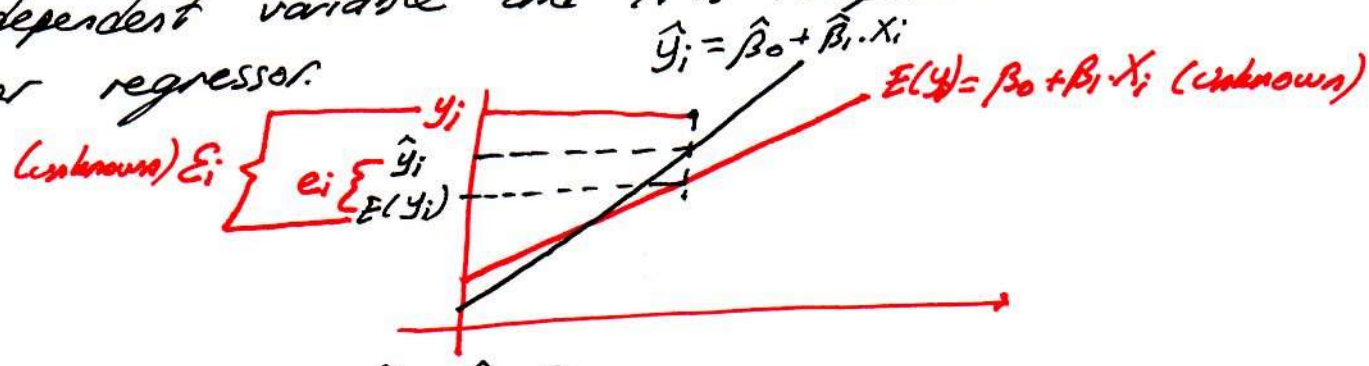
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_i$$

we have; y_i: Observed value $\quad \hat{y}_i$: Estimated value $\quad e_i = y_i - \hat{y}_i$

	(unknown Constants) Population Parameters		(Known Variables) Sample Statistics
Intercept	β_0		$\hat{\beta}_0$
Slope	β_1	Inference (i) Confidence Interval (ii) Hypothesis Testing	$\hat{\beta}_1$
Error	ε		e_i
Correlation Coefficient	ρ		r
Dependent Variable	$E(y_i)$		$\hat{y}_i$

Ex: y_i: Cummulative GPA
x_i: Weekly studying Hours.

We aim to explain y using X. So, y is dependent variable and X is independent variable or regressor.



$$y_i = \hat{\beta}_0 + \hat{\beta}_1 . X_i$$

Estimated

$\hat{\beta}_1$: Slope; Effect of a unit increase in X to y.

$\hat{\beta}_0$: Intercept (Not always interpretable): Estimated value of y when $X=0$

Ex (WSH) Let $n=20$ students are observed, Y_i: Cum. GPA and X_i: Weekly studying hours. Let,

$$\hat{Y}_i = 1{,}3 + 0{,}1 \cdot X_i$$

is sample regression function (SRF). (we will show how to find $\hat{\beta}_0$ and $\hat{\beta}_1$ later)

Sample data;

i	1	2	3	---	10	---	20
X_i	7	9	9	---	17	---	25
Y_i	2,8	2,3	2,1	---	3,1	---	3,7

Let the data is shown in ascending order of X_i.

Let $\bar{Y} = 2{,}5$

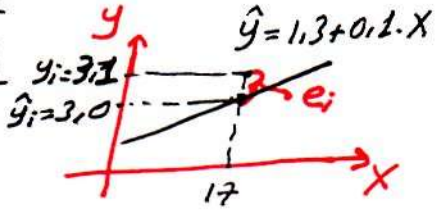
a) Interpret estimated Slope and Intercept

b) Estimate Cum. GPA of a student who studies 17 hours a week.

c) Estimate Cum. GPA of a student who studies 30 hours a week.

Ans a) Slope $=0{,}1$: If a student studies 1 more hour a week, its estimated ~~effect~~ that her Cum. GPA will increase ($\hat{\beta}_1 > 0$) by 0,1 points.

Intercept $= 1{,}3$ Not interpretable



b) $(\hat{Y} \mid X=17) = 1{,}3 + 0{,}1 \cdot 17 = 3{,}0$

c) $(\hat{Y} \mid X=30) = 1{,}3 + 0{,}1 \cdot 30 = 4{,}3 \rightarrow$ Does Not Make sense.

* Note that Regression function is valid only for the range of X: $[7; 25]$ or for values that are close to range. That's also why intercept is not meaningful. ($X=0$ is NOT close to range)

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

Least Squares Coefficient Estimators;

(C2) $\hat{\beta}_1 = \dfrac{SS_{xy}}{SS_{xx}}$; $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$

12.21 A corporation administers an aptitude test to all new sales representatives. Management is interested in the extent to which this test is able to predict their eventual success. The accompanying table records average weekly sales (in thousands of dollars) and aptitude test scores for a random sample of eight representatives.

Weekly sales	10	12	28	24	18	16	15	12
Test score	55	60	85	75	80	85	65	60

a. Estimate the linear regression of weekly sales on aptitude test scores.

b. Interpret the estimated slope of the regression line.

12.21)

a) X: Test Scores
y: weekly Sales } Because we want to estimate weekly Sales

x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
55	10	55^2	10^2	55.10
60	12	60^2	12^2	60.12
85	28	85^2	28^2	85.28
⋮	⋮	⋮	⋮	⋮
+ 60	12	60^2	12^2	60.12
$\Sigma x_i = 565$		$\Sigma x_i^2 = 40925$		$\Sigma x_i y_i = 9945$
	$\Sigma y_i = 135$		$\Sigma y_i^2 = 2553$	$\boxed{n=8}$

$SS_{xx} = 40925 - \dfrac{565^2}{8} = 1021,9$

$\bar{x} = \dfrac{565}{8} = 70,63$; $\bar{y} = \dfrac{135}{8} = 16,88$

$SS_{yy} = 2553 - \dfrac{135^2}{8} = 274,9$

$SS_{xy} = 9945 - \dfrac{565 \cdot 135}{8} = 410,6$

$\hat{\beta}_1 = \dfrac{410,6}{1021,9} = 0,402$ $\hat{\beta}_0 = 16,88 - 0,402 \cdot 70,63 = -11,51$

$$\boxed{\hat{y}_i = -11,51 + 0,402 \cdot x_i}$$

b) If test scores is increased by one unit, sales is expected to increase by 0,402.

c) Estimate weekly sales of a salesman whose test score is 70

$(\hat{y} \mid x = 70) = -11,51 + 0,402 \cdot 70 = 16,63$

12.22 It was hypothesized that the number of bottles of an imported premium beer sold per evening in the restaurants of a city depends linearly on the average costs of meals in the restaurants. The following results were obtained for a sample of $n = 17$ restaurants, of approximately equal size, where

y = Number of bottles sold per evening
x = Average cost, in dollars, of a meal

$\bar{x} = 25.5 \qquad \bar{y} = 16.0$

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = 350 \qquad \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = 180$$

a. Find the sample regression line.
b. Interpret the slope of the sample regression line.
c. Is it possible to provide a meaningful interpretation of the intercept of the sample regression line? Explain.

12.22) a) $n-1 = 16$

$SS_{xx} = \sum (x_i - \bar{x})^2 = 16 \cdot 350 = 5600$

$SS_{xy} = \sum (x_i - \bar{x}) \cdot (y_i - \bar{y}) = 16 \cdot 180 = 2880$

$\hat{\beta}_1 = \frac{2880}{5600} = 0,514$

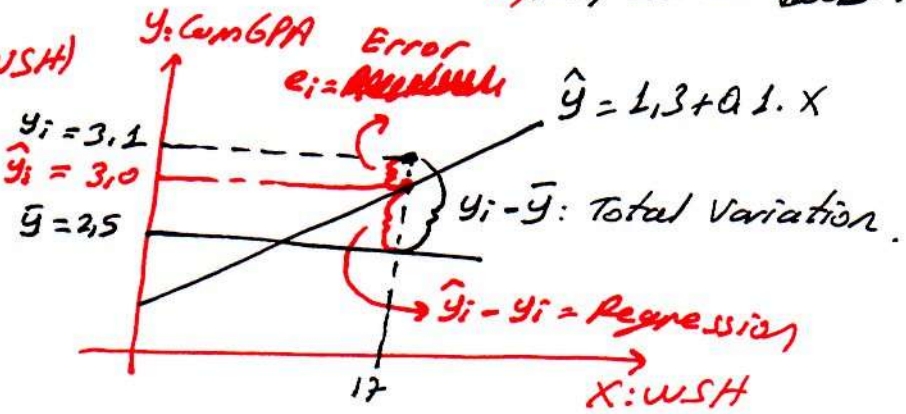
$\hat{\beta}_0 = 16 - 0,514 \cdot 25,5 = 2,893$

$\hat{y}_i = 2,893 + 0,514 \cdot X_i$

b) If Cost of Meals increases 1 unit, bottles of beer sold is expected to increase by 0,514 units.

c) No, Cost = 0 does not make sence.

Ex (WSH)



How good is our Regression model? Namely, how well can the exploratory variable Weekly Studying Hour ~~can~~ explain, or estimate Cum GPA?

The idea is; if we do NOT know weekly studying hour of a student, our best estimate for Cum. GPA would be sample mean: $\bar{y} = 2,5$. The specific student under consideration has GPA $y_i = 3,1$. The difference $y_i - \bar{y}$: Total Variation

Using WSH information, we "upgrade" our estimate $\bar{y}$ to $(\hat{y} \mid X=17) = 3,0$. X "explains" 0,5 of the total variation; however, we still have an unexplained part, $y_i - \hat{y}_i = e_i = 0,1$.

We have;

$$\underbrace{(y_i - \bar{y})}_{\text{Total Variation}} = \underbrace{(\hat{y}_i - \bar{y})}_{\substack{\text{Regression Variation} \\ \text{(Explained)}}} + \underbrace{(y_i - \hat{y}_i)}_{\substack{\text{Error Variation} \\ \text{(Unexplained)}}}$$

Taking the square and summing over all terms; (The cross product goes to 0.

$$\underbrace{\sum (y_i - \bar{y})^2}_{SST} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{SSE}$$

$$\boxed{SST = SSR + SSE}$$

where;

C3) $SST = SS_{yy}$

$SSR = \hat{\beta}_1^2 \cdot SS_{xx}$

$SSE = SST - SSR$

Coefficient of Determination;

Now, we can find a measure to our question. How much (what percentage) of total variation in y can be explained by X?

C4) $$\boxed{R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}}$$

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

* Remember; r: Correlation Coefficient.

We also have; $\boxed{R^2 = (r)^2 = \dfrac{SS_{xy}^2}{SS_{xx} \cdot SS_{yy}}}$

12.30 Find and interpret the coefficient of determination for the regression of the percentage change in the Dow-Jones index in a year on the percentage change in the index over the first five trading days of the year, continuing the analysis of Exercise 12.7. Compare your answer with the sample correlation found for these data in Exercise 12.7. Use the data file **Dow Jones**.

12.30 $SS_{yy} = 1918,1$

$SS_{xx} = 58,9$; $SS_{xy} = -75,0$; $n = 13$

$\bar{x} = \frac{16,6}{13} = 1,277$; $\bar{y} = \frac{153,6}{13} = 11,815$

$\hat{\beta}_1 = \frac{-75,0}{58,9} = -1,273$

$\hat{\beta}_0 = 11,815 + 1,273 \cdot 1,277 = 13,661$

The SRF is; $\hat{y}_i = 13,661 - 1,273 \cdot X_i$

$SST = 1918,1$

$SSR = (-1,273)^2 \cdot 58,9 = 95,65$

$SSE = 1918,1 - 95,65 = 1822,65$

$R^2 = \frac{95,65}{1918,1} = 0,0498 = (0,223)^2$

Dow Johnes index can explain only 5% of the total variation in percentage changes in the index.

Hypothesis Testing & Confidence intervals.

* Remember, we were ~~making~~ ~~[illegible]~~ ~~[illegible]~~ making inference on μ using $\bar{X}$: sample mean.

Now, we are going to make inference on β_i using $\hat{\beta}_i$. First, we need test statistics and standard error formulas.

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

C4) $\hat{\sigma}^2 = \dfrac{SSE}{n-2} = s^2$: Estimator of the Error Variance

C5) $\hat{\sigma}^2_{\hat{\beta}_1} = Var(\hat{\beta}_1) = \dfrac{\hat{\sigma}^2}{SS_{xx}} = s^2_{\hat{\beta}_1}$

$$\hat{\sigma}^2_{\hat{\beta}_0} = Var(\hat{\beta}_0) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{X}^2}{SS_{xx}}\right) = s^2_{\hat{\beta}_0}$$

* Remember; $t = \dfrac{\bar{X} - \mu}{s/\sqrt{n}} = \dfrac{\bar{X} - \mu}{s_{\bar{x}}}$ was test statistics for One Sample Mean test.

Likewise; $t = \dfrac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$ and $t = \dfrac{\hat{\beta}_0 - \beta_0}{s_{\hat{\beta}_0}}$

are test statistics for slope and intercept terms.

So, the confidence interval formulas follow;

C6) $(1-\alpha).100\%$ C.I. for β_i are;

$$\hat{\beta}_i \pm t_{\alpha/2} \cdot s_{\hat{\beta}_i} \; ; \; i = 1, 2 \; ; \; d.of\,f = n-2$$

$$\boxed{t = \frac{\hat{\beta}_i - \beta_i}{s_{\hat{\beta}_i}}}$$

in general, if model is

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$

d. of $f = n - (k+1)$

k: # of explanatory variables.

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

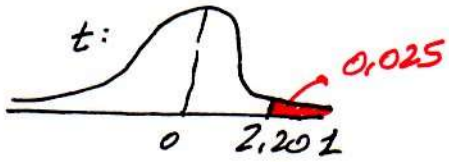
12.40 Continue the analysis of Exercise 12.30 of the regression of the percentage change in the Dow-Jones index in a year on the percentage change in the index over the first five trading days of the year. Use the data file **Dow Jones**.

a. Use an unbiased estimation procedure to find a point estimate of the variance of the error terms in the population regression.
b. Use an unbiased estimation procedure to find a point estimate of the variance of the least squares estimator of the slope of the population regression line.
c. Find and interpret a 95% confidence interval for the slope of the population regression line.
d. Test at the 10% significance level, against a two-sided alternative, the null hypothesis that the slope of the population regression line is 0.

12.40) a) $\hat{\sigma}^2 = \frac{1822,65}{11} = 165,7$

b) $\hat{\sigma}^2_{\hat{\beta}_1} = S^2_{\hat{\beta}_1} = \frac{165,7}{58,9} = 2,81$

c)



d.of f = 11

95% C.I. for β_1 is;

$-1,273 \pm 2,201 \cdot 2,81$

$(-7,66 ; 4,91)$

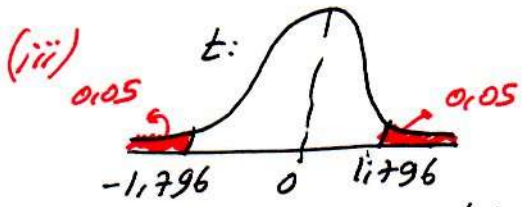
d) (i) $H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

$\alpha = 0,10$

(ii) $t = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}$; d.of f = 11

(iii)



Reject H_0 if $|t| > 1,796$

(iv) $t = \frac{-1,273 - 0}{\sqrt{2,81}} = -0,76$

Do NOT Reject H_0.

β_1 is NOT significantly different from zero.

This means that model is NOT valid at $\alpha = 0,05$

$y_i = \beta_0 + \underbrace{\beta_1}_{=0} \cdot X + \varepsilon$

$\Rightarrow$ X cannot explain y.

Confidence and Prediction Intervals;

$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_0$: Point estimate for $(y|X=x_0)$

1. Prediction interval for Individual y is;

$$\hat{y}_0 \pm t_{\alpha/2} \cdot \hat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

2. Confidence interval for Mean y is;

$$\hat{y}_0 \pm t_{\alpha/2} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

12.46 A sample of 25 blue-collar employees at a production plant was taken. Each employee was asked to assess his or her own job satisfaction (x) on a scale from 1 to 10. In addition, the numbers of days absent (y) from work during the last year were found for these employees. The sample regression line

$$\hat{y} = 12.6 - 1.2x$$

was estimated by least squares for these data. Also found were

$$\bar{x} = 6.0 \qquad \sum_{i=1}^{25} (x_i - \bar{x})^2 = 130.0 \qquad SSE = 80.6$$

a. Test at the 1% significance level against the appropriate one-sided alternative the null hypothesis that job satisfaction has no linear effect on absenteeism.

b. A particular employee has job satisfaction level 4. Find a 90% interval for the number of days this employee would be absent from work in a year.

12.46) $n = 25$

$\hat{y} = 12{,}6 - 1{,}2 \cdot x$; $\hat{\beta}_1 = -1{,}2$

$\bar{x} = 6{,}0$; $SS_{xx} = 130{,}0$; $SSE = 80{,}6$

a) $\hat{\sigma}^2 = \frac{80{,}6}{23} = 3{,}504$

$\hat{\sigma}^2_{\hat{\beta}_1} = s^2_{\hat{\beta}_1} = \frac{3{,}504}{130{,}0} = 0{,}027$

$s_{\hat{\beta}_1} = \sqrt{0{,}027} = 0{,}164$

(i) $H_0 : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

$\alpha = 0{,}05$

(ii) $t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$; d.of $f = 23$

(iii)

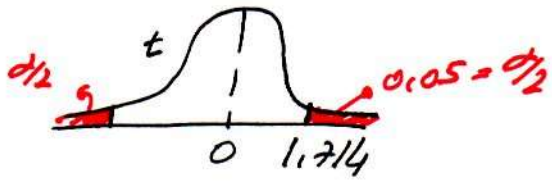
Reject H_0 if $|t| > 2{,}069$

(iv) $t = \frac{-1{,}2 - 0}{0{,}164} = 7{,}32$

(v) Reject H_0. The model is valid at $\alpha = 0{,}05$

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

b) $(\hat{y}_0 \mid X=4) = 12{,}6 - 1{,}2 \cdot 4 = 7{,}8$



$\hat{\sigma} = \sqrt{3{,}506} = 1{,}872$

90% prediction interval for individual $Y \mid X=4$ is;

$$7{,}8 \pm 1{,}714 \cdot 1{,}872 \cdot \sqrt{1 + \frac{1}{25} + \frac{(4-6)^2}{130}}$$

$$(4{,}48 \; ; \; 11{,}12)$$

c) Find a 90% C.I for mean (expected) absent days of all workers whose job satisfaction is 4

90% C.I. for mean (expected) $Y \mid X=4$ is;

$$7{,}8 \pm 1{,}714 \cdot 1{,}872 \cdot \sqrt{\frac{1}{25} + \frac{(4-6)^2}{130}}$$

$$7{,}8 \pm 0{,}85$$

$$(6{,}95 \; ; \; 8{,}65)$$

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

MULTIPLE REGRESSION

The Multiple Linear Regression Model;

we have k regressors to explain the ~~difference~~ variation in y and estimate y

Ex (WSH-continued)

X_1: weekly Studying Hour of the student

X_2: Entrance to the classes

X_3: ÖSS score

y: Cum. GPA.

Then;

Population Regression Model is;

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

(note that we have $k=3$)

Sample Regression Model;

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + e_i$$

Population Regression Function;

$$E(y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$

Sample Regression Function;

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$$

* Here, we do NOT calculate $\hat{\beta}_i$, or $S_{\hat{\beta}_i}$. Given them, we will learn how to test validity of the model, individual significance tests & confidence intervals.

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

* Interpretation of $\hat{\beta}_i$, $i=1,2,\dots,k$ has an additional comment to that of simple regression function. To illustrate, we interpret $\hat{\beta}_1$ of WSH example as follows;

"Keeping (X_2) Entrance to the classes and (X_3) ÖSS score constant, a unit increase in (X_1) WSH is expected to result in a $\hat{\beta}_1$ change in (y) Cum. GPA"

13.6 An aircraft company wanted to predict the number of worker-hours necessary to finish the design of a new plane. Relevant explanatory variables were thought to be the plane's top speed, its weight, and the number of parts it had in common with other models built by the company. A sample of 27 of the company's planes was taken, and the following model estimated:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

where

y_i = Design effort, in millions of worker-hours
x_{1i} = Plane's top speed, in miles per hour
x_{2i} = Plane's weight, in tons
x_{3i} = Percentage number of parts in common with other models

The estimated regression coefficients were
$b_0 = 150{,}4$
$b_1 = 0.661$ $b_2 = 0.065$ $b_3 = -0.018$

a. Interpret these estimates.

b. Estimate number of working hours for a plane whose top speed is 900 miles/hour, weight is 5,3 ton and have 40% common-parts with other models

13.6) Note that b_i represents $\hat{\beta}_i$.

Then;

$$\hat{y}_i = 150{,}4 + 0{,}661 \cdot X_{1i} + 0{,}065 \cdot X_{2i} - 0{,}0018 \cdot X_{3i}$$

a) Keeping weight and % of parts in common constant, a unit increase in planes top speed will increase working hours by 0,661

Keeping top speed and % of parts in common constant, a unit increase in weigth will increase working hours by 0,065

Keeping top speed and weight constant, a unit increase in % of parts in common will decrease working hours by 0,018
("is expected to" is better than "will")

b) $(\hat{y} \mid X_1 = 900,\ X_2 = 5{,}3,\ X_3 = 40)$
$= 150{,}4 + 0{,}661 \cdot 900 + 0{,}065 \cdot 5{,}3 - 0{,}018 \cdot 40 = 745$ millions hours of working.

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

* Coefficient of Determination: R^2 is found by the same way as simple regression; $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

Its interpretation is as follows;
"$100 \cdot R^2$ of the total variation in y can be explained by the regressors in the model."

For example, if $R^2 = 0{,}873$ in WSH example,
"87,3% of the total variation in Cum. GPA can be explained by WSH, Entrance to the Classes and ÖSS-Score"

* We have; $\hat{\sigma}^2 = \frac{SSE}{n-k-1}$: Estimator of error variance

Then, degrees of freedom for confidence intervals and hypothesis testing for individual β_i becomes $n-k-1$. That was why we had $n-2$ for simple regression.

* Adjusted R^2: $\bar{R}^2$ is used to compare multiple regression models with same Y: Dependent variable but different number of independent variables.
The idea is that, R^2 always increases by adding a new regressor but $\bar{R}^2$ increases if new regressor sufficiently increases the explained part.

$$\bar{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - (1-R^2) \cdot \frac{n-1}{n-k-1}$$

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

13.19 In the study of Exercise 13.6, where the least squares estimates were based on 27 sets of sample observations, the total sum of squares and regression sum of squares were found to be

$$SST = 3.881 \quad \text{and} \quad SSR = 3.549$$

a. Find and interpret the coefficient of determination.
b. Find the error sum of squares.
c. Find the adjusted coefficient of determination.

13.19) a) $R^2 = \frac{3,549}{3,881} = 0,914$

91,4% of the total variation in worker hours can be explained by plane's top speed, weight and % of the common parts in common.

b) $SSE = 3,881 - 3,549 = 0,332$

Also, estimate of error variance is;

$$\hat{\sigma}^2 = \frac{0,332}{27-3-1} = 0,0144$$

c) $\bar{R}^2 = 1 - (1 - 0,914) \cdot \frac{27-1}{27-3-1} = 0,903$

Another model explaining y can be better if its adjusted R^2 is greater than 0,903.

Testing the Validity of the Model: ANOVA

Consider WSH example. The population model is;

$$y_i = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \beta_3 \cdot X_{3i} + \varepsilon_i.$$

Note that, X_j's explain y_i via β_j's, $j = 1,2,3$.

Namely, if $\beta_1 = \beta_2 = \beta_3 = 0$, the model will be useless to explain or estimate Cum. GPA of the students.

This test is "Validity of the Model" test.

We have; (i) $H_0: \beta_1 = \beta_2 = \beta_3 = 0$

H_A: At least one β_j is Non-zero.

and "Do NOT reject H_0" means "Model is NOT valid."

"Reject H_0" means "At least one of the regressors significantly effects Cum. GPA"

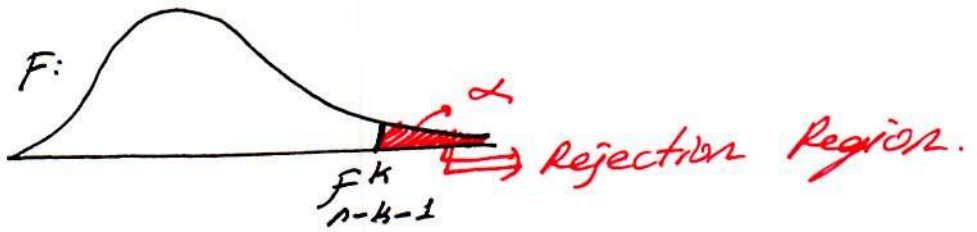
Alp Giray Özen | 0533 549 91 08 | alp@lecturemania.com | www.lecturemania.com

(ii) Our test statistics is F-test, which is calculated by constructing an Analysis of Variance (ANOVA) table. ANOVA table is as follows;

ANOVA

Source	degrees of freedom (df)	Sum of Squares (SS)	Mean Square (MS)	F
Regression	k	SSR	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Error	$(n-1)-k = n-k-1$	$SST - SSR = SSE$	$MSE = \frac{SSE}{n-k-1}$	—
TOTAL	$n-1$	SST	—	—

(iii) Since $F = \frac{MSR}{MSE}$ $\rightarrow$ d.of f $= k$, $\rightarrow$ d.of f $= n-k-1$



"Reject H_0 if $F > F^{k}_{n-k-1}$

(iv-v) We calculate F and compare with table value. If we fall in Rejection Region we conclude that the model is valid. Otherwise, the model is not valid. (Useless in explaining or estimating y)

* Note that $F = \frac{MSR}{MSE} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$

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

13.38 Refer to the study on aircraft design effort in Exercises 13.6 and 13.19.

a. Test the null hypothesis

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

b. Set out the analysis of variance table.

a. (i) $H_0: \beta_1 = \beta_2 = \beta_3 = 0$

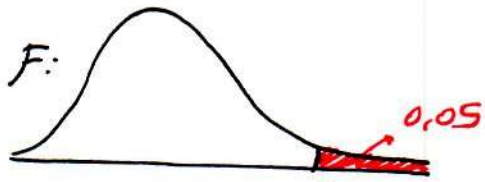
H_A: At least one β_j is Nonzero

$\alpha = 0{,}05$

(ii) $F = \dfrac{R^2/k}{(1-R^2)/(n-k-1)}$ $\rightarrow$ d.of f $= k = 3$

$\rightarrow$ d.of f $= n-k-1 = 27-3-1 = 23$

(iii)



$F^{3}_{23} = 3{,}03$

(iv) $F = \dfrac{\frac{0{,}914}{3}}{\frac{1-0{,}914}{23}} = 81{,}48$

Reject H_0 if $F > 3{,}03$

(v) Reject H_0. The model is Valid at $\alpha = 0{,}05$. (In fact, since F is too high, model is valid for any acceptable significance level.)

b) **ANOVA**

Source	d.f	S.S	M.S	F
Regression	3	3,549	$\frac{3{,}549}{3} = 1{,}183$	$\frac{1{,}183}{0{,}114} = 81{,}48$
Error	23	0,332	$\frac{0{,}332}{23} = 0{,}0144$	
Total	26	3,881		

* Note that we have found same F's.

Individual β_j's inferences;

* Confidence intervals and Hypothesis Testing for individual X_j's effect on Y is the same with that of the simple regression. We use t test;

$$\boxed{t = \frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}}} \; ; \; d \text{ of } f = n - k - 1$$

and the confidence intervals are;

$$\boxed{\hat{\beta}_j \pm t_{\alpha/2} \cdot s_{\hat{\beta}_j}}$$

* In general, the standard errors of the Estimated parameters are written in pharantesis under each estimate of the model.

13.27 In the study of Exercise 13.6, the estimated standard errors were

$$s_{b_1} = 0.099 \qquad s_{b_2} = 0.032 \qquad s_{b_3} = 0.002$$

a. Find 90% and 95% confidence intervals for β_1.
b. Find 95% and 99% confidence intervals for β_2.
c. Test against a two-sided alternative the null hypothesis that, all else being equal, the plane's weight has no linear influence on its design effort.
d. The error sum of squares for this regression was 0.332. Using the same data, a simple linear regression of design effort on percentage number of common parts was fitted, yielding an error sum of squares of 3.311. Test at the 1% level the null hypothesis that, taken together, top speed and weight contribute nothing in a linear sense to explanation of design effort, given that percentage number of common parts is also used as an explanatory variable.

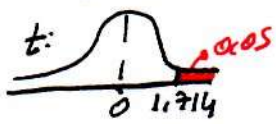
13.27) $s_{\hat{\beta}_1} = 0{,}099$; $s_{\hat{\beta}_2} = 0{,}032$; $s_{\hat{\beta}_3} = 0{,}002$

$$\hat{y}_i = 150{,}4 + \underset{(0{,}099)}{0{,}661 \cdot X_{1i}} + \underset{(0{,}032)}{0{,}065 \cdot X_{2i}} - \underset{(0{,}002)}{0{,}0018 \cdot X_{3i}}$$

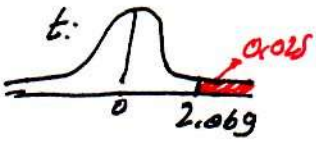
$n = 27$; $R^2 = 0{,}914$

d.of f $= 27 - 3 - 1 = 23$

$1 - \alpha = 90\%$
$\alpha/2 = 0{,}05$



$1 - \alpha = 95\%$
$\alpha/2 = 0{,}025$



$1 - \alpha = 0{,}99$
$\alpha/2 = 0{,}005$

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

a) 90% C.I. for β_1 is;

$0{,}661 \pm 1{,}714 \cdot 0{,}099$

$(0{,}491 ; 0{,}831)$

95% C.I. for β_1 is;

$0{,}661 \pm 2{,}069 \cdot 0{,}099$

$(0{,}456 ; 0{,}866)$

Note that 95% C.I. is wider because we need wider C.I. to have greater confidence level.

b) 95% C.I. for β_2 is;

$0{,}065 \pm 2{,}069 \cdot 0{,}032$

$(-0{,}001 ; 0{,}131)$

99% C.I. for β_2 is;

$0{,}065 \pm 2{,}5 \cdot 0{,}032$

$(-0{,}015 ; 0{,}145)$

Note that both C.I.'s contain 0. This means that, we cannot reject $H_0: \beta_2 = 0$ at a significance level 0,05 or lower.

c) Part (b) asserts that we cannot reject H_0 at $\alpha = 0{,}05$. Let's make the test at $\alpha = 0{,}10$.

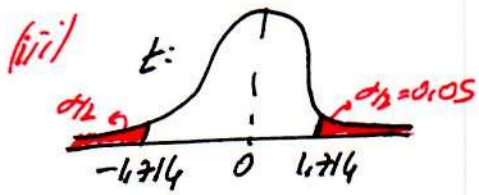
(i) $H_0: \beta_2 = 0$

$H_A: \beta_2 \neq 0$

$\alpha = 0{,}10$

(ii) $t = \dfrac{\hat{\beta}_2 - \beta_2}{s_{\hat{\beta}_2}}$

(iii)



Reject H_0 if $|t| > 1{,}714$

(iv) $t = \dfrac{0{,}065 - 0}{0{,}032} = 2{,}03$

(v) Reject H_0. Effect of ~~the~~ weight to working hours is significant at $\alpha = 0{,}10$.

Note that;

$\alpha = 0{,}05 \Rightarrow$ Do Not Reject H_0

$\alpha = 0{,}10 \Rightarrow$ Reject H_0.

Then; $0{,}05 < p\text{-value} < 0{,}10$.

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

d) Test of a Subset of Regression Coefficients;

Full Model; $y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + \alpha_1 Z_{1i} + \dots + \alpha_r Z_{ri} + \varepsilon_i$

SSE given (OR R^2 given)

Restricted Model; $y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki}$

SSE(r) given (OR $R^2_{(r)}$ given)

$H_0: \alpha_1 = \dots = \alpha_r = 0$ (Supports Restricted Model)

H_A: At least one α_j is NONZERO (Supports Full Model)

$$F = \frac{(SSE(r) - SSE)/r}{SSE/(n-k-r-1)} = \frac{(R^2 - R^2_{(r)})/r}{(1-R^2)/(n-k-r-1)}$$

Applying this test to part (d), we have

Full Model; $SSE = 0{,}332$; $y = \beta_0 + \beta_3 X_3 + \alpha_1 X_1 + \alpha_2 X_2 + \varepsilon$

Restricted Model; $SSE(r) = 3{,}331$; $y = \beta_0 + \beta_3 X_3 + \varepsilon$

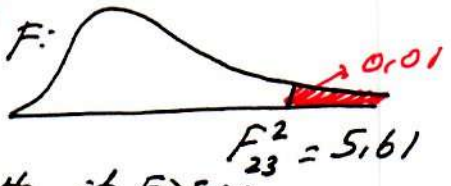
(i) $H_0: \alpha_1 = \alpha_2 = 0$

$H_A: \alpha_j \neq 0 \ \exists j = 1,2$

$\alpha = 0{,}01$

(ii) $F = \dfrac{(SSE(r) - SSE)/2}{SSE/23}$

(iii) F:



Reject H_0 if $F > 5{,}61$

(iv) $F = \dfrac{(3{,}331 - 0{,}332)/2}{0{,}332/23}$

$F = 103{,}88$

(v) Reject H_0. Data supports Full Model.

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

Reading Computer Output

13.83 A random sample of 93 freshmen at the University of Illinois was asked to rate on a scale from 1 (low) to 10 (high) their overall opinion of residence hall life. They were also asked to rate their levels of satisfaction with roommates, with the floor, with the hall, and with the resident advisor. (Information on satisfaction with the room itself was obtained, but this was later discarded as it provided no useful additional power in explaining overall opinion.) The following model was estimated:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

where

Y = Overall opinion of residence hall
x_1 = Satisfaction with roommates
x_2 = Satisfaction with floor
x_3 = Satisfaction with hall
x_4 = Satisfaction with resident advisor

Use the accompanying portion of the computer output from the estimated regression to write a report summarizing the findings of this study.

DEPENDENT VARIABLE: Y OVERALL OPINION

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	R-SQUARE
MODEL	4	37.016	9.2540	9.958	0.312
ERROR	88	81.780	0.9293		
TOTAL	92	118.79			

PARAMETER	ESTIMATE	STUDENT'S t FOR HO: PARAMETER = 0	STD. ERROR OF ESTIMATE
INTERCEPT	3.950	5.84	0.676
X1	0.106	1.69	0.063
X2	0.122	1.70	0.072
X3	0.092	1.75	0.053
X4	0.169	2.64	0.064

13.83)

$$\hat{y}_i = \underset{(0,676)}{3,950} + \underset{(0,063)}{0,106} X_{1i} + \underset{(0,072)}{0,122} X_{2i} + \underset{(0,053)}{0,092} X_{3i} + \underset{(0,064)}{0,169} X_{4i}$$

$n = d.of.f\,(Total) + 1 = 93$

$R^2 = 0,312$

* Validity of the model;

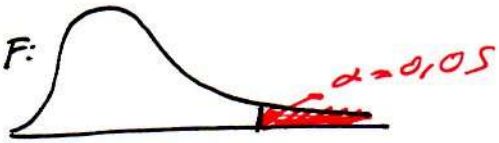
(i) $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_A: At least one $\beta_j \neq 0$

$\alpha = 0,05$

(ii) $F = \dfrac{MSR}{MSE}$ $MSR \to d.of.f = 4$, $MSE \to d.of.f = 88$

(iii) F:



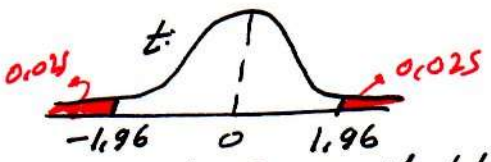
$F^4_{88} = 2,45$

(iv) $F = 9,958$

(v) Reject H_0.

* Individual β_j's.

Let $\alpha = 0,05$



Reject $H_0: \beta_j = 0$ if $|t_j| > 1,96$

$t_1 = 1,69$ NOT Sig.

$t_2 = 1,70$ NOT Sig.

$t_3 = 1,75$ NOT Sig.

$t_4 = 2,64$ Sig.

Only sat. with res. advisor is significant.

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

DUMMY VARIABLE

A dummy variable is;

$$X = \begin{cases} 1 & \text{if unit is in the category} \\ 0 & \text{o.w.} \end{cases}$$

Consider $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

where; y_i : Salary

X_{1i} : Experience

$$X_{2i} = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female.} \end{cases}$$

Then; β_2 stands for "Contribution" of being male (that is difference of male salaries).

13.74 A consulting group offers courses in financial management for executives. At the end of these courses participants are asked to provide overall ratings of the value of the course. For a sample of 25 courses the following regression was estimated by least squares.

$$\hat{y} = 42.97 + \underset{(0.29)}{0.38x_1} + \underset{(0.21)}{0.52x_2} - \underset{(0.11)}{0.08x_3} + \underset{(0.359)}{6.21x_4} \qquad R^2 = 0.569$$

where

y = Average rating by participants of the course
x_1 = Percentage of course time spent in group discussion sessions
x_2 = Money, in dollars, per course member spent on preparing course material
x_3 = Money, in dollars, per course member spent on food and drinks
x_4 = Dummy variable taking the value 1 if a visiting guest lecturer is brought in and 0 otherwise

The numbers in parentheses under the coefficients are the estimated coefficient standard errors.

a. Interpret the estimated coefficient on x_4.
b. Test against the alternative that it is positive the null hypothesis that the true coefficient on x_4 is 0.

13.74) a) "Average rating of the course" is 6,21 more for those who brought a visiting guest lecturer. (Significant at $\alpha = 0{,}05$)

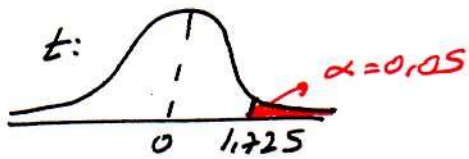
b) (i) $H_0: \beta_4 \leq 4$

$H_A: \beta_4 > 4$

$\alpha = 0{,}05$

(ii) $t = \dfrac{\hat{\beta}_4 - \beta_4}{s_{\hat{\beta}_4}}$; d.of $f = 25 - 4 - 1 = 20$

(iii) t:



Reject H_0 if $t > 1{,}725$

(iv) $t = \dfrac{6{,}21 - 0}{0{,}359} = 17{,}3$

(v) Reject H_0.

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