Chapter 12

Simple Linear Regression

(12.1-12.2-12.3-12.7)

Section 12.2

Example (1)

(Textbook page 445-12.1)

Fitting a straight line to a set of data yields the following prediction line:

$$\hat{Y}_i = 7 + 2X_i$$

a) Interpret the meaning of the Y-intercept,b_0.

b) Interpret the meaning of the slop,b_1.

c) Predict the mean value of Y for X=3

Solution:

a) Interpret the meaning of the Y-intercept,b_0.

The Y-intercept,$b_0 = 7$, implies that when the value of X is 0, the predicted mean value of Y is 7.

b) Interpret the meaning of the slop,b_1.

The slope coefficient,$b_1 = 2$, implies that for each increase of 1 unit in X the predicted mean value of Y is estimated to increase by 2 units

c) Predict the mean value of Y for X=3

$$\hat{Y} = 7 + 2(3) = 7 + 6 = 13$$

Example (2)

(Textbook page 445-12.3)

Fitting a straight line to a set of data yields the following prediction line:

$$\hat{Y}_i = 16 - 0.5X_i$$

a) Interpret the meaning of the Y-intercept,b_0.

b) Interpret the meaning of the slop,b_1.

c) Predict the mean value of Y for X=6

Solution:

a) Interpret the meaning of the Y-intercept,b_0.

The Y-intercept,b_0 =16, implies that when the value of X is 0, the predicted mean value of Y is 16.

b) Interpret the meaning of the slop,b_1.

The slope coefficient,b_1 = -0.5, implies that for each increase of 1 unit in X the predicted mean value of Y is estimated to decrease by 0.5 units

c) Predict the mean value of Y for X=6

$$\hat{Y} = 16 - 0.5(6) = 16 - 3 = 13$$

Example (3)

Fitting a straight line to a set of data yields the prediction line $\widehat{Y}_i = 7 + 2X_i$.

The values of X used to find the prediction line range from 1 to 20

a) Should this model be used to predict the mean value of Y when X equals 15?

b) Should this model be used to predict the mean value of Y when X equals 22?

c) Should this model be used to predict the mean value of Y when X equals 0?

Solution:

a) Should this model be used to predict the mean value of Y when X equals 15?

Yes

b) Should this model be used to predict the mean value of Y when X equals 22?

No

c) Should this model be used to predict the mean value of Y when X equals 0?
No

Section 12.3

Example (4)

(Textbook page 451-12.14)

If SSE=12, and SSR=28, from a sample of 4

a) Compute the coefficient of determination, r^2,and interpret its meaning.

b) Determine the standard error of the estimate

Solution:

a)

SST=SSR+SSE=28+12=40

$$r^2 = \frac{SSR}{SST} = \frac{28}{40} = 0.70$$

It means that r^2 *•100% of the variation in the dependent variable can be explained by the variation in the independent variable.*

It means that 70% of the variation in the dependent variable can be explained by the variation in the independent variable

Since the value of r^2 is close to 1, the regression model is very useful.

b) The standard error of the estimate=$S_{XY} = \sqrt{\frac{SSE}{n-2}}\sqrt{\frac{12}{4-2}} = \sqrt{\frac{12}{2}} = \sqrt{6}$ =2.45

Example (5)

(Textbook pages 449-451)

In the Sunflowers Apparel scenario, the business objective of the director of planning is to forecast annual sales for all new stores, based on the number of profiled customers who live no more than 30 minutes from a Sunflowers store. To examine the relationship between the number of profiled customers ($ millions) who live within a fixed radius from a Sunflowers store and its annual sales ($ millions), data were collected from a sample of 14 stores.

Store	Profiled Customers(X)	Annual Sales(Y)	XY	Y^
1	3.7	5.7	21.09	32.49
2	3.6	5.9	21.24	34.81
3	2.8	6.7	18.76	44.89
4	5.6	9.5	53.2	90.25
5	3.3	5.4	17.82	29.16
6	2.2	3.5	7.7	12.25
7	3.3	6.2	20.46	38.44
8	3.1	4.7	14.57	22.09
9	3.2	6.1	19.52	37.21
10	3.5	4.9	17.15	24.01
11	5.2	10.7	55.64	114.49
12	4.6	7.6	34.96	57.76
13	5.8	11.8	68.44	139.24
14	3	4.1	12.3	16.81
Total	52.9	92.8	382.85	693.9

$\hat{Y}_I$ =-1.2088 +2.0742 X_I , with $\sum_{i=1}^{n} Y_I = 92.8$, $\sum_{i=1}^{n} Y_i^2 = 693.9$, and

$\sum_{i=1}^{n} X_i Y_i = 382.85$, n=14

Find

a) SSR , b) SSE , c) SST , d) r^2 , e) S_{xy}

Solution:

a)

$$SSR = b_0 \sum_{i=1}^{n} Y_i + b_1 \sum_{i=1}^{n} X_i Y_i - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}$$

$$= -1.2088(92.8) + 2.0742(382.85) - \frac{92.8^2}{14} = 66.7854$$

b)

$$SSE = \sum_{i=1}^{n} Y_i^2 - b_0 \sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i Y_i$$

$$= 693.9 - (-1.2088 \times 92.8) - 2.0742(382.85) = 11.9822$$

c)

$$SST = SSR + SSE = 66.7854 + 11.9822 = 78.7686$$

Or

$$SST = \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n} = 693.9 - \frac{92.8^2}{14} = 693.9 - 615.13 = 78.7686$$

d)

$$r^2 = \frac{SSR}{SST} = \frac{66.7854}{78.7686} = 0.8479$$

It means that r^2 *•100% of the variation in the dependent variable can be explained by the variation in the independent variable.*

It means that 84.79% of the variation in the dependent variable can be explained by the variation in the independent variable

Since the value of r^2 , is close to 1, the regression model is very useful.

e)

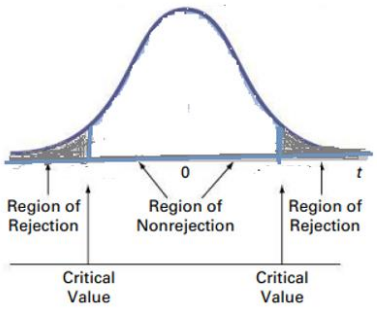
$$S_{xy} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{11.9822}{14-2}} = 0.9993$$

Section 12.7

Step (1)

H_0: $\beta_1 = 0$ (There is no linear relationship between X and Y, the slope is zero)

H_1: $\beta_1 \neq 0$ (There is linear relationship between X and Y, the slope is not zero)



Step (2) $t_{\frac{\alpha}{2},n-2}$

Step (3)

$$t_{stat} = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{b_1 - 0}{S_{b_1}} = \frac{b_1}{S_{b_1}}$$

Step (4)

Reject H_0 if $t_{stat} > t_{\frac{\alpha}{2},n-2}$, $t_{stat} < -t_{\frac{\alpha}{2},n-2}$

Step (5) Decision

Construct a 95% confidence interval estimate of the population slope, β1.

$$b_1 \pm t_{\frac{\alpha}{2},n-2} S_{b_1}$$

$$b_1 - t_{\frac{\alpha}{2},n-2} S_{b_1} < \beta_1 < b_1 + t_{\frac{\alpha}{2},n-2} S_{b_1}$$

Example (6)

(Textbook pages 460-463)

Return to example (5)

You are testing the null hypothesis that there is no linear relationship between two variables, X and Y. From your sample of n = 14, you determine that $b_1 = 2.07417$ and $S_{b_1} = 0.2536$.

a) What is the hypothesis of the test?

b) What is the value of t_{stat}?

c) At the $\alpha = 0.05$ level of significance, what are the critical values?

d) Based on your answers to (a) and (b), what statistical decision should you make?

e) Construct a 95% confidence interval estimate of the population slope, β1.

Solution:

a) What is the hypothesis of the test?

H_0: $\beta_1 = 0$ (There is no linear relationship between X and Y, the slope is zero)

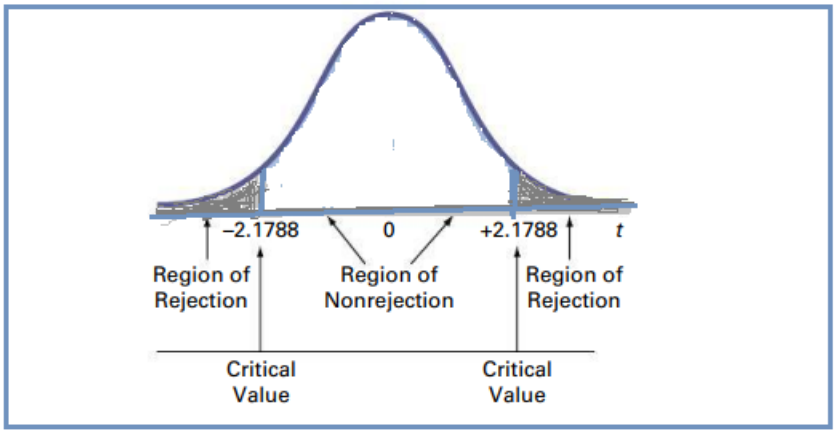
H_1: $\beta_1 \neq 0$ (There is linear relationship between X and Y, the slope is not zero)

b) What is the value of t_{stat}?

$$t_{stat} = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{b_1 - 0}{S_{b_1}} = \frac{b_1}{S_{b_1}} = \frac{2.07417}{0.2536} = 8.178$$

c) At the $\alpha = 0.05$ level of significance, what are the critical values?

$t_{\frac{\alpha}{2},n-2} = t_{\frac{0.05}{2},14-2} = t_{0.025,12} = \pm 2.1788$



d) Based on your answers to (a) and (b), what statistical decision should you make?

Since the $t_{stat} = 8.178$ is greater than the upper critical value $t_{\frac{\alpha}{2},n-2} = 2.1788$, reject the null hypothesis. There is enough evidence that there is a relationship between X and Y

e) Construct a 95% confidence interval estimate of the population slope, β1.
$b_1 \pm t_{\frac{\alpha}{2},n-2} S_{b_1} = 2.0742 \pm 2.1788(0.2536) = 2.0742 \pm 0.5526$

$1.5216 < \beta_1 < 2.6268$

At 95% level of confidence, the confidence interval for the slope is (1.5216, 2.6268)
Since this interval does not contain 0, we are 95% confident that $\beta_1 \neq 0$ and this leads to the same result.

Excel t test for the slope results for the Sunflowers Apparel data

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-1.2088	0.9949	-1.2151	0.2477	-3.3765	0.9588
X Variable 1	2.0742	0.2536	8.178	0	1.5216	2.6268

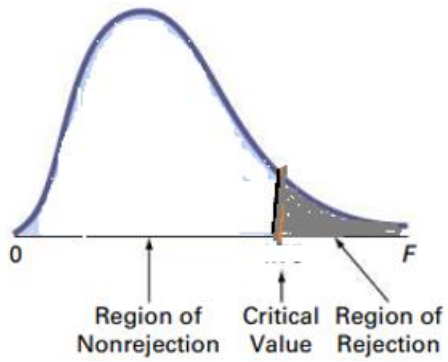
b_0 b_1 S_{b_1} $t_{stst|}$ $L < \beta_1 < U$

	Regression Statistics	
r →	Multiple R	0.9208
r^2 →	R Square	0.8479
	Adjusted R Square	0.8352
S_{xy} →	Standard Error	0.9993
	Observations	14

Step (1)

H_0: $\beta_1 = 0$ (There is no linear relationship between X and Y, the slope is zero)

H_1: $\beta_1 \neq 0$ (There is linear relationship between X and Y, the slope is not zero)



Step (2) $F_{\alpha,1\,,n-2}$

Step (3) $$F_{STAT} = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} = \frac{MSR}{MSE}$$

Step (4) Reject H_0, if $F_{STAT} > \quad F\alpha$

Step (5) Decision

Example (7)

Return to example (5)

You are testing the null hypothesis that there is no linear relationship between two variables, X and Y. From your sample of n = 14, you determine that SSR=66.7854 and $SSE = 11.9832$.

a) What is the hypothesis of the test?

b) What is the value of F_{stat}?

c) At $\alpha = 0.05$ level of significance, what are the critical values?

d) Based on your answers to (a) and (b), what statistical decision should you make?

e) Compute the correlation coefficient by first computing $r^2$and b_1=2.0742.

Solution:

a) What is the hypothesis of the test?

H_0: $\beta_1 = 0$ (There is no linear relationship between X and Y, the slope is zero)

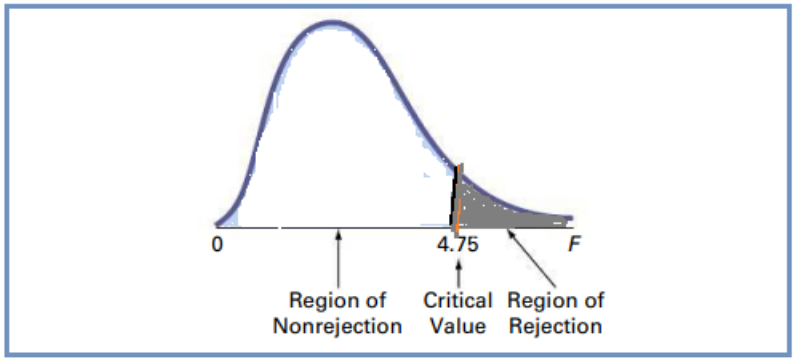
H_1: $\beta_1 \neq 0$ (There is linear relationship between X and Y, the slope is not zero)

b) $$MSR = \frac{SSR}{1} = \frac{66.7854}{1} = 66.7854$$

$$MSE = \frac{11.9832}{n-2} = \frac{11.9832}{20-2} = \frac{11.9832}{18} = 0.9986$$

$$F_{STAT} = \frac{MSR}{MSE} = \frac{66.7854}{0.9986} = 66.8792$$

c) $F_{0.05,1,12} = 4.75$



d) Reject H_0.There is enough evidence that there is a relationship between X and Y also, it is a good evidence that the fitted linear regression model is useful.

e) SST=SSR+SSE=66.7854+11.9832=78.7686

$$r^2 = \frac{SSR}{SST} = \frac{66.7854}{78.7686} = 0.8479 \qquad r = \sqrt{0.84790} = 0.9208$$

Excel F test for the slope results for the Sunflowers Apparel data

ANOVA

		df	*SS*	*MS*	*F*
SSR →	Regression	1	60.7854	60.7854	66.8792
SSE →	Residual	12	11.9832	0.9985	
SST →	Total	13	78.7686		

	Regression Statistics	
r →	Multiple R	0.9208
r^2 →	R Square	0.8479
	Adjusted R Square	0.8352
S_{xy} →	Standard Error	0.9993
	Observations	14

Example (8)

(Textbook page 465-12.40)

You are testing the null hypothesis that there is no linear relationship between two variables, X and Y. From your sample of n = 18, you determine that b1=4.5 and $S_{b_1} = 1.5$.

a) What are the hypotheses to test?

b) What is the value of t_{stat}?

c) At the $\alpha = 0.05$ level of significance, what are the critical values?

d) Based on your answers to (a) and (b), what statistical decision should you make?

e) Construct a 95% confidence interval estimate of the population slope, β1.

Solution:

a) What are the hypotheses to test?

H_0: $\beta_1 = 0$ (There is no linear relationship, the slope is zero)

H_1: $\beta_1 \neq 0$ (There is linear relationship, the slope is not zero)

b) What is the value of t_{stat}?

$$t_{stat} = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{b_1 - 0}{S_{b_1}} = \frac{b_1}{S_{b_1}} = \frac{4.5}{1.5} = 3$$

c) At the $\alpha = 0.05$ level of significance, what are the critical values?

$$t_{\frac{\alpha}{2},n-2} = t_{\frac{0.05}{2},18-2} = t_{0.025,16} = \pm 2.1199$$

d) Based on your answers to (a) and (b), what statistical decision should you make?

Since the $t_{stat} = 3$ is greater than the upper critical value $t_{\frac{\alpha}{2},n-2} = 2.1199$, reject the null hypothesis. There is evidence that there is a relationship between X and Y

e) Construct a 95% confidence interval estimate of the population slope, β1.

$$\beta_1 = b_1 \pm t_{\frac{\alpha}{2},n-2} S_{b_1} = 4.5 \pm 2.1199(1.5) = 4.5 \pm 3.17985$$

$$1.32 < \beta_1 < 7.67$$

At 95% level of confidence, the confidence interval for the slope is (1.32, 7.67)

Since this interval does not contain 0, we are 95% confident to reject H_0, or, 95% confident that there is a relationship between X and Y.

Example (9)

(Textbook page 465-12.41)

You are testing the null hypothesis that there is no linear relationship between two variables, X and Y. From your sample of n = 20, you determine that SSR=60 and $SSE = 40$.

a) What is the hypothesis of the test?

b) What is the value of F_{stat}?

c) At the $\alpha = 0.05$ level of significance, what are the critical values?

d) Based on your answers to (a) and (b), what statistical decision should you make?

e) Compute the correlation coefficient by first computing r^2 and assuming that b_1 is negative.

Solution:

a) What is the hypothesis of the test?

H_0: $\beta_1 = 0$ (There is no linear relationship, the slope is zero)

H_1: $\beta_1 \neq 0$ (There is linear relationship, the slope is not zero)

b) $MSR = \frac{SSR}{1} = \frac{60}{1} = 60$

$MSE = \frac{SSE}{n-2} = \frac{40}{20-2} = \frac{40}{18} = 2.222$

$F_{STAT} = \frac{MSR}{MSE} = \frac{60}{2.222} = 27$

c) $F_{0.05,1,18} = 4.41$

d) Reject H_0.There is evidence that there is a relationship between X and Y

e) $r^2 = \frac{SSR}{SST} = \frac{60}{100} = 0.6$ $\quad\quad$ $r = -\sqrt{0.60} = -0.7746$