

Some Calculations to do the Regression example worked by the computer in the handout

Row	x_i	y_i	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	100	40	90000	6000
2	200	50	40000	2000
3	300	50	10000	1000
4	400	70	0	0
5	500	65	10000	500
6	600	65	40000	1000
7	700	80	90000	6000
	$\sum x_i = 2800$	$\sum y_i = 420$	$\sum (x_i - \bar{x})^2 = 280,000$	$\sum (x_i - \bar{x})(y_i - \bar{y}) = 16,500$

Now lets compute the slope coefficient using formula (14.6) on page 543.

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b_1 = \frac{16500}{280,000}$$

$$b_1 = .0589286$$

Since the mean of X is 400, and the mean of Y is 60, formula (14.7) gives the intercept.

$$b_0 = \bar{y} - b_1\bar{x} = 60 - (.0589286)(400) = 36.4286$$

We can use the estimated regression line to generate predicted values, using the fact that

$$\hat{Y}_i = b_0 + b_1X_i = 36.4286 + .0589286X_i$$

Row	$y - \hat{y}$	$y - \bar{y}$	$\hat{y} - \bar{y}$	$(y - \hat{y})^2$	$(y - \bar{y})^2$	$(\hat{y} - \bar{y})^2$
1	-2.3214	-20	-17.6786	5.389	400	312.532
2	1.7857	-10	-11.7857	3.189	100	138.903
3	-4.1071	-10	-5.8929	16.869	100	34.726
4	10.0000	10	0.0000	100.000	100	0.000
5	-0.8929	5	5.8929	0.797	25	34.726
6	-6.7857	5	11.7857	46.046	25	138.903
7	2.3214	20	17.6786	5.389	400	312.532
				$\sum (y_i - \hat{y}_i)^2$ = 177.68	$\sum (y_i - \bar{y})^2$ = 1150.0	$\sum (\hat{y}_i - \bar{y})^2$ = 972.32

By equation (14.12) on page 554, $r^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$, which in this case is

$$r^2 = 972.32/1150.0 = .845496$$

561. To calculate the mean square error (our estimate of σ^2) we use formula (14.15) on page 561.

$$s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{177.68}{5} = 35.536$$

To get the standard error of estimate, use formula (14.16) on page 561 - (take the square root)

$$s = \sqrt{s^2} = \sqrt{35.536} = 5.96121$$

To get the standard deviation of b_1 , s_{b_1} , use formula (14.18) on page 562.

$$\begin{aligned} s_{b_1} &= \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} \\ &= \frac{5.96121}{\sqrt{280,000}} = .0112656 \end{aligned}$$

A Confidence interval for the slope coefficient.

The $(1 - \alpha)\%$ Confidence interval for β_1 can be created using the following formula:¹

$$\beta_1 = b_1 \pm t_{\frac{\alpha}{2}} s_{b_1}$$

The t value in question has 5 degrees of freedom in this case; there were 7 observations, but two were “used up” in estimating b_0 and b_1 , leaving 5. Therefore if one wanted a 95% Confidence interval, it would be

$$\beta_1 = b_1 \pm t_{\frac{\alpha}{2}} s_{b_1}$$

$$\beta_1 = .0589286 \pm (2.571)(.0112656)$$

$$\beta_1 = .0589286 \pm .0289639$$

If you were testing the following hypothesis $H_0 : \beta_1 = 0$ using the t-statistic, the observed value of the t statistic is $t_{obs} = \frac{b_1 - 0}{s_{b_1}} = \frac{.0589286 - 0}{.0112656} = 5.231$. The p-value for this test is the probability of an

outcome this far or further from the null of zero, which is $P(t < -5.231 \cup t > 5.231)$. Evaluating this using Minitab, we discover the p-value is $2 \times .0017 = .0034$.

The F statistic is calculated by formula (14.21) on page 565,

¹ A similar, less commonly used formula gives us a confidence interval for the intercept.

$$\beta_0 = b_0 \pm t_{\frac{\alpha}{2}} s_{b_0}$$

The formula for computing s_{b_0} is not given in the book, but Minitab computes it.

$$F_{df_1, df_2} = \frac{MSR}{MSE} = \frac{\sum(\hat{y}_i - \bar{y})^2 / df_1}{\sum(y_i - \hat{y}_i)^2 / df_2}$$

where df_1 is the regression degrees of freedom: in simple regression $df_1 = 1$
and df_2 is our ordinary degrees of freedom, $n-2$.

$$\text{Here } F_{1,5} = \frac{972.32/1}{177.68/5} = 27.3615$$

Now we turn to predicting for the value $x_0 = 550$. The predicted value is

$$\hat{y} = b_0 + b_1x_0 = 36.4286 + (.0589286)(550) = 68.8393$$

The 95% Confidence interval for the mean value of y is given by formulas (14.23) and (14.24) on page 570.

$$\hat{y}_p \pm t_{\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

$$68.8393 \pm (2.571)(5.96121) \sqrt{\frac{1}{7} + \frac{(550 - 400)^2}{280,000}}$$

$$68.8393 \pm 7.24098$$

The 95% Prediction interval for an Individual value of Y is given by formulas (14.26) and (14.27) on page 572.

$$\hat{y}_p \pm t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

$$68.8393 \pm (2.571)(5.96121) \sqrt{1 + \frac{1}{7} + \frac{(550 - 400)^2}{280,000}}$$

$$68.8393 \pm 16.9507$$

We can calculate the leverage of each observation using formula (14.33) on p. 592. The actual calculations are done in the worksheet below.

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

Rows	X_i	$(x_i - \bar{x})^2$	h_i
1	100	90000	0.464286
2	200	40000	0.285714
3	300	10000	0.178571
4	400	0	0.142857
5	500	10000	0.178571
6	600	40000	0.285714
7	700	90000	0.464286

$$\bar{x} = 400 \quad \sum (x_i - \bar{x})^2 = 280,000$$

Finally, we use formula (3.10) and (3.12) on page 96 to calculate the correlation coefficient between X and Y:

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y}) / (n-1)}{\sqrt{\sum (x_i - \bar{x})^2 / (n-1)} \sqrt{\sum (y_i - \bar{y})^2 / (n-1)}}$$

$$r_{xy} = \frac{16,500/6}{\sqrt{280,000/6} \sqrt{1150/6}}$$

$$r_{xy} = \frac{2750}{(216.025)(13.8444)} = .919509$$

We can verify the relationship between the correlation coefficient and the Coefficient of Determination, as stated in formula (14.13) on page 555.

$$r_{xy} = (\text{sign of } b_1) \sqrt{\text{Coefficient of Determination}} = \pm \sqrt{r^2}$$

$$.919509 = +\sqrt{.845496}$$

As may be mentioned in class, the other way to calculate the F statistic is by using the (equivalent) formula that:

$$F_{df_1, df_2} = \frac{r^2 / df_1}{(1-r^2) / df_2} = \frac{.845496 / 1}{(1-.845496) / 5} = 27.3615$$

And we know for the special case of simple regression that

$$t = (\text{sign of } b_1) \sqrt{F_{1, df_2}} = +\sqrt{27.3615} = 5.23082$$