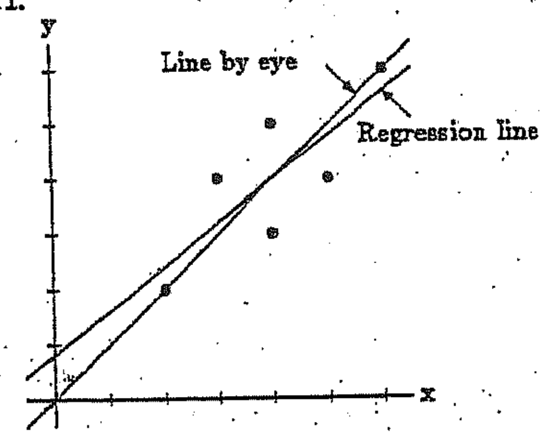
Chapter 7

Regression and Correlation

7.1 Answers to Exercises

7.1.1.



a.

b. The line you draw by eye probably goes through the points (2,2) and (6,6) with data points evenly scattered on both sides. This line would have an equation of $y = 0 + 1x$.

c.

y	x	xy	x^2
2	2	4	4
4	3	12	9
5	4	20	16
3	4	12	16
4	5	20	25
6	6	36	36
24	24	104	106

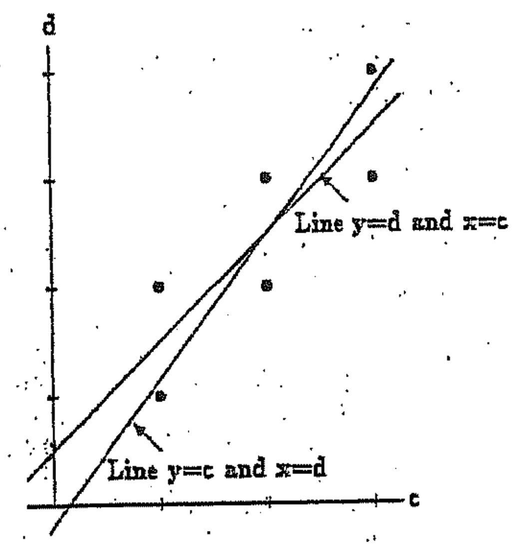
$$b = \frac{\sum xy - (\sum x)(\sum y)/n}{\sum x^2 - (\sum x)^2/n} =$$

$$\frac{104 - (24)(24)/6}{106 - (24)(24)/6} = 0.8 \text{ and}$$

$$a = \frac{\sum y}{n} - b\frac{\sum x}{n} = \frac{24}{6} - (0.8)\frac{24}{6} = 0.8 \rightarrow$$

$y = 0.8 + 0.8x$ Notice that this line is not the same as the line you drew based on eyeing the data points. That is because your eye tended to try to minimize the perpendicular distance of the points from the line as well as using the symmetry of the data points. The regression looks only at the vertical distance of the points from the line and "minimizes" that distance. Hence, the differing lines.

7.1.2.



a.

b. Regression line is $d = 0.5 + 1c$ and correlation is 0.8528.

c. Regression line is $c = \frac{1}{11} + \frac{8}{11}d$ and correlation is 0.8528. Notice that the lines do not coincide on the graph, but that the correlation is identical in both parts. The regression line when y=d and x=c looks at the vertical distance of points from the line, while the regression line when y=c and x=d looks at the horizontal distance of points from the line. The correlation in either case is measuring how close the points are to a perfect line and this does not change with the designation of y and x.

7.1.3.

a. Sales= -3971.5091+2.0364 Year, correlation is 0.9938, $s_e = s_{y|x} = 0.7324$ and $s_b = 0.0806$

b. -0.5291

c. Sales= -0.6+2.0364 x, correlation is 0.9938, $s_e = s_{y|x} = 0.7324$ and $s_b = 0.0806$ Notice that everything remains the same except the intercept. The new intercept should be the predicted sales in 1950 because x=0 corresponds to 1950 but the values are different due to rounding [actual intercept in (a) is -3971.5090909... and actual slope in (a) is 2.0363636...].

d. Sales= -102.4103+2.0364 x, correlation is 0.9938, $s_e = s_{y|x} = 0.7324$ and $s_b = 0.0806$

7.1.4.

a. Lb= -0.5333+2.0061 Ft, correlation is 0.9933, $s_e = s_{y|x} = 0.7497$ and $s_b = 0.0825$

b. Lb= -0.5333+0.1672 In, correlation is 0.9933, $s_e = s_{y|x} = 0.7497$ and $s_b = 0.0069$ Notice that the correlation remains the same because you are dealing with the same data, only in different units. The intercept also remains the same because when Ft=0, In=0 and y is still Lb. The slope is now one twelfth as big because a change of one foot is a change of twelve inches. s_b is also one twelfth as big because it is in the same units as the slope. s_e remains the same because the units of y are the same.

c. Oz= -8.5333+32.0970 Ft, correlation is 0.9933, $s_e = s_{y|x} = 11.9960$ and $s_b = 1.3207$ Notice that the correlation remains the same because you are dealing with the same data, only in different units. The intercept and slope are 16 times bigger because 16 oz=1 Lb. s_b is also 16 times bigger because it is in the same units as the slope. s_e is 16 times bigger because the units have changed by 16 times.

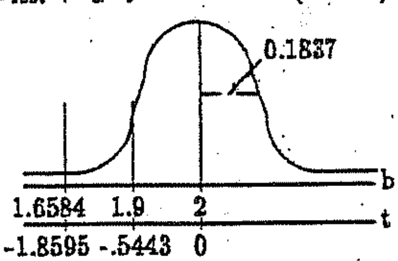
d. Oz= -8.5333+2.6747 In, correlation is 0.9933, $s_e = s_{y|x} = 11.9960$ and $s_b = 0.1101$

e. For length in inches, the regression equation is \$= -2.1333+0.6687 In, correlation is 0.9933, $s_e = s_{y|x} = 2.9990$ and $s_b = 0.0275$. For length in feet, the regression equation is \$= -2.1333+8.0242 Ft, correlation is 0.9933, $s_e = s_{y|x} = 2.9990$ and $s_b = 0.3302$

7.1.5.

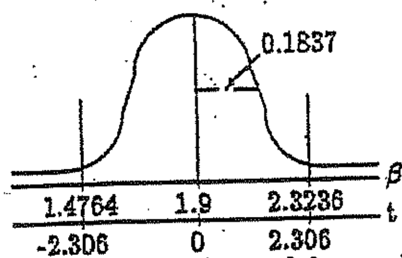
a. Quiz= -0.9+1.9 Homework

b. $H_o : \beta \geq 2$ vs. $H_a : \beta < 2$ From the sample, $b = 1.9, s_e = s_{y|x} = 0.8216$ and $s_b = 0.1837$ From the t table with $n-2 = 8df$ we get $t_{cr} = -1.8595$. Based on the sample $t^* = \frac{b-2}{s_b} = -0.5443$. You may sometimes wish to carry out the test in terms of b and b_{cr} rather than t^* and t_{cr}. In that case, $b_{cr} = \beta_{test} + t_{cr}s_b = 2 - 1.8595(0.1837) = 1.6584$



Because $b_{cr} = 1.6584$ is less than $b = 1.9$ and $t_{cr} = -1.8595$ is less than $t^* = -0.5443$ our conclusion is: do not reject the null hypothesis. Therefore, believe that each additional homework problem could lead to an increase of at least 2 points on the quiz.

c. Need confidence interval for β. $b \pm ts_b \rightarrow$
$1.9 \pm 2.306(0.1837) \rightarrow 1.9 \pm 0.4236 \rightarrow$
$1.4764 < \beta < 2.3236$



d. Need confidence interval for prediction of y given x. $\hat{y} \pm ts_{pred\ y|x} \rightarrow$

$$(a + bx) \pm ts_{y|x}\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x - \bar{x})^2}} \rightarrow$$

$$(-0.9 + 1.9[4]) \pm$$
$$(2.306)(0.8216)\sqrt{1 + \tfrac{1}{10} + \tfrac{(4-3)^2}{20}} \rightarrow$$
$$6.7 \pm (2.306)(0.8216)\sqrt{1.15} \rightarrow$$
$$6.7 \pm 2.306(0.8810) \rightarrow$$
$$4.6683 < y_{pred} < 8.7317$$

7.1.6.

a. Mistakes=16.45 - 1.65 Time

b. $H_o : \beta \leq -2$ vs. $H_a : \beta > -2$
$t^* = 1.74$ is greater than $t_{cr} = 1.397$ and $b = -1.65$ is greater than $b_{cr} = -1.72 \rightarrow$
Conclusion: reject H_o and believe that for each additional minute operators does not make at least 2 fewer mistakes.

c. $0.1757 < \mu_{y|x} < 3.0243$

7.2 Answers to Problems

7.2.1.

a. Expenses = 15.5476 - 0.5439 Mortgage, correlation is 0.9784, $s_e = s_{y|x} = 0.7312$ and $s_b = 0.0406$

b. Expenses = 4.1333 + 1.0667 Tax, correlation is 0.9688, $s_e = s_{y|x} = 0.8756$ and $s_b = 0.0964$

c. Mortgage is a better predictor because the correlation between Mortgage and Expenses is higher than the correlation between Tax and Expenses.

d. The magnitude of the slope does not indicate which is the better predictor because the independent variable are in different units. You must look at the slope in relation to s_b in order to decide. For instance, the slope for Mortgage is -0.5439 while its s_b is 0.0406. Hence, the slope for Mortgage is 13.3966 standard deviations away from zero. On the other hand, the slope for Tax is 1.0667 which seems larger, but its s_b is 0.0964. Therefore, the slope for Tax is 11.0654 standard deviations away from zero. Since a slope of zero indicates no relation, the variable whose slope is farther from zero has a stronger relation with the dependent variable Expenses.

e. Since the dependent variable Expenses has remained the same, the smaller s_e indicates the stronger relationship.

f. C.I. for $\beta \rightarrow \beta_L = 0.8443, \beta_U = 1.2889$ Expenses would differ by between \$8.44 and \$12.89

g. $\beta_L = -0.6801, \beta_U = -0.3745$ Expenses would differ between \$20.40 and \$11.24

7.2.2.

a. Commission = 0.5714 + 2 Sales

b. \$106.16 to \$408.11

c. ~~\$204.11 to \$225.26~~ This interval is much narrower, but it may not be legitimate to assume that all sales by a given saleperson yielded identical commissions. \$205.08 to \$223.49

d. ~~\$173.09 to \$255.48~~ This is not as narrow as the previous result because the assumption about identical commissions was wrong. \$179.74 to \$248.83

e. ~~\$692.30 to \$1021.92~~ \$718.96 to \$995.32

f. \$773.40 to \$940.88 This interval is narrower than the one above because the regression assumes roughly equal commissions for each additional sale.