Math 17, Section 2 – Spring 2011

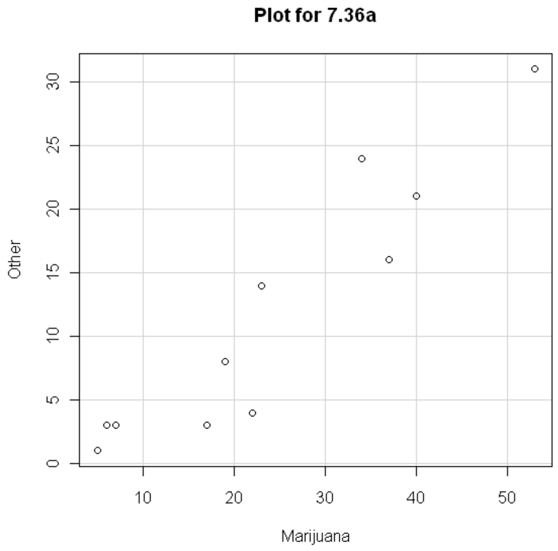
Homework 8 Solutions

<u>Assignment</u>
Chapter 7: 7.36, 7.40
Chapter 8: 8.14, 8.16, 8.28, 8.36 (a-d), 8.38, 8.62
Chapter 9: 9.4, 9.14

Chapter 7

7.36]

a) A scatterplot is given below. It doesn't become clear which is the explanatory variable and which is the response until part (d). Either plot is fine.



b) Using Rcmdr, the correlation is 0.9341.

```
>
cor(Drugs[,c("Marijuana","Ot
her")], use="complete.obs")
          Marijuana  Other
Marijuana    1.0000 0.9341
Other        0.9341 1.0000
```

c) We have a positive, fairly strong linear association between marijuana and other drugs. Countries with higher percentages of teens who have used marijuana tend to have higher percentages of teens that have used other drugs.

d) These results do not confirm that marijuana is a "gateway drug". An association exists between the percent of teens that have used marijuana and the percent of teens that have used other drugs. This does not mean that one caused the other.

7.40]

a) Winning teams generally enjoy greater attendance at their home games. The association between home attendance and number of wins is positive, somewhat straight, and moderately strong.

b) The association between winning and home attendance has the strongest correlation, with $r = 0.697$. This is only slightly higher than the association between home attendance and scoring runs, with $r = 0.667$.

c) The correlation between number of runs scored and number of wins is $r = 0.605$, indicating a possible moderate association. However, since there is no scatterplot of wins vs. runs provided, we can’t be sure the relationship is straight. Correlation may not be an appropriate measure of the strength of the association.

Chapter 8

8.14] Residuals.

a) The curved pattern in the residuals plot indicates that the linear model is not appropriate. The relationship is not linear.

b) The fanned pattern indicates heteroscedastic data. This is a fancy statistics term that means the equal variances assumption has been violated.

c) The scattered residuals plot indicates an appropriate linear model.

8.16] Roller coaster.

a) The explanatory variable (x) is initial drop, measured in feet, and the response variable (y) is duration, measured in seconds.

b) The units of the slope are seconds per foot.

c) The slope of the regression line predicting duration from initial drop should be positive. Coasters with higher initial drops probably provide longer rides.

8.28] Last ride.

a) According to the linear model, the duration of a coaster ride is expected to increase by about 0.242 seconds for each additional foot of initial drop.

b) $\widehat{Duration} = 91.033 + 0.242(Drop)$
$\widehat{Duration} = 91.033 + 0.242(200) = 139.433$

According to the linear model, a coaster with a 200 foot initial drop is expected to last 139.433 seconds.

c) $\widehat{Duration} = 91.033 + 0.242(150) = 127.333 \approx 2.12$ minutes

According to the linear model, a coaster with a 150 foot initial drop is expected to last 127.333 seconds. The advertised duration is shorter, at 120 seconds.

The difference is 120 seconds – 127.333 seconds = – 7.333 seconds. This is called a residual.

8.36 (a-d)] Interest Rates and mortgages again.

a) Yes, the regression seems appropriate. Both interest rate and total mortgages are quantitative, and the scatterplot looks pretty straight. The spread is fairly constant, and there are no outliers.

b) We can use the summary data to find the regression equation.

$$b_1 = r\frac{s_y}{s_x} = -0.84\left(\frac{23.86}{2.58}\right) = -7.768$$
$$b_0 = \bar{y} - b_1\bar{x} = 151.9 - (-7.768)(8.88) = 220.88$$

The regression equation is: $\widehat{Mortgage} = 220.88 - 7.768(Interest)$

c) With an interest rate of 20%, we'd predict a mortgage amount of
$\widehat{Mortgage} = 220.88 - 7.768(20) = 65.52$ million.

d) Yes, I would have reservations about this prediction. We are extrapolating beyond the range of our data. We cannot be sure the linear relationship still holds.

8.38] Online clothes II.

a) We can use the summary data to find the regression equation.

$$b_1 = r\frac{s_y}{s_x} = 0.722\left(\frac{253.62}{16952.50}\right) = 0.0108$$
$$b_0 = \bar{y} - b_1\bar{x} = 572.52 - (0.0108)(50343.40) = 28.81$$

The regression equation is: $\widehat{Purchases} = 28.81 + 0.0108(Income)$

b) Yes, the regression seems appropriate. Both variables are quantitative, and the scatterplot looks pretty straight. The spread is fairly constant, and there are no outliers.

c) With an income of \$20,000, we'd predict a total yearly purchases amount of
$\widehat{Purchases} = 28.81 + 0.0108(20000) = 244.81.$

With an income of \$80,000, we'd predict a total yearly purchases amount of
$\widehat{Purchases} = 28.81 + 0.0108(80000) = 892.81.$

d) This is the R^2 value. $R^2 = r^2 = 0.722^2 = 0.5213$. This means that 52.13% of the variation in purchases is explained by the linear regression with income.

e) This regression *may* be useful. The R^2 value isn't great, but the line is still explaining something.

8.62] Gators

a) Weight is the proper dependent variable. The researchers can estimate length from the air, and use length to predict weight

.

b) The correlation between an alligator's length and weight is
$r = \sqrt{R^2} = \sqrt{0.836} = \pm 0.914$. Since length and weight are positively associated, the correlation is r = 0.914.

c) The linear regression model that predicts and alligator's weight from its length is

$$\widehat{Weight} = -393 + 5.9\ (Length)$$

d) For each additional inch in length, the model predicts an increase of 5.9 pounds in weight.

e) The estimates made using this model should be fairly accurate. The model accounts for 83.6% of the variability in weight. However, care should be taken. With no scatterplot, and no residuals plot, we cannot verify the regression condition of linearity. The association between length and weight may be curved, in which case, the linear model is not appropriate.

Chapter 9

9.4] HDI revisited.

a) Fitting a linear model to the association between the number of cell phones and HDI would be misleading, since the relationship is not straight.

b) The residuals plot will be curved downward. Imagine a linear regression line through the points. At the left side, the line will be above the data point. Residuals will be negative. In the middle of the plot the line will be below the data points, so the residuals will be positive. At the right of the plot, the line is again above the data points, so the residuals will be negative again. So the residuals will go from negative to positive to negative, a downward curve.

9.14] The extra point revisited.

1) -0.45 → d. Point d is influential. Its addition will pull the slope of the regression line toward point d, resulting in the steepest negative slope, a slope of –0.45.

2) -0.30 → e. Point e is very influential, but since it is far away from the group of points, its addition will only pull the slope down slightly.
The slope is –0.30.

3) 0.00 → c. Point c is directly below the middle of the group of points. Its position is directly below the mean of the explanatory variable. It has no influence. Its addition will leave the slope the same, 0.

4) 0.05 → b. Point b is almost in the center of the group of points, but not quite. It has very little influence, but what influence it has is positive. The slope will increase very slightly with its addition, to 0.05.

5) 0.85 → a. Point a is very influential. Its addition will pull the regression line up to its steepest positive slope, 0.85.