**Openintro questions 7.1 to 7.29**

7.1 a. For residuals will lie very close to the regression line and no potential outliers.

b. For lower values of x the residual plots are away from the regression line as compared to higher values of x.

7.2

a. For higher values of x the residual plots are lie away from the regression line and variability is higher.

b. The residual plot seems to follow little curve but for lower values of x there are potential outliers lying far below the regression line.

7.3 a. Linear model won’t fit. The relationship is strong.

b. Linear model will fit. The relationship is strong.

c. The variability is high we can try fitting linear model here. The relationship is weak.

d. Linear model won’t fit. The relationship is moderate.

e. The relationship is strong. Linear model will fit well.

f. The relationship is weak and it’s hard to fit a linear model here.

7.4

a. Strong. No linear.

b. Strong. No linear.

c. Strong. Yes linear.

d. Weak. Try fitting linear.

e. Moderate. Linear will fit.

f. Moderate. Linear will fit.

7.5 a. Exam2 has the strongest correlation with final exam grade as the points lies more closer to the regression line, reducing the residual distance.

b. The variability is constant along the regression line.

7.6 a. Husband’s ages and wife ages are strongly related to each other as points are closely tied with no potential outliers.

b. Husband’s and wife’s heights are moderately related to each other as points are scattered all over the place.

c. Age plot shows a stronger correlation because of strong positive relationship between ages.

d. No it doesn’t affect the correlation.

7.7 a.4

b.3

c.1

d. 2

7.8 a.2

b. 4

c. 3

d. 1

7.9 a. True

b. False, Correlation is the measure of linear association between two variables.

7.10 Either of them will convert to the other’s unit and then compare the data points.

7.11 a. Height and fastest speed seems to be strongly correlated specially for the height values between 60 and 70.

b. We can see a linear positive trend.

c. Gender can be a confounding variable.

7.12

a. The volume and height have a moderate positive relationship.

b. The diameter and volume have a strong positive relationship.

c. We can use diameter to develop a association but can’t predict certainly.

7.13 a. The distance and travel time have a weak positive relationship with each other.

b. The relationship will remain same.

c. It will be same.

7.14 a. They have a weak negative linear relationship. We should note a potential outlier with temp value around 52.

b. No change

c. No change

7.15 a. The relationship is moderate, linear and positive.

b. No change

7.16 a. The relationship is strong, positive and linear.

b. No change

7.19 y = weight (kg) x=height (cm)

Slope = y/x = kg/cm

y- Intercept = kg

correlation = no units

7.20 The uncertainty is higher when there is a lot of scatter around the regression line.

7.21 Residual = observed – predicted. Since, the residual is negative the predicted value is higher hence we have over estimated the value.

7.22 observed – predicted = observed – 1.5 = residual. Since the residual is positive the observed value is greater than the predicted value and we underestimated the value from regression line.

7.23 a. The relationship between no of tourists and spending is linear, strong, positive with no potential outliers.

b. Explanatory variable is number of tourists. Response variable is spending.

c. Since, both the variables have a strong positive relationship we can think of drawing a regression line here.

d. The residual plot doesn’t show linearity so fitting a least regression line might not be good here.

7.24

a. Calories and carbs have a moderate, positive relationship. There are two possible potential outliers which can be seen from the scatter plot.

b. Explanatory – No of calories. Response- Carbs.

c. From the scatter plot the relationship is seen to be linear and positive so we can try fitting a regression line here.

d. The residual histogram doesn’t show a linear relationship. The residual plot shows a high variability for large values of calories.

7.25 a. Mean travel time= 129mins Travel time sd=113mins

Mean distance = 108miles Sd distance=99miles

R=0.636

1. y-129 = b1 \* (x-108)

b1 = sy/sx \* R = 113/99 \* 0.636 = 0.72593 y-129 = 0.725(x-108)

1. For each additional mile the time will increase by 0.726 minutes.

Intercept = b0 = 51. When the distance traveled is 0 the travel time is expected to be 51 minutes.

1. R^2 = 0.636^2 = 0.40. Accounts for variability. 40% of the variability in the travel time is accounted for by the model.
2. Distance = 103miles y=51+0.726\*103=125.778
3. Residual = observed – predicted = 168-126=42 is the residual. Since it is positive we have under estimated the value.
4. No, this needs extrapolation.

7.26 mean sg = 107.20 cm sd sg=10.37 mean hei=171.14 sd hei=9.41 R=0.67

a. y-171.14 = 0.61(x-107.20) b1= 9.41/10.37 \* R = 0.6079749

y-171.14 = 0.061x - 65.392 => y = 0.061x + 105.748

b. slope b1 = 0.61 intercept = 105.75

c. R^2 = 0.67^2 = 0.4489

d. height = 105.75 + 0.61\*100 = 166.75

e. 160-166.75 = -6.75. It is negative we have over estimated the height of the boy.

f. No this needs extrapolation.

7.27 Both have a weak positive relationship with high variability for large values of calories.

7.28 a. correlation = sqrt(R) = sqrt(0.72) = 0.8485281

b. b1 and b0.

y – 38.8 = b1(y-30.8) b1 = 16.9/26.7\*R= 0.538015

y = 0.54x + 22.168

c. b0 = 22.168

d. slope = 0.54

e. 40-30.8 = 9.2. We underestimated the value

40-38.8=We underestimated the value

7.29

a. y=-29.901 + 2.559x

b. -29.901. For 0% poverty the annual murders will be negative.

c. For every additional percent poverty there will be additional 2.559% change in annual murders.

d. 68.89% variability for annual murders is covered by this model.

e. r= sqrt(0.68)=0.82