STA101 Problem Set 8 - KEY

Summer I, 2021, Duke University

Exercises from the OpenIntro book - 100pts total

Problems include: Chapter 8 exercises 8.4, 8.21, 8.29, 8.42 and 2 additional problems

8.4 (12pts total)

(a) **(1pt for strength, 1pt for reasonableness of linear model)** Strong relationship, but a straight line would not fit the data.

(b) **(1pt for strength, 1pt for reasonableness of linear model)** Strong relationship, but a straight line would not fit the data.

(c) **(1pt for strength, 1pt for reasonableness of linear model)** Strong relationship, and a linear fit would be reasonable.

(d) **(1pt for strength, 1pt for reasonableness of linear model)** Weak relationship, and trying a linear fit would be reasonable.

(e) **(1pt for strength, 1pt for reasonableness of linear model)** Weak relationship, and trying a linear fit would be reasonable.

(f) **(1pt for strength, 1pt for reasonableness of linear model)** Moderate relationship, and a linear fit would be reasonable.

8.21 (16pts total)

(a) **(4pts for reasonable description)** There is a positive, very strong, linear association between the number of tourists and spending.

(b) **(2pts "tourists" for explanatory, 2pts for "spending" for response)** Explanatory: number of tourists (in thousands). Response: spending (in millions of US dollars).

(c) **(4pts for reasonable explanation)** We can predict spending for a given number of tourists using a regression line. This may be useful information for determining how much the country may want to spend in advertising abroad, or to forecast expected revenues from tourism.

(d) **(2pts for "no", 2pts for explanation)** Even though the relationship appears linear in the scatterplot, the residual plot actually shows a nonlinear relationship. This is not a contradiction: residual plots can show divergences from linearity that can be difficult to see in a scatterplot. A simple linear model is inadequate for modeling these data. It is also important to consider that these data are observed sequentially, which means there may be a hidden structure not evident in the current plots but that is important to consider.

8.29 (8pts total)

(a) **(4pts for reasonable description)** There is a negative, moderate-to-strong, somewhat linear relationship between percent of families who own their home and the percent of the population living in urban areas in 2010. There is one outlier: a state where 100% of the population is urban. The variability in the percent of homeownership also increases as we move from left to right in the plot.

(b) **(2pts for "leverage", 2pts for "influential")** The outlier is located in the bottom right corner, horizontally far from the center of the other points, so it is a point with high leverage. It is an influential point since excluding this point from the analysis would greatly affect the slope of the regression line.

8.42 (22pts total)

(a) **(4pts)** $r = \sqrt{R^2} = -0.849$

(b) **(3pts for slope, 3pts for intercept)** $b_1 = \frac{s_y}{s_x} r = \frac{16.9}{26.7}(-0.849) = -0.537$
$b_0 = \overline{y} - \overline{x} b_1 = 38.8 - 30.8(-0.537) = 55.34$

(c) **(4pts)** For a neighborhood with 0% reduced-fee lunch, we would expect 55.34% of the bike riders to wear helmets.

(d) **(4pts)** For every additional percentage point of lunch there is a decrease of 0.537 percentage points in helmet.

(e) **(2pts for residual, 2pts for interpretation)** $e = 40 - \widehat{y} = 30 - (30 \cdot -0.537 + 55.34) = 6.14$. There are 6.14% more bike riders wearing helmets than predicted by the regression model in this neighborhood.

Problem 5 (22pts total)

(a) **(2pts for slope and intercept, 2pts for correct representation of regression line)** First calculate the slope

$$b_1 = \frac{s_y}{s_x} R = \frac{3.495}{2.573}(0.91) = 1.236$$

Next make use of the fact that the regression line passes through the point $(\overline{x}, \overline{y})$.

$$\overline{y} = b_0 + b_1 \overline{x}$$

$$15.58 = b_0 + (1.236)(12.74)$$

$$b_0 = -0.167$$

The regression line can be written as

$$\widehat{\text{Frontal Lobe}}\text{ Size} = -0.167 + 1.236\text{Rear Width}$$

(b) **(2pts for slope interpretation, 2pts for intercept interpretation)** $b_1 =$ For each 1mm increase in the rear width, the frontal lobe size increases by 1.236mm.
$b_0 =$ Crabs with with a rear width of 0mm have a frontal lobe size of -0.167mm. Here, the y-intercept serves only to adjust the height of the line and is meaningless by itself.

(c) **(2pts for R^2, 2pts for interpretation)** $R^2 = 0.91^2 = 0.8281$. Approximately 82.81% of the variation in frontal lobe sizes is accounted for by the model, i.e., explained by rear width.

(d) **(2pts)**

$$\widehat{\text{Frontal Lobe}} \text{ Size} = -0.167 + (1.236)(14.5) = 17.755\text{mm}.$$

(e) **(2pts for residual, 2pts for interpretation)** The residual can be calculated as $e_i = y_i - \widehat{y}_i = 16.8 - 17.755 = -0.955$. A negative residual means that the model overestimated the frontal lobe size.

(f) **(2pts for "no", 2pts for explanation)** No. A rear width of 2.69mm is far away from the typical observations in the data used to fit the regression line, so estimating the frontal lobe size associated to this rear width would require extrapolation. The linear model may no longer hold outside the range of the data.

Problem 6 (20pts total)

(a) **(6pts for hypotheses)** If we write down the linear model as

$$\widehat{\text{Distance}} = b_0 + b_1 \text{Speed},$$

then the hypotheses are $H_0 : \beta_1 = 0$ and $H_A : \beta_1 \neq 0$.

(b) **(4pts for "reject H_0")** The p-value for this test is approximately 0, therefore we reject H_0. The data provide convincing evidence that car speed is a significant predictor of stopping distance.

(c) **(3pts for CI, 2pts for interpretation)** $n = 50$, $df = 48$, $T^*_{48} = 2.01$.

$$CI = 3.9324 \pm (2.01)(0.4155) = (3.097, 4.768)$$

For each additional mile per hour, the distance to stop the car is expected to be take on average 3.097 to 4.768 feet more feet to stop.

(d) **(3pts for "yes", 2pts for explanation)** Yes, we rejected H_0 and the confidence interval does not include 0.