# Assignment 5

Sri Harsha Sudalagunta

2023-03-25

## Problem 1.

1. (3 points)

Identify relationships,

For each of the six plots, identify the strength of the relationship (e.g. weak, moderate, or strong) in the data and whether fitting a linear model would be reasonable.

(a) The strength of the relationship seems to be strong but the linear model doesn't fit the data.

(b) The strength of the relationship is strong and fitting a linear model will be reasonable.

(c) The strength of the relationship is weak and may be can try fitting the linear model but the model performance will be low.

(d) The strength of the relationship is moderate and a linear model would not fit the data.

(e) The Strength of the relationship is strong and fitting a linear would be reasonable

(f) The strength of the relationship is weak, trying a linear fit would be reasonable.

## Problem 2.

2. (2 pts)

Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender for 507 physically active individuals.28 The scatterplot below shows the relationship between height and shoulder girth (over deltoid muscles), both measured in centimeters.

(a) Describe the relationship between shoulder girth and height. (1 point)

The relationship between the shoulder girth and height is strong and linear relationship, as we can see the positive slope between them.

(b) How would the relationship change if shoulder girth was measured in inches while the units of height remained in centimeters? (1 point)

The change in the units doesn't change the relationship between the height and the shoulder girth. Even if we change the unit of height, the relationship remains the same.

## Problem 3

3. (1 pt)

Suppose we fit a regression line to predict the shelf life of an apple based on its weight. For a particular apple, we predict the shelf life to be 4.6 days. The apple's residual is -0.6 days. Did we over or under-estimate the shelf-life of the apple? Explain your reasoning.

The shelf-life of the apple is over-estimated. Because we calculate the residual as observed - predicted. So, the negative i.e., -0.6 means we over predicted the shelf life by 0.6 days.

## Problem 4.

4. (2 pts)

Guppies are small, brightly colored tropical fish often seen in freshwater fish aquariums. A study was conducted on 147 male guppies to examine the relationship between coloration and heterozygosity; heterozygosity refers to the condition of having different alleles at a given genetic locus. The guppies were randomly sampled from a river in the wild. In the initial stage of the study, researchers examined whether length and height are linearly associated. The mean length is 1261.21 cm, with standard deviation 95.62 cm. The mean height is 201.75 cm, with standard deviation 20.68. The correlation between length and height is 0.85.

(a) From a visual inspection, does it seem like the line is a reasonable fit for the data? (0.5 pts)

Yes, I think the line is reasonable fit for the data.

(b) Write the equation of the regression line for predicting length from height. (1 pt)

The equation for the regression line predicting length from height is

Y = b0+ b1.x

Where

x is the height, for predicting the length

b0 is the intercept

b1 is the slope of the line

b1 = r(sy/sx)

b1 = 0.85(95.62/20.68)

b1 = 3.93

b0 = Y-b1(x)= 1261.21-3.93(201.75) = 1261.21-792.87 = 468.34

The equation of the regression line for predicting length from height is

Y = 468.34 + 3.93(x)

(c) Estimate the predicted mean length of a guppy with height 180 cm. (0.5 pts)

The predicted mean length of a guppy with height 180cm is The equation is Y = 468.34 + 3.93(x)

Y = 468.34 +3.93(180) = 1175.74

## Problem 5.

5. (2 pts)

Visualize the residuals.

The scatter plots shown below each have a superimposed regression line. If we were to construct a residual plot (residuals versus x) for each, describe what those plots would look like.

(a) The residual plot looks some what randomly distributed residuals around the 0.

(b) The residual will show a fan shape, with higher variability for smaller values of the x.

## Problem 6.

6. The scatter plot below shows the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain.29 Since Starbucks only lists the number of calories on the display items, we are interested in predicting the amount of carbs a menu item has based on its calorie content. (2 points)

(a) Describe the relationship between number of calories and amount of carbohydrates (in grams) that Starbucks food menu items contain.

The relationship between the calories and the carbohydrates is strong, positive and linear relationship.

(b) In this scenario, what are the explanatory and response variables?

The explanatory variable is number of calories and the response variable is number of carbohydrates.

(c) Why might we want to fit a regression line to these data?

As there is a relationship between the number of calories and amount of carbohydrates, we can build a regression line.

(d) Do these data meet the conditions required for fitting a least squares line?

No, the data doesn't meet the conditions required for fitting a least squares line.

## Problem 7.

7. Based on the scatter plot and the residual plot provided, describe the relationship between the protein content and calories of these menu items, and determine if a simple linear model is appropriate to predict amount of protein from the number of calories. (1 point)

There is a positive and upward trend relationship between the protein content and calories. we can see the variance is more for higher value counts of calories. We should be cautious while using the simple linear model, as we can see there is an increased variance with increase in the calorie count.

## Problem 8.

8. Identify the outliers in the scatter plots shown below and determine what type of outliers they are. Explain your reasoning. (3 points)

(a) The outlier is in the bottom right. we can see it is far away from the remaining data, so it is a point with high leverage. It is an influential point, which affected the slope of the regression line.

(b) The outlier is in the bottom right. It is far away from the center of the data, which is a point with high leverage. As the outlier is within the regression line, it doesn't have much much influence on the slope of the line.

(c) The observation is seen in the center of the data in the x-axis direction, so it doesn't have high leverage. This point wont have much effect on the slope of the line and it is not an influential point.

(d) The outlier is seen in the top left. It is far away from the centre of the data, so it is a point with high leverage. It is an influential point, which will affect the slope of the regression line.

(e) The outlier is seen in the bottom left. It is far away from the remaining data, so it is a point with high leverage. It is an influential point, which will affect the slope of the regression line.

(f) The observation is seen in the center of the data in the x-axis direction, so it doesn't have high leverage. This point wont have much effect on the slope of the line and it is not an influential point.

## Problem 9.

9. Many people believe that gender, weight, drinking habits, and many other factors are much more important in predicting blood alcohol content (BAC) than simply considering the number of drinks a person consumed. Here we examine data from sixteen student volunteers at Ohio State University who each drank a randomly assigned number of cans of beer. These students were evenly divided between men and women, and they differed in weight and drinking habits. Thirty minutes later, a police officer measured their blood alcohol content (BAC) in grams of alcohol per deciliter of blood.30 The scatter plot and regression table summarize the findings.

(a) Describe the relationship between the number of cans of beer and BAC. (0.5 points)

From the plot, we can see that there is a strong positive relationship between the cans of beer and BAC.

(b) Write the equation of the regression line. Interpret the slope and intercept in context. (0.5 points)

The regression equation is

Y = b0 + b1(x)

From the regression table, the value of intercept is –0.0127 and the value of slope is 0.0180

Therefore,

Y = (-0.0127) + 0.0180(x)

(c) Do the data provide strong evidence that drinking more cans of beer is associated with an increase in blood alcohol? State the null and alternative hypotheses, report the p-value, and state your conclusion. (2 points)

Null Hypothesis:

There is no increase in blood alcohol with drinking more cans of beers.

Alternate Hypothesis:

There is an increase in blood alcohol with drinking more cans of beers.

As the given p-value is less than 0.05,we can reject the null hypothesis and conclude that there is strong evidence that drinking more cans of beer is associated with an increase in blood alcohol.

(d) The correlation coefficient for number of cans of beer and BAC is 0.89. Calculate R 2 and interpret it in context. (1 points)

Given R = 0.89

R2 = 0.89*0.89

R2 = 0.7921

The value of coefficient of determination is 0.7921, it can be interpreted as 79.21% of the total observed variance in the independent variable (BAC) is being explained by the independent variable