Week 7 Homework

Steven Simonsen

2024-02-27

8.6 Husbands and wives, Part I. The Great Britain Office of Population Census and Surveys once collected data on a random sample of 170 married couples in Britain, recording the age (in years) and heights (converted here to inches) of the husbands and wives.5 The scatterplot on the left shows the wife's age plotted against her husband's age, and the plot on the right shows wife's height plotted against husband's height.

(a) Describe the relationship between husbands' and wives' ages.

The relationship between husbands' and wives' ages appears to be positive, very strong, and linear. There are some outliers in which both positive residuals (wives' age significantly greater than husbands' age), and negative residuals (husbands' age significantly greater than wives' age) exist. Although these outliers exist, they do not appear to be statistically significant as they do not deviate very far from the main cluster of data points.

(b) Describe the relationship between husbands' and wives' heights.

The relationship between husbands' and wives' heights also appears positive, weak, and possibly linear. If greater analysis were to be conducted, I would expect R (correlation) to be positive, but a lesser value than the graph used to compare ages. Again, there are outliers creating both positive and negative residuals. Additionally, some of the outliers appear that they could be statistically significant. This appears to happen in more on the side of negative residuals, meaning the husbands' height is significantly greater than the wives' height.

(c) Which plot shows a stronger correlation? Explain your reasoning.

Plot a appears to show a stronger correlation. The scatter of various plots comprising the overall dataset appear to be tightly grouped, whereas plot b appears to show a much less tightly grouped dataset. Additionally, the outliers appear to be more closely grouped to the larger cluster of data points in plot a. In plot b, there appear to be outliers further away from the larger cluster of data points, resulting in less correlation. For these reasons, plot a appears to show a stronger correlation.

(d) Data on heights were originally collected in centimeters, and then converted to inches. Does this conversion affect the correlation between husbands' and wives' heights?

No, changing the conversion between units does not change the form, direction, or strength of the relationship between the two variables.

8.22 Nutrition at Starbucks, Part I. The scatterplot below shows the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain.15 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.

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

The relationship between number of calories and amount of carbohydrates (in grams) is positive, weak to moderate strength, and possibly linear, although the variability of errors also appears as though it could be related to the value of x, or calories in this case.

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

Explanatory: Calories. Response: Carbs (grams)

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

We can possibly predict carbs (grams) based on the number of calories food items contain. This could be useful information for all types of people, one in particular being nutritionists helping to shape a client's diet to properly balance nutrients. I would like to emphasize the words "possibly predict" because the visible correlation based on the plots provided appears to be moderate, at most.

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

No, while the data may be slightly linear, I would not justify the linearity enough to fit a least squares line, which assumes linearity. Additionally, there is not constant variability. As previously described, the variability of errors appears to be related to the value of x, or calories. In other words, as the number of calories in a menu item increases, the variability in the number of carbs (grams) from the regression line also appears to increase. Finally, the residual data does not appear to be normal. It is skewed left, and has a bi-modal distribution. This is, in part, due to how far some of the outliers in the initial plot lie in relation to the regression line.

8.32 Beer and blood alcohol content. 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. The scatterplot and regression table summarize the fndings.

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

The relationship between the number of cans of beer and BAC appears positive, moderate, and linear. There is one outlier that appears to be influential, and I would expect this to pull the regression line up on the right side of the graph.

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

BAC (grams / deciliter) hat = -0.0127 + 0.0180 x Cans of beer.

Slope: For each additional can of beer, the model predicts the BAC to be 0.0180 additional grams / deciliter.

Intercept: Students who consume 0 cans of beer have a BAC of -0.0127. It is obviously not possible to possess a negative BAC, therefore the intercept exists only as a means of adjusting the height of the line and is meaningless by itself.

(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.

H0=The true slope coefficient of Cans of Beer is 0 (B1=0). HA=The true slope coefficient of Cans of Beer is different than 0 (B1 not= 0). Looking at the second row of the table (beers), the p-value for the two-sided alternative hypotheses (B1 not= 0) is incredibly small, so we reject H0. The data provide convincing evidence that Cans of Beer and BAC (grams / deciliter) are positively correlated. The true slope parameter is therefore greater than 0.

(d) The correlation coefficient for number of cans of beer and BAC is 0.89. Calculate R2 and interpret it in context.

0.89ˆ2 Rˆ2 = 0.79

Approximately 79% of the variability in BAC can be explained by the Cans of Beer a person consumes.

(e) Suppose we visit a bar, ask people how many drinks they have had, and also take their BAC. Do you think the relationship between number of drinks and BAC would be as strong as the relationship found in the Ohio State study?

Considering the R squared value of 0.79 and small p-value of approximately 0, I would expect the relationship between number of drinks and BAC to be near as strong as the relationship found in the Ohio State study. Since the p-value is 0, this would mean that in cases where H0 is true, we would not incorrectly reject it more than 5% of the time. Given that 79% of the variability in BAC is explained by Cans of Beer in the Ohio State model, this can be used to predict a strong relationship between the two variables outside of the study. However, it should be noted that the R squared model does not necessarily predict how "good" the model is. It may be possible that other unseen factors come into play, such as the weight or age of the person. For example, if the bar visited is outside of a college campus, the mean for a persons age may differ drastically, causing varying results. Similarly, the mean weight of Ohio State students may be higher or lower than the mean weight of the participants from another bar, also potentially causing varying results.