charleyferrari_week7hw

Charley Ferrari

November 3, 2015

Question 7.24, Nutrition at Starbucks, Part I

The scatterplot below shows the relationship between the number of calories and amount of carbohydrates Starbucks food menu items contain. Since Starbucks only lists the number of calories on the display items, we are interested in predicing the amount of carbs a menu item has b ased on its calorie content.

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

This relationship appears to be linear, with an medium sized R-squared value. The residuals are pretty normally distributed, indicating that there are no underlying patterns we're missing.

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

In this example, calories is the explanatory variable and amount of carbs is the response variable.

c. Why might we want to fit a regression line to these data? The data appear to be linearly related, so a regression line may help us estimate the amount of carbs in random menu items based on the calorie count.

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

Not necessarily. The relationship seems to be normal and there are nearly normal residuals, but the variability isn't constant, and seems to increase with higher calorie foods. We can't really comment on the independence of the observations, but given the nature of the study (different types of food being the cases,) this might be a sound assumption.

7.26: Body Measurements, Part III

The mean shoulder girth is 107.2cm with a standard deviation of 10.37cm. The mean height is 171.14cm with a standard deviation of 9.41cm. The correlation between height and shoulder girth is 0.67.

a. Write the equation line for predicting height.

b_1, the slope, is equal to $\frac{s_y}{s_x}R$, so:

```
b1 <- (9.41/10.37)*0.67
```

$\bar{x}$ and $\bar{y}$ are both on the line, so we can use the point slope form of the equation:

$$y - \bar{y} = 0.6(x - \bar{x})$$
$$y - 171.14 = 0.6(x - 107.2)$$
$$y = 0.6x + 106.82$$

b. Interpret the slope and the intercept in this context.

The slope in this case explains how much taller someone with different shoulder girths might be. In other words, given a change in shoulder girth, the slope will tell us the expected height difference (not in a causal way...)

The intercept isn't too important in this case, since these measurements are never expected to be 0.

c. Calculate R^2 of the regression line for predicting height from shoulder girth, and interpret it in the context of the application

```
R2 <- 0.67^2

R2
```

```
## [1] 0.4489
```

R-squared is a measure of how muuch of the data's variation is described by this relationship. An R-squared of 0.44 means that 44% of the data's variation is explained by using this information.

d. A randomly selected student from your class has a shoulder girth of 100cm. Predict the height of this student using the model.

```
height = 0.6*100 + 106.82
height
```

```
## [1] 166.82
```

e. The student from part d is 160cm tall. Calculate the residual, and explain what this residual means.

The residual in this case is 6.82cm, and refers to the deviation of this particular datapoint. This particular student didn't vary too far from the model.

f. A one year old has a shoulder girth of 56cm. Would it be appropriate to use this linear model to predict the height of this child?

We don't have enough information to answer this question. If the samples were chosen solely from adults, it would not be appropriate to use this linear model to predict the height of this child. What we should do is add age as a second variable, so we can further clarify the model as it applies to age.

7.30: Cats, Part I

The following regression output is for predicting the heart weight of cats from their body weight. The coefficients are estimated using a dataset of 144 domestic cats.

a. Write out the linear model

$$HeartWt = 4.034 * BodyWt - 0.357$$

b. Interpret the intercept.

As with the previous problem, this intercept doesn't mean much, it simply clarifies the relationship. We're not expected to see cats with 0 weight!

c. Interpret the slope.

The slope shows the amount of change in heart weight expected per change in body weight (adjusted for the different units of measurement).

d. Interpret R^2

The R^2 shows the relationship between these two variables. In this case, 65% of the variation in heart weight is explained by a cat's body weight.

e. Calculate the correlation coefficient.

This is just R:

```
R <- sqrt(0.6466)

R
```

```
## [1] 0.8041144
```

7.40: Rate my Professor

Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the usue of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because those measures may reflect the influence of non-teaching related characteristics, such as the physical appearance of the instructor. Researchers at University of Texas, Austin collected data on teaching evaluation score and standardized beauty score for a sample of 463 professors. The scatterplot below shows the relationship between these variables, and also provided is a regression output for predicting teaching evaluation score from beauty score.

a. Given that the average standardized beauty score is -0.883 and average teaching evaluation score is 3.9983, calculate the slope:

$$y - \bar{y} = b_1(x - \bar{x})$$

b. Do these data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive? Explain your reasoning.

The very low P-value for the slope suggests the slope is definitely positive (at least to 4 decimal places!)

c. List the conditions required for linear regression and check if each one is satisfied for this model based on the following diagnostic plots.

Linearity, Nearly normal residuals, constant variability, Independent Observations

Linearity: The data does appear to be linear, although there is a high variability among the data.

Nearly normal residuals: The residuals appear to be normal enough.

Constant variability: The variability seems constant, but also seems rather high. The scatterplot of the original data doesn't show a "tight" relationship. There appear to be very high p-values for the intercept and slope however.

Independent Observations: This can't be proven, but given the fact that this is a random sample of professors one can assume that these observations are indeed independent.

summary(m1)