IS606 Homework 7

Daniel Dittenhafer

November 5, 2015

7.24 Nutrition at Starbucks, Part I (p363)

a) Describe the relationship between number of calories and amount of carbohydrates (in grams) that Starbucks food menu items contain. Calories and carbohydrates have a positive relationship. As calories increase, so do carbohydrates.

b) In this scenario, what are the explanatory and response variables? The explanatory variable is Calories on the x axis, and the response variable is Carbohydrates on the y axis.

c) Why might we want to fit a regression line to these data? We might be interested in predicting the number of carbohydrates in a particular food item based on the number of calories.

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

- **Linearity**: The data is not clearly linear, but there does appear to a general trend. It appears more fan-like than linear.
- **Nearly normal residuals**: The residuals distribution appears nearlly normal.
- **Constant variability**: I don't think I would go so far as constant variability. The residuals on the left of the Residuals scatterplot (lower calories) are smaller, and trend larger as we move up the calories axis.
- **Independent observations**: Each menuitem is presumably independent of the next, but they are all Starbucks menu items (as opposed to separate vendors).

In conclusion, the data don't seem to sufficiently meet the conditions for fitting a least squares line due primarily to non-satisfaction of the Linearity and Constant Variability conditions.

7.26 Body measurements, Part III (p364)

Mean shoulder girth is 107.20 cm. Standard deviation: 10.37 cm. Mean height is 171.14. Standard deviation: 9.41 cm. Correlation between height and shoulder girth is 0.67.

a) Write the equation of the regression line for predicting height. First we compute the Beta estimates:

```
b1 <- (9.41 / 10.37) * 0.67
b0 <- (b1 * (0 - 107.2)) + 171.14
```

$$\hat{y} = 105.9650878 + 0.6079749x$$

b) Interpret the slope and the intercept in this context. The intercept is the height in centimeters at girth of 0 cm. The slope is the number of of centimeters increase in height for each increase in shoulder girth.

c) Calculate R^2 of the regression line for predicting height from shoulder girth, and interpret it in the context of the application. First compute the R^2 using the given correlation value:

```r
R2 <- 0.67^2
```

The $R^2 = 0.4489$. This means the linear model explains ~45% of the variation of the height data.

d) A randomly selected student from your class has a shoulder girth of 100cm. Predict the height of this student using the model. First we define a function of the model, then call it to get our prediction:

```r
heightFromShoulderGirth <- function(girth)
{
  height <- b0 + b1 * girth
  return(height)
}

predHeight <- heightFromShoulderGirth(100)
predHeight
```

```
## [1] 166.7626
```

e) 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? Using the previously defined function, we get the predicted value out of interest. Given the data set's lowest shoulder girth value was > 80cm, any value below this boundary may not be handled appropriately. In addition, a one year old is much different than physically active individuals which were the basis for the sample. In conclusion, I don't believe it would be appropriate to use this linear model to predict the height of this child.

```r
predHeight <- heightFromShoulderGirth(56)
predHeight
```

```
## [1] 140.0117
```

7.30 Cats, Part I (p365)

a) Write out the linear model.

$$\hat{y} = -0.357 + 4.034x$$

b) Interpret the intercept The value -0.357 tells us the model will predict a negative heart weight when the cat's body height is 0. This of course doesn't make any sense since the heart is part of the body, therefore this value is not meaningful when standing alone.

c) Interpret the slope The slope of 4.034 tells us that the heart weight increases by 4.034 grams for each 1kg of body weight increase.

d) Interpret R^2 The $R^2 = 64.66\%$ tells us that the linear model describes 64.66% of the variation in the heart weight.

e) Calculate the correlation coefficient. The correlation coefficient would be the square root of the R^2 value.

```r
r <- sqrt(0.6466)
r
```

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

7.40 Rate my professor (p369)

a) Given that the average standardized beauty score is -0.0883 and average teaching evaluation score is 3.9983, calculate the slope. Alternatively, the slope may be computed using just the information provided in the model summary table. Using an algebraic manipulation of the slope, intercept from of the linear model in `R` code:

```r
b1 <- (3.9983 - 4.010) / -0.0883
b1
```

```
## [1] 0.1325028
```

We find our estimate of $\beta_1 = 0.1325028$.

b) Do these data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive? Explain your reasoning. Given the p-value for the slope is ≈ 0.0000, this provides strong evidence that the slope is not 0 on a two tail test. Looking at the positive t-value of 4.13 and half the two-tail p-value (still very close to zero), this provides strong evidence of a positive relationship between teaching evaluationa and beauty.

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**: There is a weak trend in the scatterplot. We are not provided the correlation coefficient nor the R^2, therefore we will accept, with concerns, the linearity condition is satisfied.
- **Nearly normal residuals**: As shown in the residuals distribution and Q-Q plot, they are in fact nearly normal.
- **Constant variability**: The scatterplot of the residuals does appear to have constant variability.
- **Independent observations**: Assuming independence due to no clear evidence one way or the other. 463 professors would likely be $< 10\%$ of nationwide professors, so this is a plus.