DATA 606 Homework 7

Ilya Kats

April 23, 2017

7.24 Nutrition at Starbucks, Part I.

a. The relationship between number of calories and amount of carbs may be linear, but not very strong. In fact, to me it looks like the small cloud of points in the lower left corner (low calorie, low carb) is forcing the linear relationship. It is positive if it exists.

b. Number of calories is the explanatory variable. Amount of carbs is the response variable.

c. We may want to fit a regression line to have a way to predict the amount of carbs based on number of calories (perhaps for a new food item or for an item where amount of carbs is not readily available).

d. **Linearity:** As described in part (a), there may be a weak linear relationship. **Nearly normal residuals:** The histogram of the residuals is not completely symmetrical and may not necessarily be nearly normal. **Constant variability:** Based on the residual plot, I believe there is no contant variability as there are significantly more points on the right (with larger residuals) than on the left of the plot. **Independent observations:** Observations are independent since food items and their nutritional information does not depend on each other. *I believe conditions **are not met** because of lack of constant variability and distribution of residuals that is not nearly enough normal.*

7.26 Body measurements, Part III.

$\bar{x} = 107.2$, $\bar{y} = 171.14$

$s_x = 10.37$, $s_y = 9.41$

$R = 0.67$

a. $b_1 = \frac{s_y}{s_x}R = \frac{9.41}{10.37} * 0.67 = 0.6079749$

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

$y - 171.14 = 0.6079749 * (x - 107.2)$

$y = 0.6079749 * x - 0.6079749 * 107.2 + 171.14$

$y = 105.9651 + 0.6079749 * x$

Equation of the regression line:

$\widehat{height} = 105.9651 + 0.6079749 * shoulder\ girth$

b. The **slope** means that for every additional centimeter of shoulder girth the average height increases by 0.6079749 centimeters. The **intercept** means that for a shoulder girth of 0 centimeters, the average height is 105.9651 centimeters (obviously not a realistic point that serves only as a description of the regression line equation).

c. $R^2 = 0.67^2 = 0.4489$ 44.89% of the variation in height is explained by shoulder girth.

d. If $x = 100$, then height is predicted to be $\hat{y} = 105.9651 + 0.6079749 * 100 = 166.7626$ cm.

e. The residual is $e_i = y_i - \hat{y} = 160 - 166.7626 = -6.7626$ cm. A negative residual means that the model overestimated the height.

f. The data covers shoulder girth in the range of about 85 cm to 135 cm. The value of 56 cm lies far outside of the range. It is **not appropriate** to use this linear model to predict the height.

7.30 Cats, Part I.

a. $\widehat{heart\ weight} = -0.357 + 4.034 * body\ weight$

b. The intercept means that for a body weight of 0 kg, the average heart weight is -0.357 grams. It is an obviously theoretical example useful only to intepret the linear model.

c. The slope means that for each additional kilogram of body weight, the average heart weight of a cat increases by 4.034 grams.

d. 64.66% of the variability in heart weight of cats can be explained by body weight.

e. Correlation coefficient $R = \sqrt{0.6466} = 0.8041144$.

7.40 Rate my professor.

a. Considering that $\bar{x} = -0.0883$ and $\bar{y} = 3.9983$ and that point $(\bar{x}, \bar{y})$ is located on the regression line, we have $3.9983 = 4.010 + b_1 * (-0.0883)$, so the slope $b_1 = \frac{3.9983-4.01}{-0.0883} = 0.1325028$.

b. Since the slope is positive the relationship is positive. If we set up a hypothesis test with $H_0 : \beta_1 = 0$ and $H_A : \beta_1 > 0$, then based on the summary table the $p-value$ is nearly 0. And this is for a two-sided test, so it'll be even closer to 0 for a one-sided test. We reject the null hypothesis. There is convincing evidence that the relationship between teaching evluation and beauty is positive.

c. **Linearity:** Based on the scatterplot, there may be a weak linear relationship. There is no evident pattern in the residual plot. **Nearly normal residuals:** The histogram of the residuals exhibits a left skew. Additionally, the points seem to move away from the normal probability line on each end. However, the bulk of the data is very close to the line. I would conclude that the distribution of residuals is nearly normal. **Constant variability:** Based on residual plot, there appears to be constant variability in the data. **Independent observations:** Observations are not a time series, and can be assumed to be independent (unless there is evidence that students copied each other's evaluations). *I believe all conditions are satisfied for this linear model.*