Homework\_7

Priya Shaji

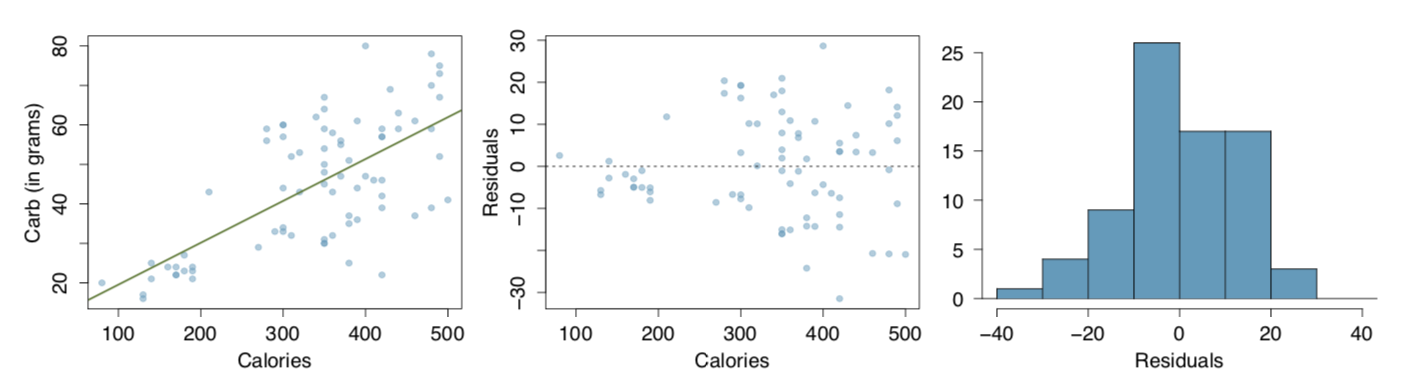
4/3/2019

## Introduction to Linear Regression

### 7.24

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.21 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.



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

Answer a)

There is positive linear relationship between calorie and carbohydrates.

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

Answer b)

x - axis is Calorie which is explanatory variable

y-axis is Carbohydrate which is response variable.

1. Why might we want to fit a regression line to these data?

Answer c)

We would like to predict the amount of carbs based on calorie count.

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

Answer d)

Linearity: The data should show a linear trend. The scatterplot suggests that the data show a moderate to weak linear relationship.

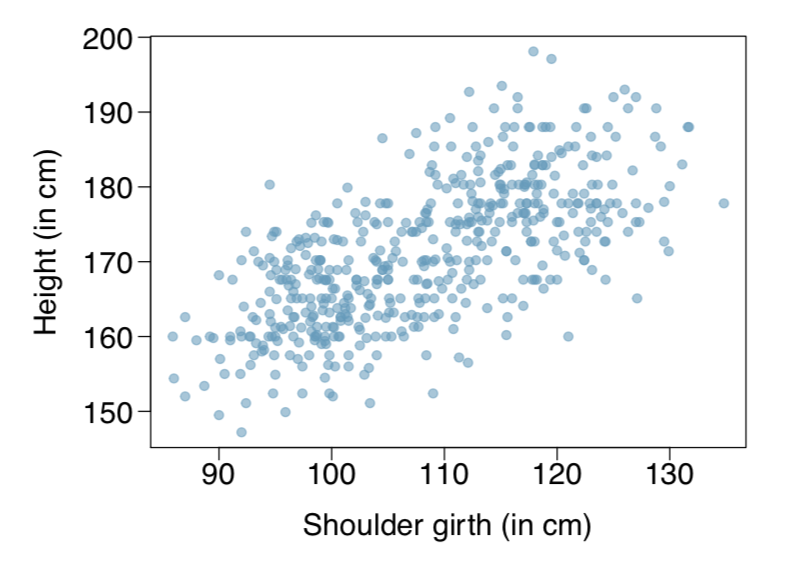
Nearly normal residual: From the histogram for residual, the residuals in this case have a slightly left skewed.

Constant variability: The residual plot suggests that there is no constant variability for residuals. The data fit the linear model much better for lower number of calories than for higher as shown by much larger residuals value.

Independent observation: Its assumed that each menu item is independent of the other, but they are all Starbucks menu items.

### 7.26

Body measurements, Part III. Exercise 7.15 introduces data on shoulder girth and height of a group of individuals. The mean shoulder girth is 107.20 cm with a standard deviation of 10.37 cm. The mean height is 171.14 cm with a standard deviation of 9.41 cm. The correlation between height and shoulder girth is 0.67.



1. Write the equation of the regression line for predicting height.

Answer a)

# The general equation for the regression line: y = B0 + B1\*x where B0 and B1 represent two model parameters.  
# X is the explanatory or predictor variable and y is the response.  
# B1 is the slope, which also equals: (sample y standard deviation) / (sample x standard deviation) \* R.  
# R is the correlation between the two variables.  
# R ranges from -1 to 1, with -1 being completely negative correlation and +1 being a completely positive correlation.  
  
# Shoulder girth is the explanatory variable  
shoulder.mean <- 107.20 # in cm  
shoulder.SD <- 10.37 # in cm  
  
# Height is the response  
height.mean <- 171.14 # in cn  
height.SD <- 9.41 # in cm  
  
# R for correlation  
R <- 0.67  
  
# Calculate the slope (or otherwise known as B1)  
B1 <- R \* (height.SD/shoulder.SD)  
  
# Now to calculate for B0, we will use the values (x,y) = (107.20, 171.14). They are also the mean values, and they lie along the regression line.  
# Now to rearrange the equation to solve for B0. B0 = y - B1\*x  
B0 <- 171.14 - B1 \* 107.20  
  
B1;B0

## [1] 0.6079749

## [1] 105.9651

Therefore, the general regression line for predicting height is:

ŷ =105.8445+0.6091∗‘shouldergirth‘

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

Answer b)

The intercept (105.9651) is the height in centimeters at girth of 0 cm. The slope (0.6079749) is the number of centimeters increase in height for each increase in shoulder girth. Not all values make sense when we plug them into this linear regresion equation. For example, a person with a shoulder width of 0 cm (which can’t happen) would indicate a height of 105.965 cm. This does not make sense.

1. Calculate R2 of the regression line for predicting height from shoulder girth, and interpret it in the context of the application.

Answer c)

R.squared <- R^2  
paste("R squared: ", round(R.squared,3))

## [1] "R squared: 0.449"

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

Answer d)

student.shoulder <- 100 # in cm  
student.height <- B0 + B1 \* student.shoulder  
paste("According to the model, the estimated height of a student with a shoulder girth of 100 cm is: ", round(student.height,3), "cm.")

## [1] "According to the model, the estimated height of a student with a shoulder girth of 100 cm is: 166.763 cm."

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

Answer e)

# The residual = time(actual) - time(expected)  
residual <- 160 - student.height  
paste("The residual is: ", round(residual,3))

## [1] "The residual is: -6.763"

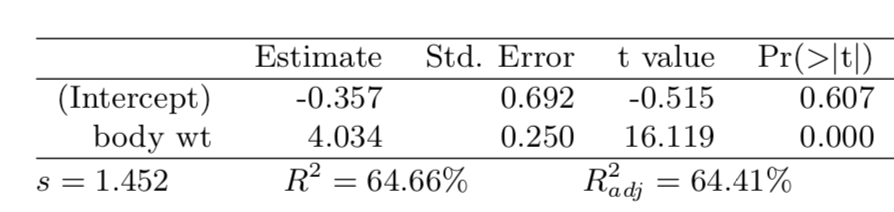
1. A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

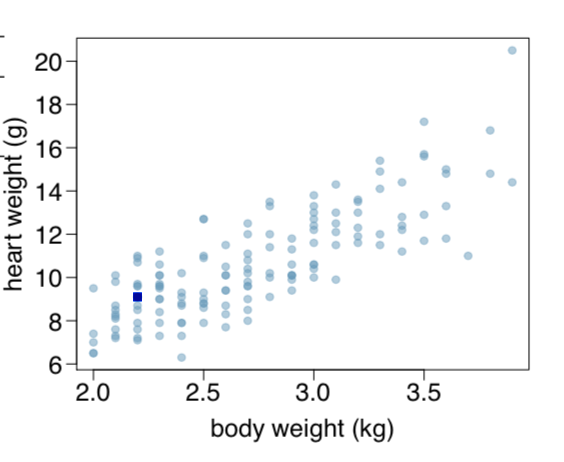
Answer f)

As per excercise 7.15, The original data set had a response variable values between ~80 and 140 cm. A measure of 56 is outside the sample and we would require extrapolation and would not be appropriate. There may be many confounding factors such as malnourishment, developmental delays, congenital diseases, etc. that may not be accounted for in this current linear regression model.

### 7.30

Cats, Part I. The following regression output is for predicting the heart weight (in g) of cats from their body weight (in kg). The coefficients are estimated using a dataset of 144 domestic cats.





1. Write out the linear model.

Answer a)

As stated in Question 7.26 for the linear model equation y = B0 + B1 \* x, where B0 is the height of the intercept and B1 is the slope. Therefore, by looking at the table, the linear regression model here is:

Heart Weight (g) = -0.357 + 4.034 \* `Body Weight (kg)

1. Interpret the intercept.

Answer b)

Heart Weight (g) = -0.357 + 4.034 \* `Body Weight (kg)

Expected heart weight in cats with 0 kg body weight is -0.357 g. This is not a meaningful value, it just serves to adjust the height of the regression line.

1. Interpret the slope.

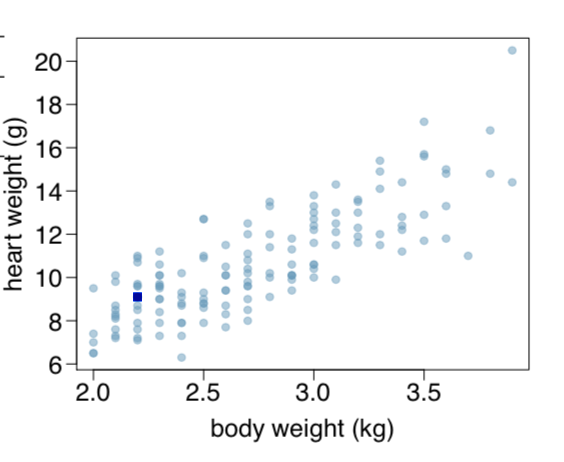
Answer c)

Heart Weight (g) = -0.357 + 4.034 \* `Body Weight (kg)

For each additional kg increase in body weight, we expect an additional 4.034 grams in the heart weight.

1. Interpret R2.

Answer d)



Body weight (in kg) explains 64.66% of the variability in the heart weight (in g) of the cat.

1. Calculate the correlation coefficient.

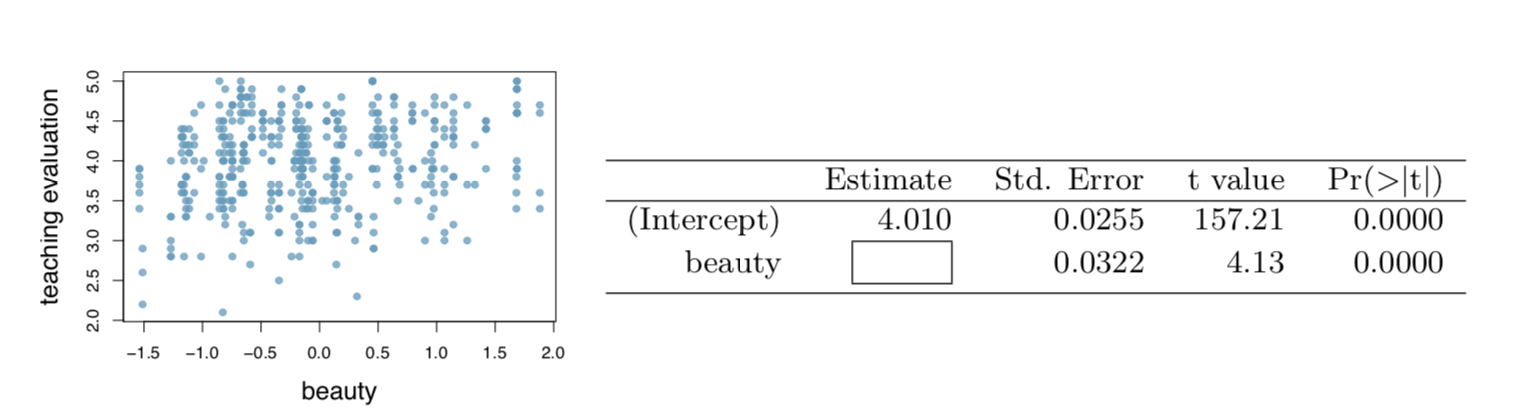
Answer e)

paste("Correlation coefficient: ", round(sqrt(.6466),3))

## [1] "Correlation coefficient: 0.804"

### 7.40

Rate my professor. Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures may reflect the influence of non-teaching related characteristics, such as the physical ap- pearance of the instructor. Researchers at University of Texas, Austin collected data on teaching evaluation score (higher score means better) and standardized beauty score (a score of 0 means average, negative score means below average, and a positive score means above average) for a sample of 463 professors.24 The scatterplot below shows the relationship between these variables, and also provided is a regression output for predicting teaching evaluation score from beauty score.



1. 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.

Answer a)

Slope can be calculated from the linear model equation y = B0 + B1 x Here, they provide enough information so that we can solve B1 (the slope) We can substitute x with the average standardized beauty score (explanatory variable) = -0.0883 and we can substitute y with the average teaching evaluation score of 3.9983 (response). We know that the averages lie on the linear regression model.We also know the Y intercept. It is on the table. Y intercept = 4.010 Therefore, the equation is now: 3.9983 = 4.010 + B1 \* (-0.0883) Now solve for B1 to find the slope.

prof.B1 <- (3.9983 - 4.010)/(-0.0883)  
paste("B1 or the slope: ", round(prof.B1, 3))

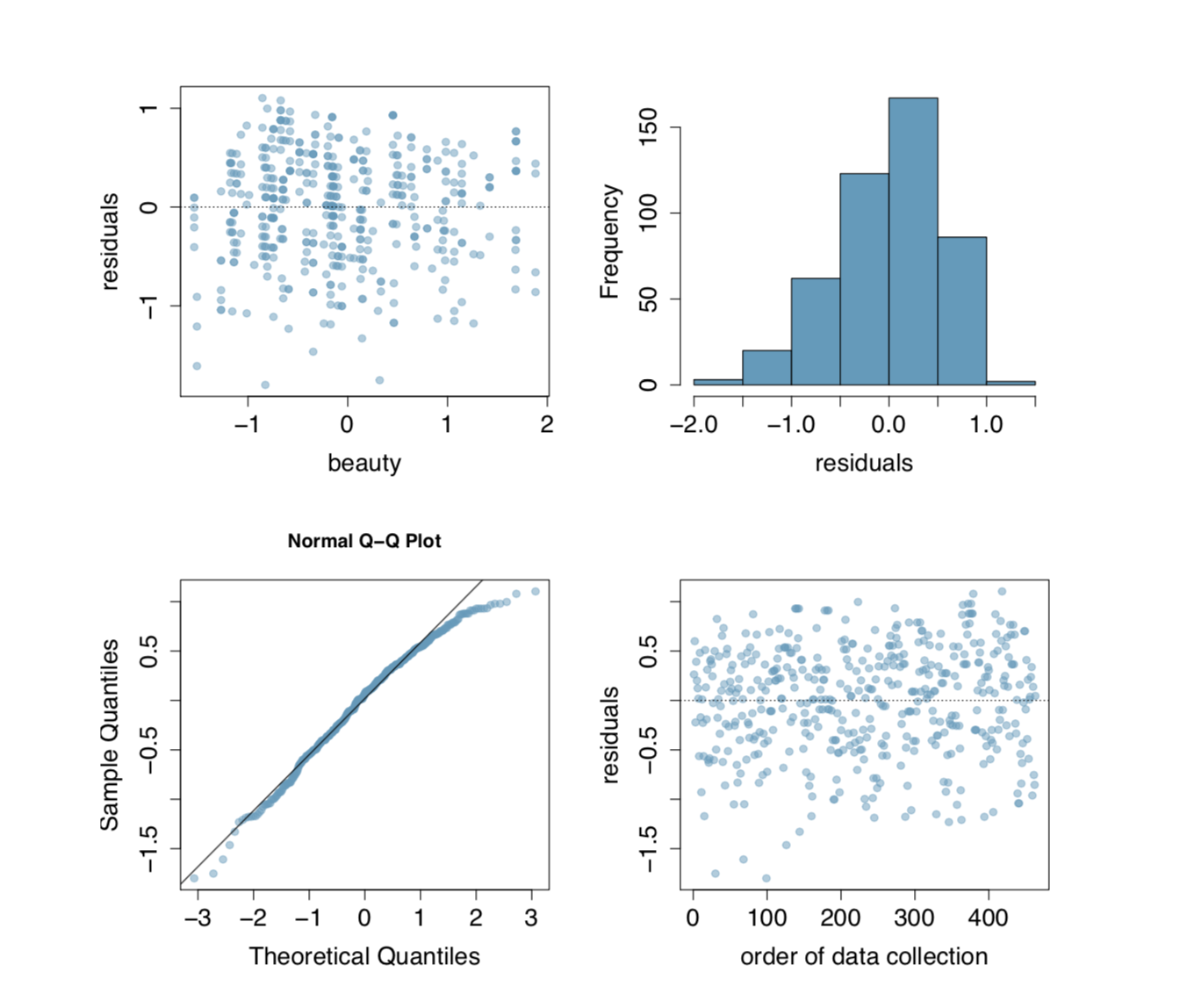
## [1] "B1 or the slope: 0.133"

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

Answer b)

While on visual inspection, if anything, it looks like there may be a near zero slope level. However, the data from the summary chart does suggest that there is some correlation between beauty and teaching evaluation. The P values in the summary chart demonstrates p’s that are ~0, which suggests that we should reject the null hypothesis (null hypothesis being that there is NO correlation).

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



Answer c)

There are four criteria that needs to be satisfied prior to using the least squares linear regresion method.

Linearity : There does appear to be some linearity if we take a horizontal slope and place a horizontal line roughly right through the data points

Nearly normal residuals : The residuals do appear to be nearly normal, by looking at its distribution on the histogram.

Constant variability : By looking at the residual plots, there appears to be constant variability throughout

Independent observations : Looking at the order of data collection, there doesn’t seem to an obvious pattern that would suggest that these observations were dependent on each other. It is likely that these were all independent observations.

Hence, we can perform the linear regression.