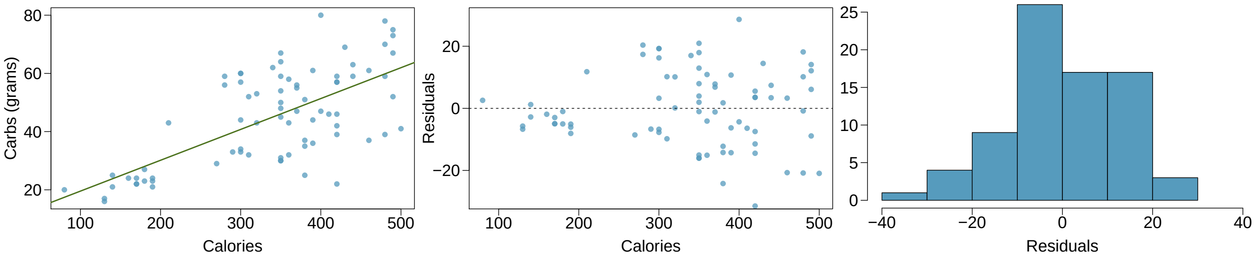
DATA606_HW8_RJM

RJM

2019-11-11

Nutrition at Starbucks, Part I. (8.22, p. 326) The scatterplot below shows the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain. 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.



8.22 (a)

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

The relationship between calories and carbs seems to be positive overall but there is some variability especially at the higher calories level where the data points exist on both sides of the regression line.

8.22 (b)

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

The explanatory variables are calories along the x-axis and carbs are the response variable on the y-axis.

8.22 (c)

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

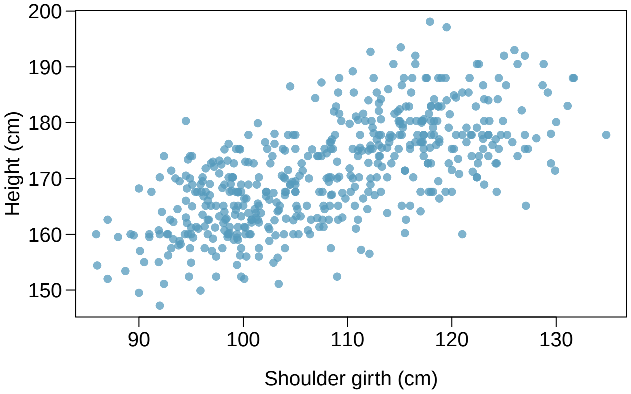
A regression line helps us understand the relationship between the two variables and gives an estimate of the amount of carbs based on the number of calories.

8.22 (d)

(d) Do these data meet the conditions required for fitting a least squares line? The conditions for fitting a least squares line are as follows: Linearity: The linear relationship is not particularly strong as the variability increases with the increase in the number of calories. Normality: The distribution is quite normal as could be seen in the residuals plot and bar chart, but there is a small skew towards left.

Constant variability: The variability seems to increase with an increase in the amount of calories, so it is not constant. Independence: The independence seems to be there as the number of calories are dependent on the type of food on the menu. Conclusion: It seems that the least squares test could be conducted but the outcome might not be as accurate.

Body measurements, Part I. (8.13, p. 316) Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender for 507 physically active individuals.19 The scatterplot below shows the relationship between height and shoulder girth (over deltoid muscles), both measured in centimeters.



8.13 (a)

(a) Describe the relationship between shoulder girth and height. The relationship between should girth and height is mostly positive but there is some variability as the shoulder girth increases.

8.13 (b)

(b) How would the relationship change if shoulder girth was measured in inches while the units of height remained in centimeters? There will be a scaling issue as inches are nearly 2.5 times bigger than centimeters. The relationship will remain the same but visualization will show steeper change in height with every change in the shoulder girth.

Body measurements, Part III. (8.24, p. 326) Exercise above 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.

8.24 (a)

(a) Write the equation of the regression line for predicting height.

```
mean_shoulder <- 107.20
sd_shoulder <- 10.37
mean_height <- 171.14
sd_height <- 9.41
r_sh_h <- 0.67

b_1 <- r_sh_h * (sd_height / sd_shoulder)
b_1
```

```
## [1] 0.6079749
```

```
b_0 <- mean_height - b_1 * mean_shoulder
b_0
```

```
## [1] 105.9651
```

```
# The equation for the regression line is:
# height = b_0 + b_1 * x
# height = 105.97 + 0.608x
```

8.24 (b)

(b) Interpret the slope and the intercept in this context. For every 1 cm increase in the shoulder girth there is an increase of 0.608 in height.

8.24 (c)

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

```
r_squared <- r_sh_h^2
r_squared
```

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

```
# R-squared is 0.449.
# The above r-squared value could be interpreted to explain the 44.9% of variation in the linear
# model.
```

8.24 (d)

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

```
x_shoulder <- 100

height_student <- b_0 + b_1*x_shoulder
height_student
```

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

```
# The height of the student is 166.76 cm.
```

8.24 (e)

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

```
real_height <-  160

res_height <- real_height - height_student
res_height
```

```
## [1] -6.762581
```

```
# The residual is -6.76 which means that the model is over estimating the height of the student.
```

8.24 (f)

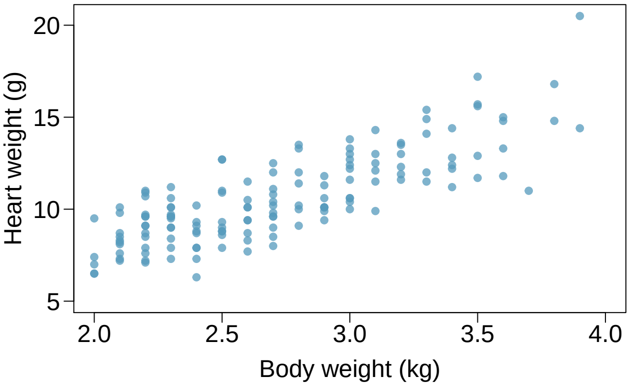
(f) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

We can observe from the graph that all the data points are above 80 cm for the shoulder girth. Therefore, 56 cm would not be within the range and we will not be able to use this model to predict the height of a child based on that size.

Cats, Part I. (8.26, p. 327) 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.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.357	0.692	-0.515	0.607
body wt	4.034	0.250	16.119	0.000

$s = 1.452$ $R^2 = 64.66\%$ $R^2_{adj} = 64.41\%$



8.26 (a)

(a) Write out the linear model.

Based on the information in the table provided, the linear model is as follows:

heart_weight = -0.357 + 4.034 * body_weight

8.26 (b)

(b) Interpret the intercept.

If the body weight is 0 then the heart weight will be -0.357gms. This is not a realistic outcome but the linear model sets this condition. However, there is no chance of getting body weight to 0.

8.26 (c)

(c) Interpret the slope.

For every 1 kg increase in the body weight, corresponding heart weight will increase by 4.034 gms.

8.26 (d)

(d) Interpret R^2.

The change in body weight explains the 64.66% of the variation in the heart weight.

8.26 (e)

(e) Calculate the correlation coefficient.

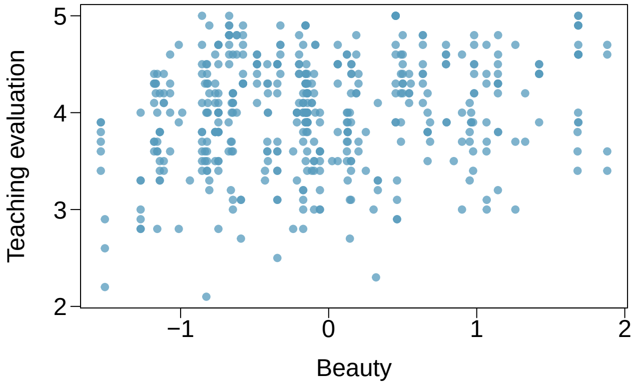
```
# The correlation coefficient or R is calculated as follows:
r_sq_weight <- .6466
r_weight <- round(sqrt(r_sq_weight),3)
r_weight
```

```
## [1] 0.804
```

```
# The coefficient of correlation is 80.4% which means that the data point on body weight explain
# 80.4% of variance in the heart weight.
```

Rate my professor. (8.44, p. 340) 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 appearance 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. The scatterplot below shows the relationship between these variables, and also provided is a regression output for predicting teaching evaluation score from beauty score.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.010	0.0255	157.21	0.0000
beauty		0.0322	4.13	0.0000



8.44 (a)

(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.

```
# From the given information:
# Avg teaching evaluation score or Y = 3.9983
Y_teach <- 3.9983
# b_0 = 4.010
b_0_t <- 4.010

# x = -0.0883
x_eval <- -0.0883

# slope or b_1 = ?

# Writing the linear model equation using the above info:

# Y_teach = b_0_t + b_1_t * x_eval
# From the above equation:

b_1_t <- (Y_teach - b_0_t) / x_eval
b_1_t
```

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

```
# The slope is 0.133.
```

8.44 (b)

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

While the plot appears to suggests that there is a very slight positive relationship, the equation confirms that the relationshop between the appearance of instructors and the evaluation has a small positive association.

8.44 (c)

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

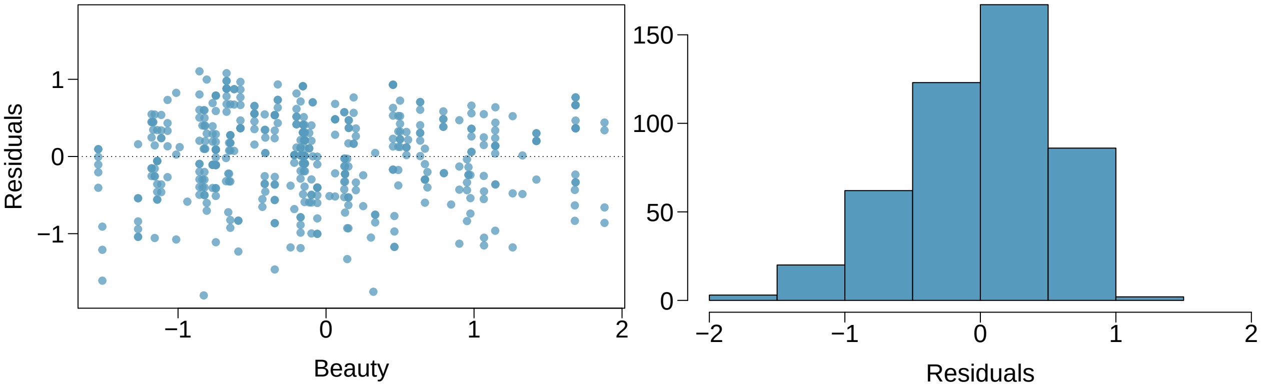
The conditions reuired for the linear regression are as follows:

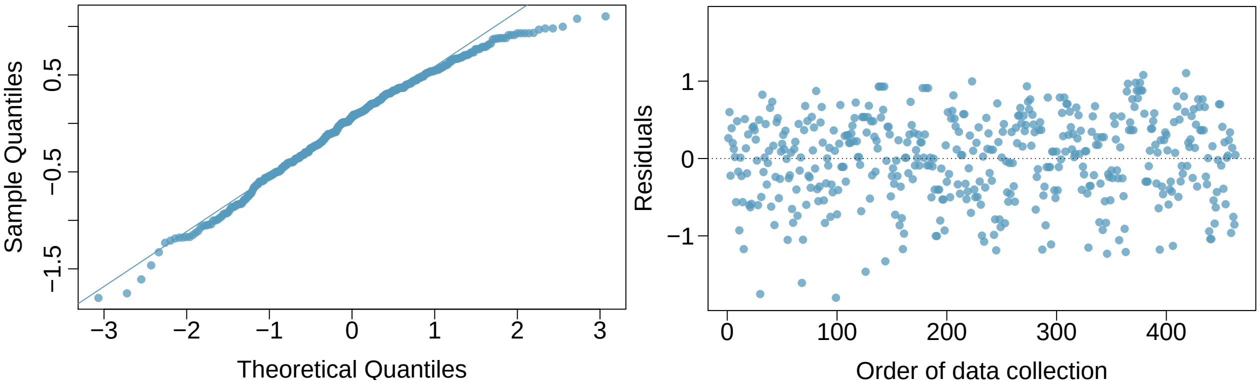
Linearity: It is possible to have a regression line run through the data points based on the equation of linear regression, so this condition is met.

Normality: The distribution of residuals seems to be normal. The histogram appears to be skightly skewed to the left but the distribution seems quite normal.

Constant variability: The variability seems to constant from the plot with more variance at the end points of data.

Independence: The observations seem to be independent of each other and the sample size is big enough to have the required indpendence.




Sample Quantiles
0.5
−0.5
−1.5
Theoretical Quantiles
−3
−2
−1
0
1
2
3
Residuals
1
0
−1
Order of data collection
0
100
200
300
400