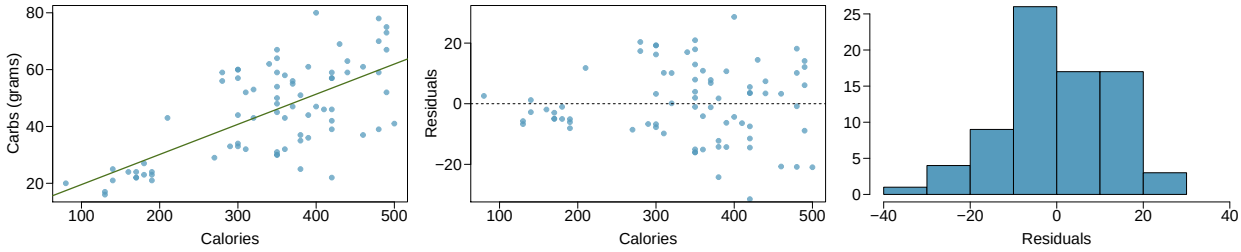
DS606-HW8

George Cruz

11/22/2020

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.



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

There appears to be a linear relationship between calories and carbohydrates.

(b) In this scenario, what are the explanatory and response variables? Calories is the explanatory variable and amount of carbohydrates the response variable.

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

We might want to predict the amount of carbohydrate (in grams) by looking at the number of calories in the item.

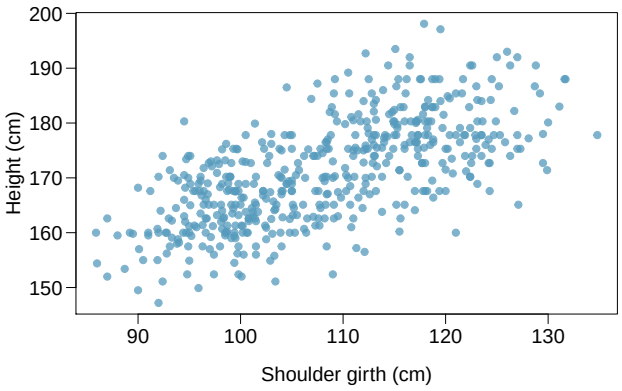
(d) Do these data meet the conditions required for fitting a least squares line?

When fitting a least squares line, we generally require

Linearity. The data appears to show a linear trend. **Nearly normal residuals.** The residuals appear to be somewhat normal. **Constant variability.** This appear to not be true. The variability of the data increases with larger values of x.

No The data does not meet the Constant Variability requirement.

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. The scatterplot below shows the relationship between height and shoulder girth (over deltoid muscles), both measured in centimeters.



(a) Describe the relationship between shoulder girth and height. There is a positive relationship between shoulder girth and height.

(b) How would the relationship change if shoulder girth was measured in inches while the units of height remained in centimeters?

```
par(mar = c(3.8, 3.8, 0.5, 0.5), las = 1, mgp = c(2.7, 0.7, 0),
    cex.lab = 1.25, cex.axis = 1.25)

library(tidyverse)
```

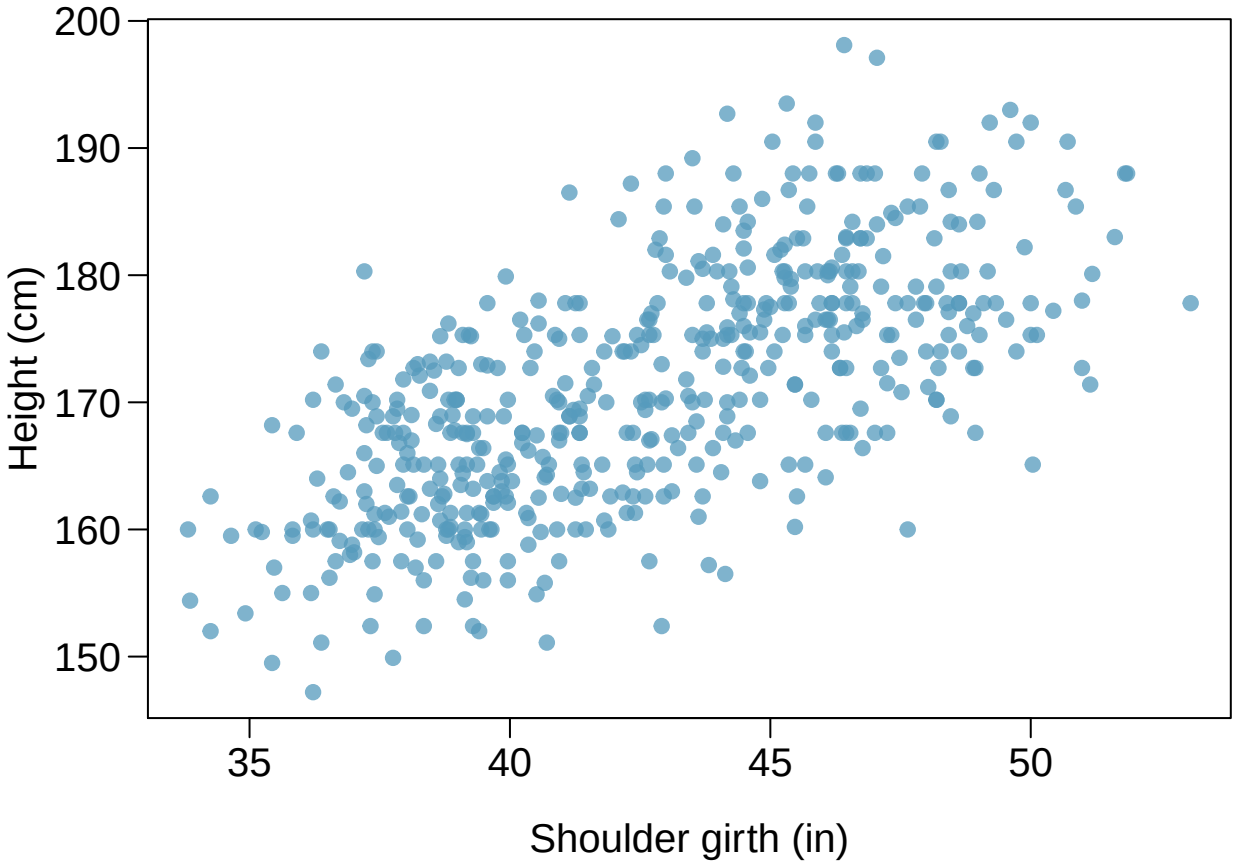
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.0
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
bdims2 <- bdims %>%
  rowwise() %>%
  mutate(sho_inch = (sho_gi * 0.393701))

plot(bdims2$hgt ~ bdims2$sho_inch,
     xlab = "Shoulder girth (in)", ylab = "Height (cm)",
     pch = 19, col = COL[1,2])
```



The positive relationship remains but the length of the graph gets shortened and we might get a steeper slope.

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.

(a) Write the equation of the regression line for predicting height. y = B0 + B1 * x

B1 = Sy /Sx * R

```
mean_x <- 107.2
sd_x <- 10.37
mean_y <- 171.14
sd_y <- 9.41
r <- 0.67

b1 <- ( sd_y / sd_x ) * r
b0 <- mean_y - (b1*mean_x)

c(b0, b1)
```

```
## [1] 105.9650878   0.6079749
```

y = 105.966 + 0.608 * x

(b) Interpret the slope and the intercept in this context. The slope means that for every cm increase in shoulder girth, we can expect an increase of 0.608cms in height.
The intercept means that for a shoulder girth of 0, we will have a height of 105.966cm. This is not possible, but allows to adjust for the prediction.

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

```
r ** 2
```

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

R^2 is 0.449 this means the regression line accounts for roughly 45% of the variance height predicted from shoulder girth.

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

```
rdm_sd_ht <- 105.966 + 0.608 * 100
rdm_sd_ht
```

```
## [1] 166.766
```

We would predict this random student to be 166.766 cm tall.

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

```
rdm_sd_res <- 160 - rdm_sd_ht
rdm_sd_res
```

```
## [1] -6.766
```

The residual is -6.766. Because it is negative, it means we overestimated the height.

(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?

```
105.966 + 0.608 * 56
```

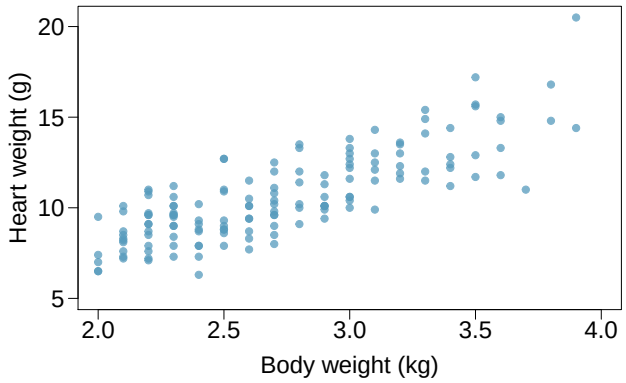
```
## [1] 140.014
```

Seeing as it would predict a height of 140cm, it does not seem an appropriate model. 56cm falls almost 5 standard deviations from the mean and it's outside the observation range.

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

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



(a) Write out the linear model. y = -0.357 + 4.034 * x

(b) Interpret the intercept. The intercept is -0.357. This means that when the weight of the cat is 0, the weight of the heart will be -0.357. This will not be observable in real life as the cat will never weight 0.

(c) Interpret the slope. The slope is 4.034, this means that the heart weight will increase 4.034g for every kg increase in cat's weight.

(d) Interpret R^2. The R^2 value is 0.6466, this means that 64.66% of the cat heart's weight is explained by the cat's weight.

(e) Calculate the correlation coefficient.

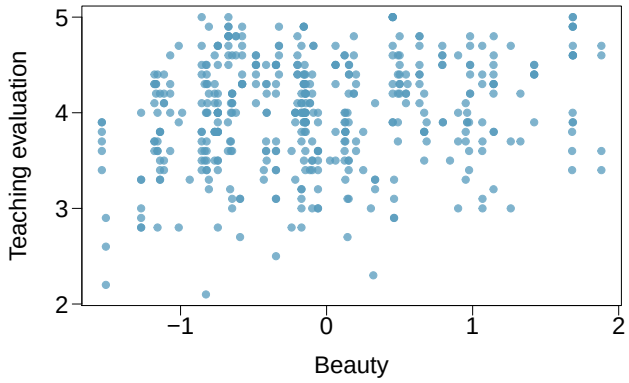
```
r <- 0.6466 ** 0.5
r
```

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

The correlation coefficient is **0.8041**.

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



(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.

```
b0 <- 4.010
x <- -0.0883
y <- 3.9983

b1 <- (y - b0) /x
b1
```

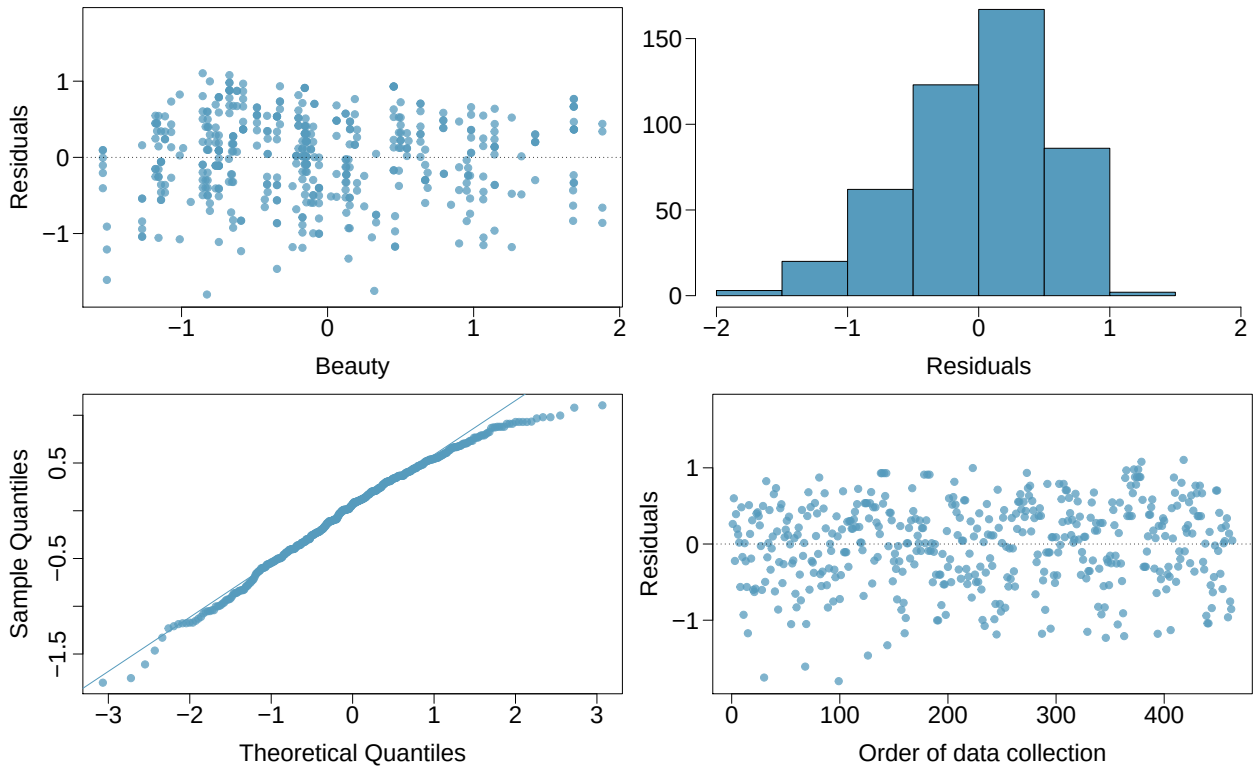
```
## [1] 0.1325028
```

The slope is 0.1325

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

Yes, the slope is positive.

(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: No, It's not certain that we are seeing a linear trend.

Nearly normal residuals: Yes, The residuals appear nearly normal.

Constant variability: Yes, The scatterplot of the residuals does appear to have constant variability.

Independent observations: Yes, The variables are independent of each other.

```
summary(lm(formula=courseevaluation~btystdave, data=prof_evals_beauty))
```

```
##
## Call:
## lm(formula = courseevaluation ~ btystdave, data = prof_evals_beauty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.80015 -0.36304  0.07254  0.40207  1.10373
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.01002    0.02551 157.205  < 2e-16 ***
## btystdave    0.13300    0.03218   4.133 4.25e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5455 on 461 degrees of freedom
## Multiple R-squared:  0.03574,    Adjusted R-squared:  0.03364
## F-statistic: 17.08 on 1 and 461 DF,  p-value: 4.247e-05
```

Based on the Multiple R^2 being 0.03 we must assume that there is not a linear relation between the two variables.