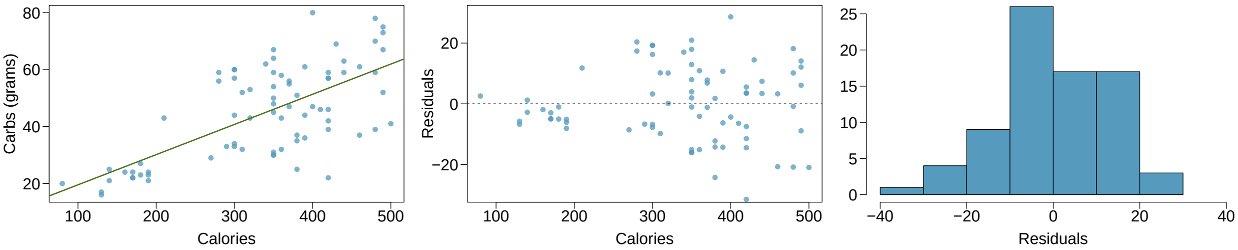
CUNY SPS DATA606 HW8- Introduction to Linear Regression

Chinedu Onyeka

October 21st, 2021

Problem 1:

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.
(b) In this scenario, what are the explanatory and response variables?
(c) Why might we want to fit a regression line to these data?
(d) Do these data meet the conditions required for fitting a least squares line?

Solution 1: 1a) There is a strong positive linear relationship between number of calories and amount of carbohydrates(grams) that Starbucks food menu items contain.

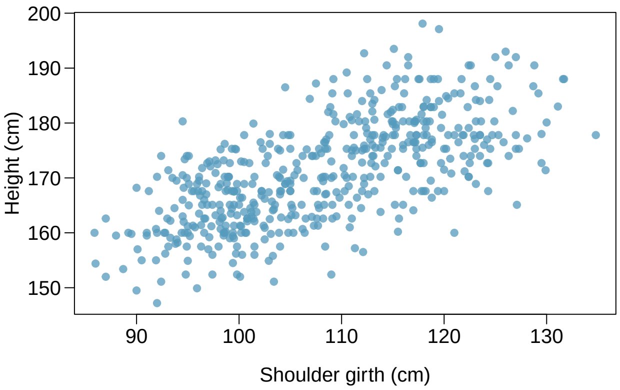
1b) The explanatory variable is "Calories" while the response variable is "Carbs(grams)".

1c) We might want to fit a regression line to these data so that we can use the regression line to predict the amount of Carbs(grams) for a given Starbucks food menu that contains a certain number of Calories.

1d) Yes. The data shows a linear trend, the observations are independent, there are no outliers that tend to be too far from the line, and there seems to be constant variability of the points about the least square line. Therefore, the data meet the conditions required for fitting a least square line.

Problem 2:

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.

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

Solution 2: 2a) The relationship between shoulder girth and height is a linear relationship with a strong positive correlation.

2b) The relationship will not change. However, the slope will be different. The slope in this case will be steeper (higher slope).

Problem 3:

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.
(b) Interpret the slope and the intercept in this context.
(c) Calculate R^2 of the regression line for predicting height from shoulder girth, and interpret it in the context of the application.
(d) A randomly selected student from your class has a shoulder girth of 100 cm. Predict the height of this student using the model.
(e) The student from part (d) is 160 cm tall. Calculate the residual, and explain what this residual means.
(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?

Solution 3: 3a)

height = slope*(shoulder girth) + intercept

```
mean_sg <- 107.20 # Mean shoulder girth in cm
sd_sg <- 10.37 # standard deviation of shoulder girth
mean_h <- 171.14 # Mean height in cm
sd_h <- 9.41 # standard deviation of height
R <- 0.67 # Correlation between height and shoulder girth.
slope <- round(R * (sd_h/sd_sg), 5)
# point slope equation of a line: y - y1 = m(x - x1); y = mx + y1 - mx1; intercept = y1 - mx1
# => intercept = mean_height - slope*(mean_sg)
intercept <- round((mean_h - slope*mean_sg), 2)
# equation of line: height = slope*shoulder_girth + intercept
paste0("The equation for predicting the height is: ",
       "height = ", as.character(slope)," * shoulder_girth +  ", as.character(intercept))
```

```
## [1] "The equation for predicting the height is: height = 0.60797 * shoulder_girth +  105.97"
```

3b)

The intercept is the minimum height possible for an individual with no shoulder girth while the slope is the amount that the height of an individual will increase by for every 1 cm increase in shoulder girth.

3c)

```
R_squared <- R**2
paste0("The R square is ", R_squared, ". That is, the regression line explains about ",
       R_squared*100, "% of variabilities in height")
```

```
## [1] "The R square is 0.4489. That is, the regression line explains about 44.89% of variabilities in
```

3d)

```
shoulder_girth <- 100
height <- round((slope*shoulder_girth + intercept), 2)
paste0("The height of a randomly selected student of shoulder girth ", shoulder_girth, "cm is: ",
       height, "cm")
```

```
## [1] "The height of a randomly selected student of shoulder girth 100cm is: 166.77cm"
```

3e)

```
height_pred <- height
height_actual <- 160
residual <- round((height_actual - height_pred),2)
paste0("The residual for this student is ", residual, "cm. This means that the error in predicting",
       " the height of this student is ", residual,
       "cm. The height of this student has been overestimated by the model")
```

```
## [1] "The residual for this student is -6.77cm. This means that the error in predicting the height of
```

3f)

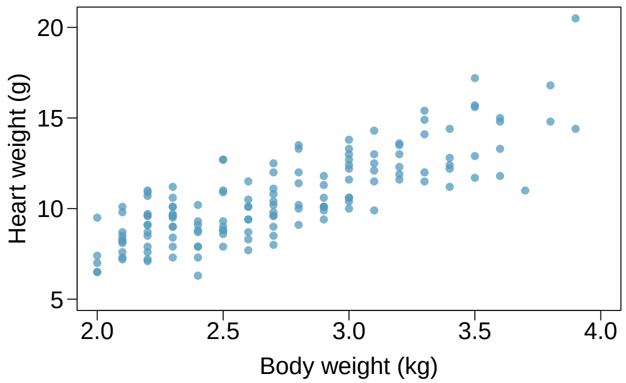
No. The scatter-plot of height vs shoulder girth shows that the model was trained on shoulder height starting from around mid 80 - 90 cm. This means that a should girth of 56cm is below the minimum height height used to train this model. Hence, it would not be appropriate to use this model to predict the height of this child.

Problem 4:

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.
(b) Interpret the intercept.
(c) Interpret the slope.
(d) Interpret R^2.
(e) Calculate the correlation coefficient.

Solution 4: 4a)

heart_weight = intercept + slope*body_weight

```r
intercept_cat <- -0.357
slope_cat <- 4.034
paste0("The linear model is: ",
       "heart_weight = ",as.character(intercept_cat), " + ",
       as.character(slope_cat)," * body_weight")
```

```
## [1] "The linear model is: heart_weight = -0.357 + 4.034 * body_weight"
```

4b)

The intercept is the minimum possible heart weight for a cat. i.e. If a cat is to have a body weight of zero kg, that cat would a heart weight equal to the intercept. In reality, it would not be possible for a cat to weigh zero kg, but that is what the intercept means. Essentially, the intercept is the base heart weight of a cat.

4c)

The slope is the amount by which the heart weight of a cat will increase by if the body weight of the cat were to increase by 1kg.

4d)

The R squared is percentage of variability in heart weight explained by the model. In this case, the R squared is 64.66% which means that about 64.66% of variations in cat heart weight is explained by the linear model.

4e)

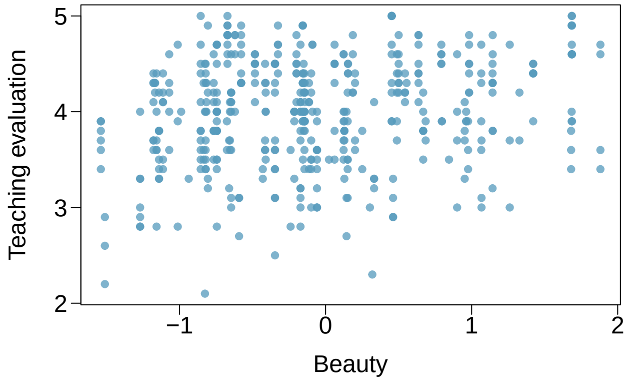
```r
# correlation coefficient
R_squared_cat <- 0.6466
R_cat <- round(sqrt(R_squared_cat), 3) # correlation coefficient
paste0("The correlation coefficient is ", R_cat)
```

```
## [1] "The correlation coefficient is 0.804"
```

Problem 5:

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.

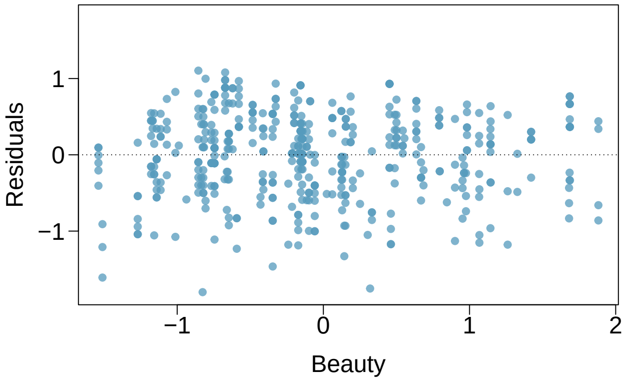
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.010	0.0255	157.21	0.0000
beauty	□	0.0322	4.13	0.0000

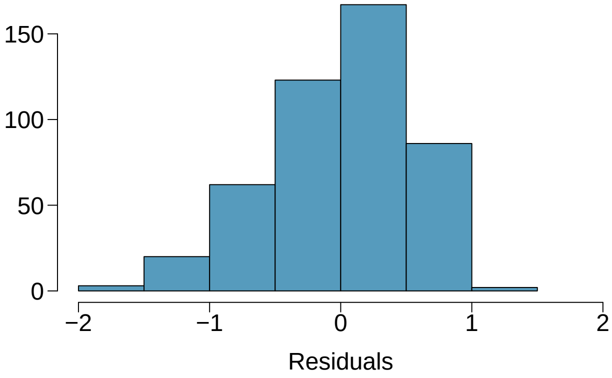


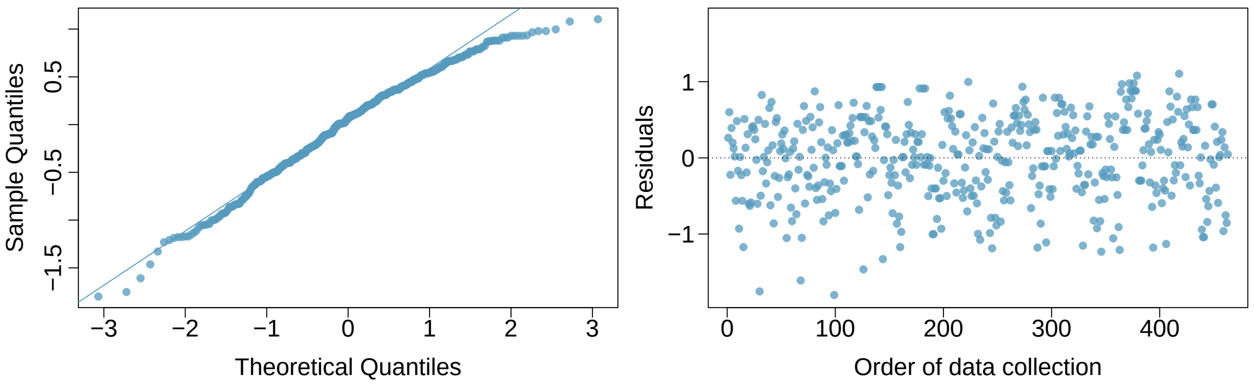
(a) Given that the average standardized beauty score is -0.0883 and average teaching evaluation score is 3.9983, calculate the slope. Alternatively, the slope may be computed using just the information provided in the model summary table.

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

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







Solution 5: 5a)

Teaching_eval = intercept_teaching + slope_teaching*Beauty_score
=> slope_teaching = (Teaching_eval - intercept_teaching)/Beauty_score

```
Beauty_score <- -0.0883
intercept_teaching <- 4.010
Teaching_eval <- 3.9983
slope_teaching <- (Teaching_eval - intercept_teaching)/Beauty_score
paste0("The slope of the regression model is ", round(slope_teaching,4))
```

```
## [1] "The slope of the regression model is 0.1325"
```

5b)

Yes. It may not be clearly visible from the scatter plot. However, when calculated, the slope is 0.1325 which is greater than zero and this implies a *weak positive* relationship between teaching evaluation and beauty. Also, from the calculated t value which is 4.13, we can see that the p value is 0.0000 and this is statistically significant. Hence, we can say that the data provides convincing evidence of a positive relationship.

5c)

Conditions for linear regression:
1. Linearity: The data should show a linear trend. The scatter plot of the Teaching evaluation vs Beauty shows a slightly linear trend.
2. Constant variability: The variability of points around the least square line remains nearly constant. There appears to be constant variability in the residuals and the points show almost similar variability from the zero line.
3. Independent observations: Since data was collected from 463 different professors, there is no clear indication that the teaching evaluation score of one professor is dependent on others. Therefore, I can assume that the observations are independent.
4. Nearly normal residuals: There is no outliers that appears to be a potential concern. The histogram of residuals shows a nearly normal distribution of the residuals.Also, the qq plot also implies a nearly normal distribution.

The conditions for linear regression is satisfied for this dataset based on the diagnostic plot.