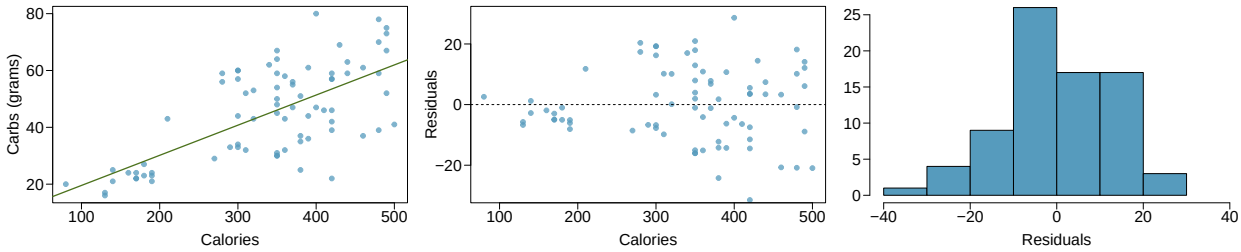
Chapter 8 - Introduction to Linear Regression

Title: CUNY SPS MDS DATA607_HW8"

Author: Charles Ugiagbe

Date: "10/27/2021"

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:(Nutrition at Starbucks)

1a)

There is a positive linear relationship between the calories content and the carbs

1b)

Explanatory Variable is the number of Calories content

Response Variable is the "amount of Carbs"

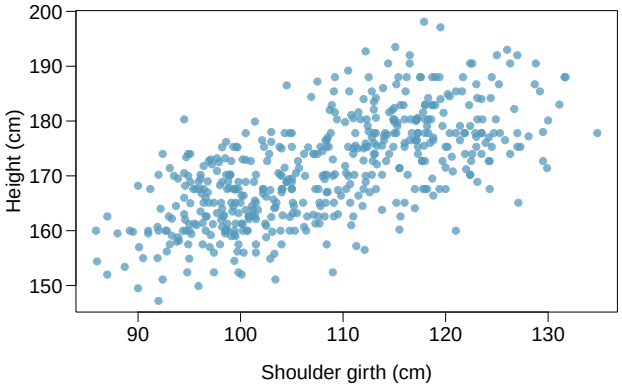
1c)

by fitting a regression line to these data, we can predict the amount of carbs a menu item has based on its calorie content.

1d)

No, it partly met the linearity and normality condition, but fail the constant variability condition. Base on the second residual plot, the absolute value of residual increase as value of calories increase, which violates the constant variability condition.

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: (Body measurements, Part I)

2a)

There is a positive linear relationship between Shoulder girth and height. As shoulder girth increase, the height also increases.

2b)

Converting inches to centimeters will not change the relationship between shoulder girth and height. Only the dimentsion will change. the plot will remain positively linear.

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:(Body measurements, Part III)

3a) height = 105.965 + 0.607975 * shoulder_girth

```
R <- 0.67
ybar <- 171.14
xbar <- 107.20
sy <- 9.41
sx <- 10.37

b1 <- R * sy / sx
b0 <- ybar - xbar*b1

b1
```

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

```
b0
```

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

3b)

The `slope` is 0.60797, means that for one 1cm increase / decrease on shoulder girth, the height is estimated to increase / decrease by 0.60797cm.

The `intercept` 105.97 describes the average outcome of Height if shoulder girth = 0

3c)

```
R^2
```

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

R^2 is 0.4489. This mean proportion of the variability in height that is explained by the shoulder girth is 0.4489

3d)

Height is 166.77

```r
shoulder_girth = 100
height = 105.965 + 0.607965 * shoulder_girth
height
```

```
## [1] 166.7615
```

3e)

$e_i = y_i - \hat{y}_i$

$e_i = 160 - 166.762$

$e_i = 6.762$

Residual between actual and estimate is -6.762cm. The negative residual of -6.762 mean the model overestimates the height.

3f)

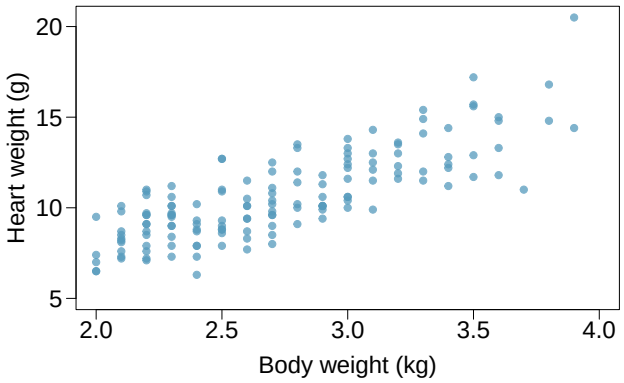
No, the observed value is too far away from the scope of current Model. The linear model is not effective on predict the height on such observed values. The calculation will require extrapolation

---

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: (Cats, Part I)

4a) linear model given as:

heart_weight = -0.357 + 4.034 * body_weight

4b)

The `intercept` -0.357 describes the average outcome of Heart weight if body weight = 0

4c)

The `slope` of 4.034, means that for one 1kg increase / decrease on body_weight, the heart_weight is estimated to increase / decrease by 4.034g.

4d)

R^2 is 0.4489. This mean proportion of the variability in heart_weight that is explained by the body_weight is 0.4489

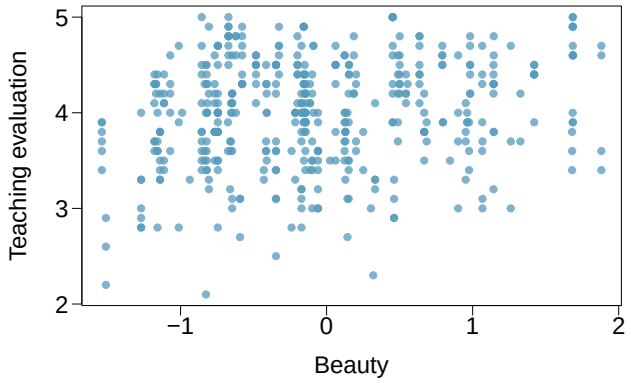
4e)

```
cor(cats$Hwt, cats$Bwt)
```

```
## [1] 0.8041274
```

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: (Rate my professor)

5a)

The slope of the model is 0.133

```
summary(m_eval_beauty)
```

```
##
## Call:
## lm(formula = eval ~ 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 ***
## beauty       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
```

```
cor(beauty, eval)
```

```
## [1] 0.1890391
```

5b)

The summary above shows that the p-value of the slope is very small and less than the confidence value. Also, the correlation coefficient between beauty and teaching evaluation is greater than Zero(0). Hence it is significant that there is a positive relationship between eval and beauty.

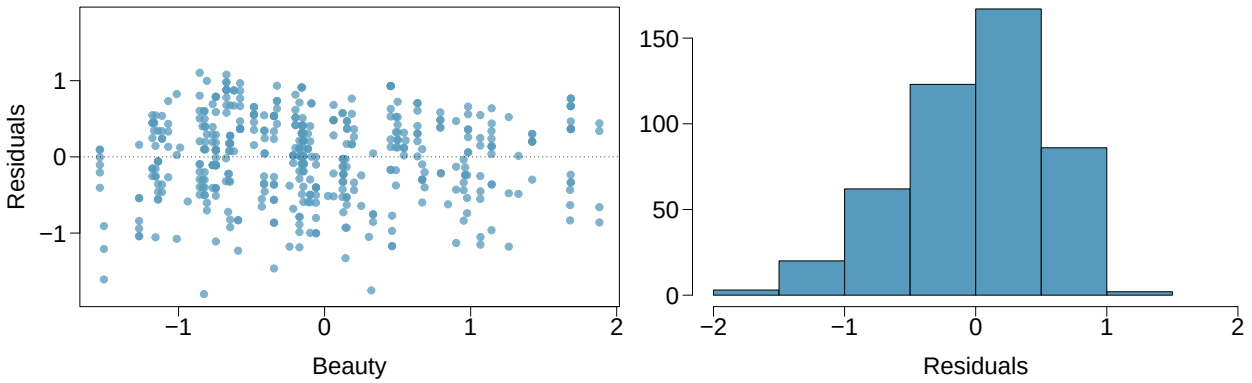
5c)

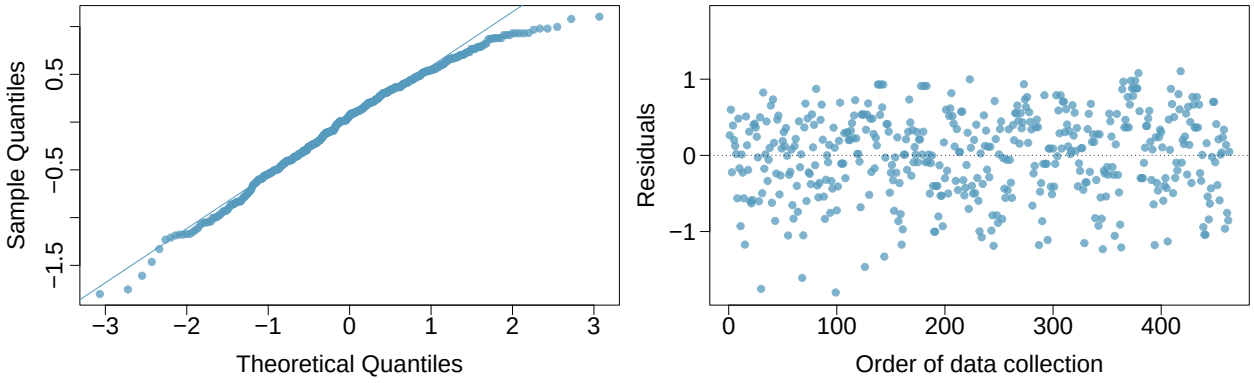
Linearity: The residuals in the scatter plot distributed randomly around y=0. No obvious shapes or patterns is found, which shows strong linearity between beauty score and teaching evaluation score.

Nearly normal residuals: Histogram seems nearly normal with negative skew

Constant Variability Condition: Most of the residuals are within the arrange of -1 and 1. Therefore the constant variability condition appears to be met.

Independent observations: could not be exactly determined, could be assumed that successive observation are independent.




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