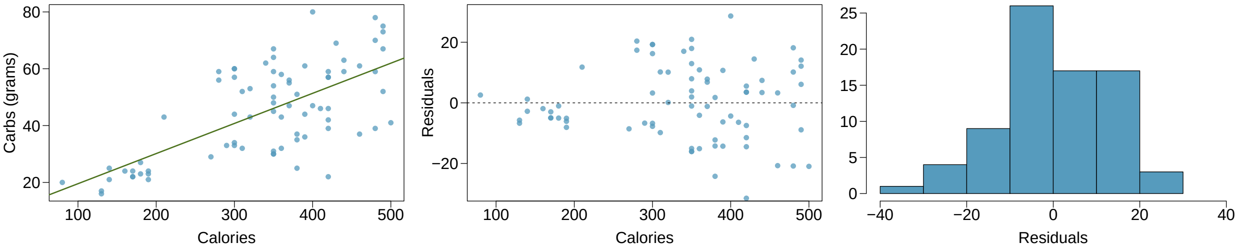
Chapter 8 - Introduction to Linear Regression

Donald Butler

11/07/2021

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.

The relationship between the number of calories and the amount of carbohydrates is positive. As the number of calories increases, the amount of carbohydrates increases.

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

The explanatory variable is the number of calories in the Starbucks food item and the response variable is the amount of carbohydrates the item contains.

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

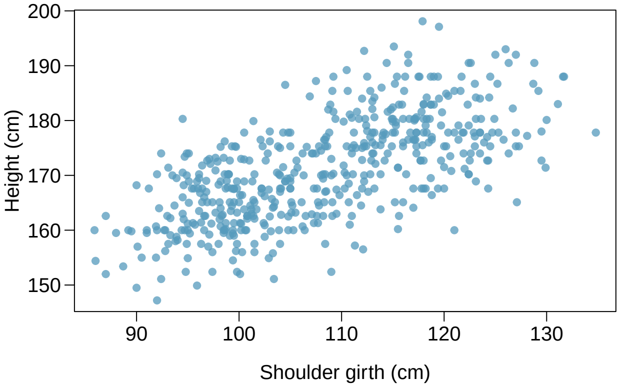
Since Starbucks is not listing the number of carbohydrates for each menu item, just the number of calories, we might want to fit a regression line to try to estimate the amount of carbohydrates in each menu item.

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

Linearity: The relationship appears to follow a linear trend.
Nearly normal residuals: The histogram of the residuals appear to be nearly normal centered at 0.
Constant Variability: The lower calorie items do have less residual variability, but not significantly so.
Independent observations: The observations are not based on a time-series and are independent of each other.

The four conditions are satisfied and a least squares regression can be applied.

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 appears to be a positive linear 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?

The relationship would not change, just the slope of the regression line by the ratio of 1 inch to 1 cm, about 1/2.5.

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.

```
s_x <- 10.37
mean_x <- 107.20
s_y <- 9.41
mean_y <- 171.14
r <- 0.67

m <- (s_y/s_x) * r
intercept <- mean_y - m * mean_x
```

$$hei\hat{g}ht = 105.965 + .608 \times shoulder\_girth$$

(b) Interpret the slope and the intercept in this context.

The intercept is the height estimate for an individual with 0 shoulder girth. The slope is the increase in height attributed to a unit increase in shoulder girth.

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

$R^2 = 0.4489$

This indicates that 44.89% of variations in height can be explained by variations in 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.

```
(height_estimate <- intercept + m * 100)
```

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

The model would predict a height of 166.7625805 cm.

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

```
(residual <- 160 - height_estimate)
```

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

The residual of -6.7625805 indicates that the model overestimated the students height.

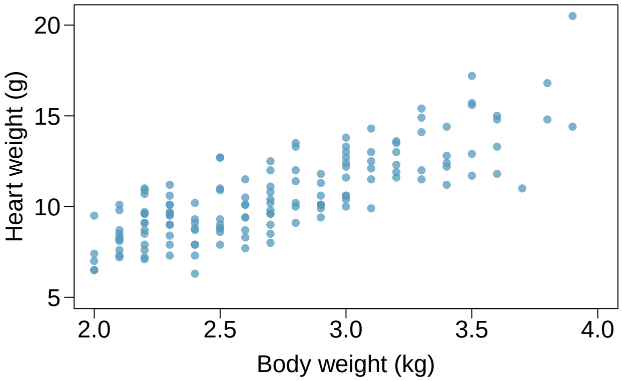
(f) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

A child with shoulder girth of 56 cm is significantly lower than the observations used to construct the model. Since we would need to extrapolate from the data our model is based on, it would not be an appropriate estimate.

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.

$$\widehat{heart}_w t = -0.357 + 4.034 \times body\_weight$$

(b) Interpret the intercept.

A cat with body weight of 0 would have a heart weight of -0.357. This is obviously not a valid weight for a cat.

(c) Interpret the slope.

For each increase of 1 kg in a cat's body weight, the heart weight will increase by 4.034 grams.

(d) Interpret R^2.

64.66% of the variability in the weight of a cat's heart is explained by variations in the cat's body weight.

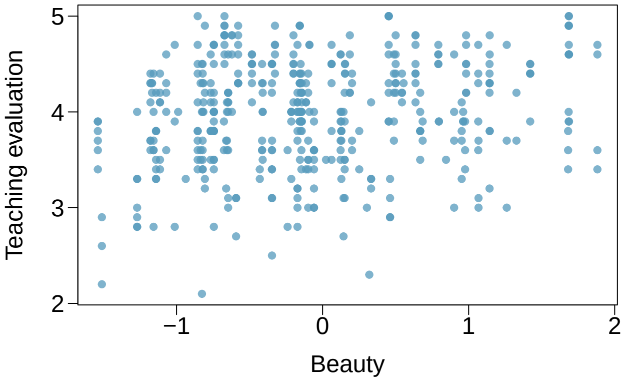
(e) Calculate the correlation coefficient.

```r
sqrt(.6466)
```

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

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.

```
summary(m_eval_beauty)
```

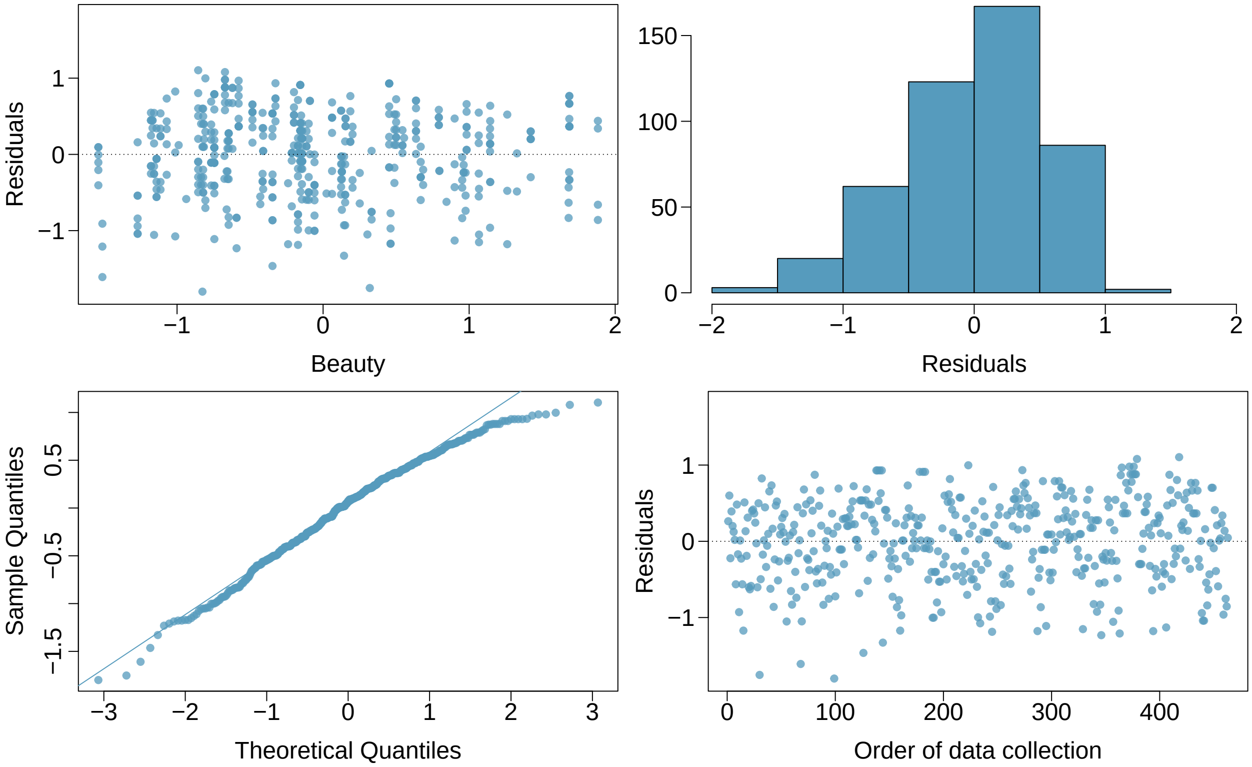
```
##
## Call:
## lm(formula = eval ~ beauty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.80015 -0.36304  0.07254  0.40207  1.10373
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.01002    0.02551 157.205  < 2e-16 ***
## beauty       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
```

The slope is 0.133 which indicates that a unit increase in beauty score will result in a .133 increase in teaching evaluation score.

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

Yes, since the p-value of the beauty coefficient is less than .05, we conclude that there is a positive correlation.

(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: There appears to be a linear trend.
Nearly normal residuals: The histogram of the residuals appear to be nearly normal centered at 0.
Constant Variability: The residuals appear to have constant variability.
Independent observations: The observations are not based on a time-series and are independent of each other.