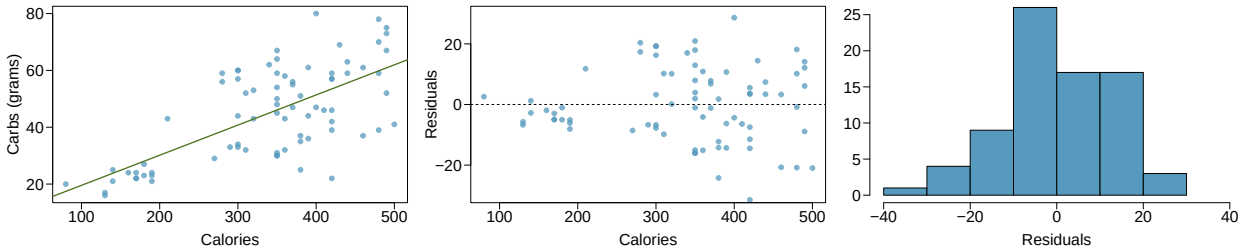
Chapter 8 - Introduction to Linear Regression

David Simbandumwe

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.

- **direction - positive**
- **form - linear**
- **strength - strong**

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

- **explanatory - calories**
- **response - carbs**

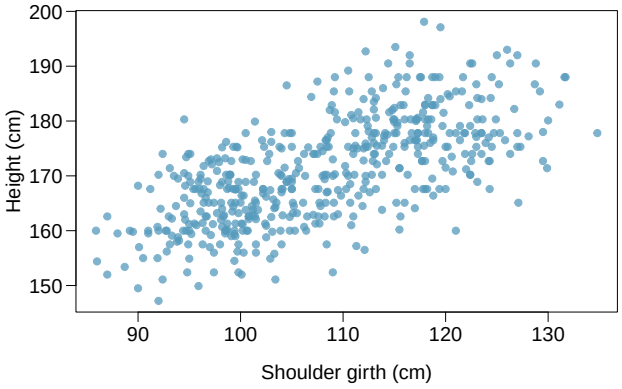
(c) Why might we want to fit a regression line to these data?

- **for prediction purposes so that we can predict the carbs for a food item based on the calory count**

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

- **yes - it is a linear relationship with a normally distributed residual, the observations are independent, constant variability, no outliers**

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.

- **the relationship is a strong 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 should scale based on the changing measures. The intercept should remain unchanged however the slope of the regression line should be larger. since each unit change in shoulder girth would be associated with larger change in height**

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.

$$y = b_0 + b_1 \times x_0$$

- **slope**

$$b_1 = \frac{S_y}{S_x} \times R$$

- **intercept**

$$y - y_0 = b_1 \times (x - x_0)$$

```
Sy <- 9.41
Sx <- 10.37
R <- 0.67

(b1 <- (Sy/Sx)*R)
```

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

```
y_mean <- 171.14
x_mean <- 107.20
x <- 0


(y_int <- b1 * (x - x_mean) + y_mean)
```

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

$$y = 105.965 + 0.608 \times x_0$$

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

- **for a girth of 0 the predicted height is the intercept (105.965) and for every 1 cm increase in girth the predicted heigth should increase by 0.608**

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

```
R <- 0.67

(R_squared = R^2)
```

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

- **R Squard:** 0.449

- **The regression line explains 44.9% of the variability in height**

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

$$y = 105.965 + 0.608 \times x_0$$

```
x <- 100

(y = 105.965 + 0.608 * x)
```

```
## [1] 166.765
```

- **height: 166.765**

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

$$e_x = y_x - \hat{y_x}$$

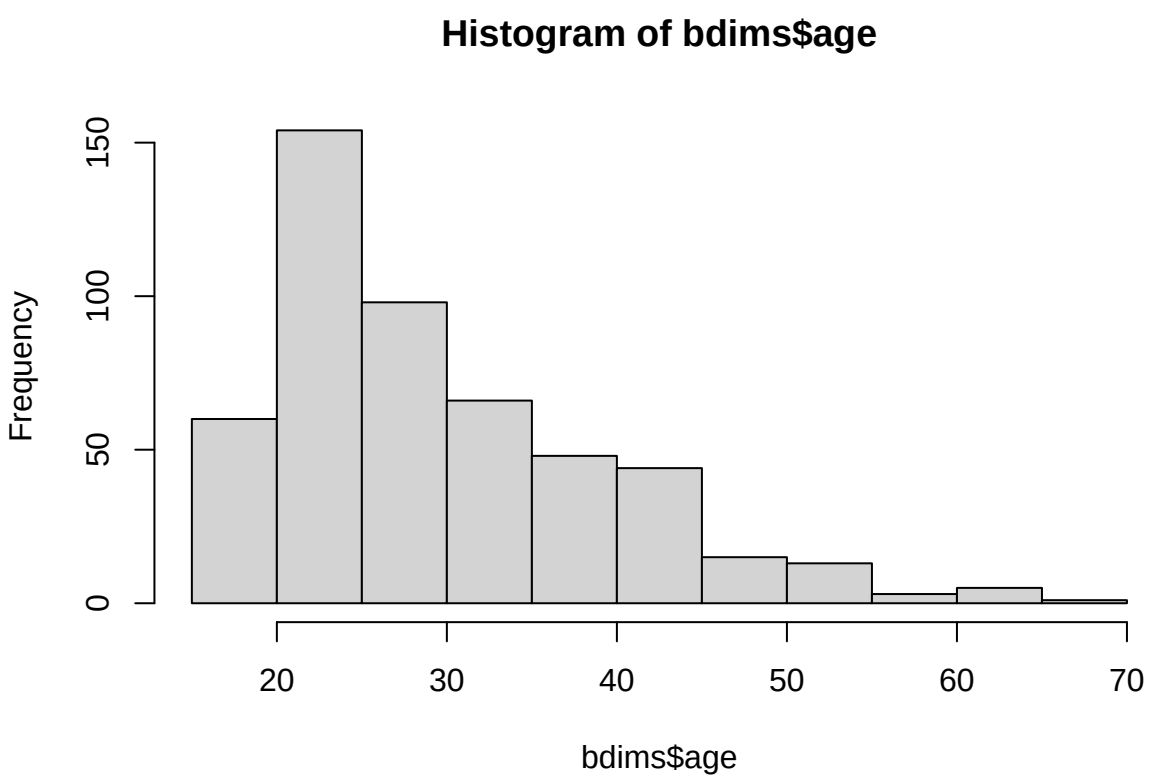
```
(e_x <- 160 - 166.765)
```

```
## [1] -6.765
```

- **residual: -6.765**
- **The regression line has an -6.765cm error when it is used to predict the height of the student with a 100cm girth**

(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?

```
hist(bdims$age)
```




```
summary(bdims$age)
```

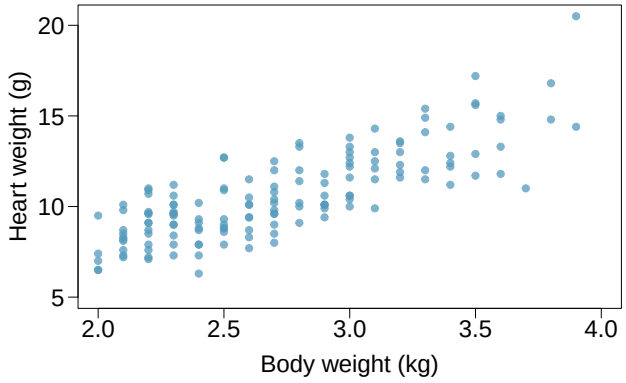
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   23.00   27.00   30.18   36.00   67.00
```

- **The dataset has a min age of 18 and a max age of 67 a child one years old is outside of the modeled data**

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.

$$y = b_0 + b_1 \times x_0$$

- **intercept: -0.357**
- **slope: 4.034**

$$y = -0.357 + 4.034 \times x_0$$

(b) Interpret the intercept.

- **intercept: -0.357**
- **The intercept is the predicted heart weight of a cat with 0 body weight. It is a theoretical prediction because a weigth of 0 for a living animal is an impossiblity**

(c) Interpret the slope.

- **slope: 4.034**
- **For each kg increase in body weight the predicted heart weigth will increase by 4.034grams**

(d) Interpret R^2.

- **The model predicts 64.66% of the relationship between body weight and heart weight**
- **r squared: 0.6466**

(e) Calculate the correlation coefficient.

$$r = \sqrt{r^2}$$

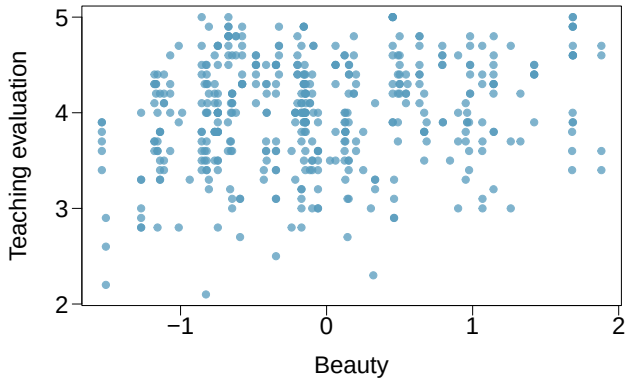
```r
(r <- sqrt(0.6466))
```

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

- **correlation coefficient: 0.804**

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.

$$y - \bar{y} = b_1 \times (x - \bar{x})$$

$$b_1 = \frac{y - \bar{y}}{x - \bar{x}}$$

```
x <- 0
x_bar <- -0.0883
y <- 4.010
y_bar <- 3.9983

(b1 <- (y - y_bar) / (x - x_bar))
```

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

- **slope: 0.133**

(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 small at 0.133**

- **the p-value is 0 so we would reject the null hypothesis in favor of the alternate hypothesis. There is evidence to support the positive relationship between beauty and evaluation**
- **the plot shows no clear associati between the beauty score and evaluation but the intercept and the p-value suggest otherwise**

(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 - data shows a linear relationship between the variables**
- **near normal residual - the residuals should follow a normal distribution**
- **constant variability - the variability of points around the least square lines should be consistent**
- **independent observations - observatiosn should be independent**