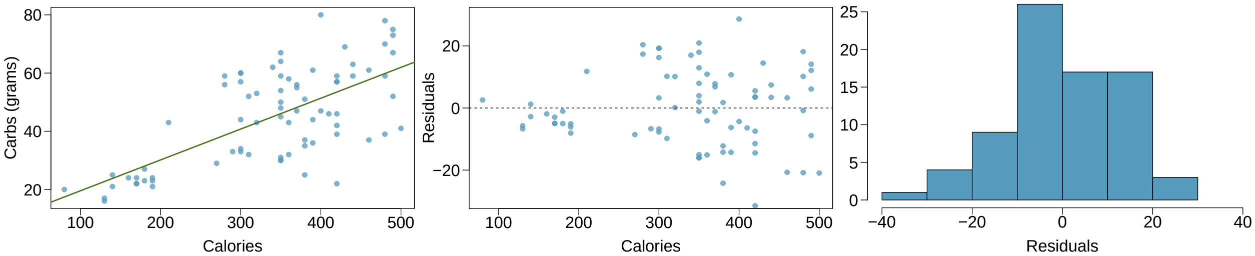
Chapter 8 - Introduction to Linear Regression

Leticia Salazar

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.

Based on the graph, there's an upward positive trend, as the amount of calories increase so do the carbs.

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

Explanatory Variable: in this scenario, the calories is the explanatory variable Response Variable: the carbs would be the response variable in this scenario

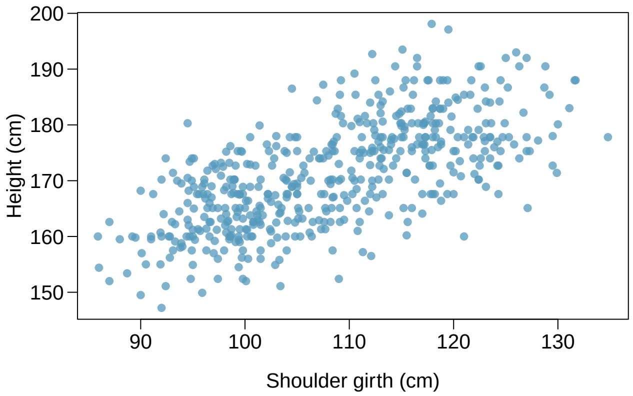
(c) Why might we want to fit a regression line to these data?

We would want to fit a regression line in order to make better predictions about the data. By doing so in this data set we'd be able to predict the amount of carbs based on the calories.

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

Based on the plot, the data fits a linear plot, the residuals appear to be normal and we cannot achieve constant variability. We cannot be sure if the variables are independent since this is from the Starbucks menu. Therefore, it does not satisfy the conditions required for a fitting a least squares line.

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.

Based on the plot, it seems to have a positive upward trend where as the shoulder girth increases so does the height.

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

The relationship wouldn't change if the units changed to inches for shoulder girth. It will still show a positive upward trend.

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.

The equation of regression line for predicting height is: y = b0 + b1*x or height = 105.97 + 0.608* girth.

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

The slope is .608, which is the rate of increase for the height in centimeters and the intercept is 105.97 is the height in centimeters of the shoulder girth.

```
# Shoulder girth mean and standard deviation
sg_mean <- 107.20
sg_sd <- 10.37

# Height mean and standard deviation
h_mean <- 171.14
h_sd <- 9.41

# Correlation between height and shoulder girth
corr <- 0.67

# Calculate slope
b1 <- (h_sd / sg_sd) * corr
b1
```

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

```
# Calculate Intercept
b0 <- h_mean - (b1 * sg_mean)
b0
```

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

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

The R^2 is 45% meaning that 45% of the variation is found within this data.

```
# $R^2$
r_sqr <- corr^2
r_sqr
```

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

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

The predicted height of the student is 166.76 cm.

```r
# shoulder girth
g <- 100

# Predict height
predict_height <- b0 + b1 * g
predict_height
```

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

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

The residual is -6.76 meaning that the model overestimated the height of the student.

```r
# Calculate the residual
act_height <- 160
residual <- act_height - predict_height
residual
```

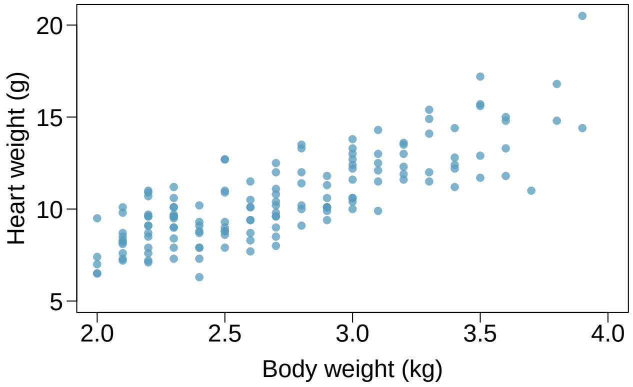
```
## [1] -6.762581
```

(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?

Being that a one year old's shoulder girth is 56 cm, it is otuside of the range in the sample 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 = b0 + b1*x or heart weight(g) = -0.357 + 4.034* body weight(kg)

(b) Interpret the intercept.

With 0kg being the body weight, the intercept is -0.357 of the heart weight. Since there is no such thing as a 0 kg body weight, this result is not meaningful.

(c) Interpret the slope.

For the slope, for every additional body weight(kg) there is an expected additional 4.034g increase in heart weight.

(d) Interpret R^2.

The body weight explains the 65% variability in the heart weight of the cats.

(e) Calculate the correlation coefficient.

The correlation coefficient is 0.804 and being that it is greater than 0.5 there is a positive correlation.

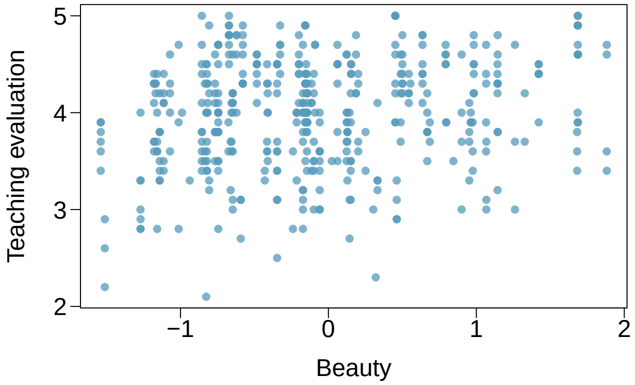
```r
R2 <-  0.6466

# Correlation Coefficient
cc <- sqrt(R2)
cc
```

```
## [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.

The slope is -0.13

```
# y = b0 + b1*x
b0 <- 4.010
y <- 3.9983
x <- 0.0883

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

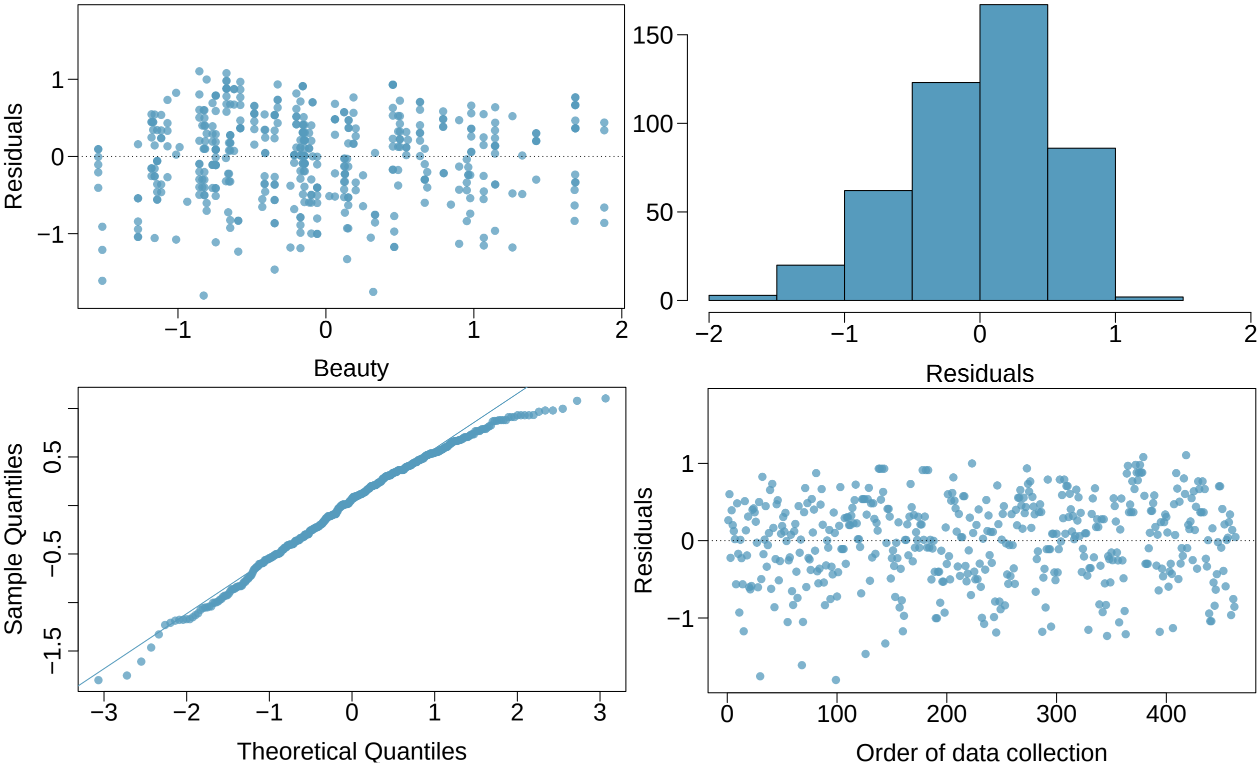
```
## [1] -0.1325028
```

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

The slope found is positive because the p-value is almost 0.

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

Based on the plots below, the conditions for linear regression are met. The data is almost linear, it's a random observation and assumes that the points are independent from each other and the distribution appears to be normal.


Residuals
Beauty
Residuals
Sample Quantiles
Theoretical Quantiles
Residuals
Order of data collection