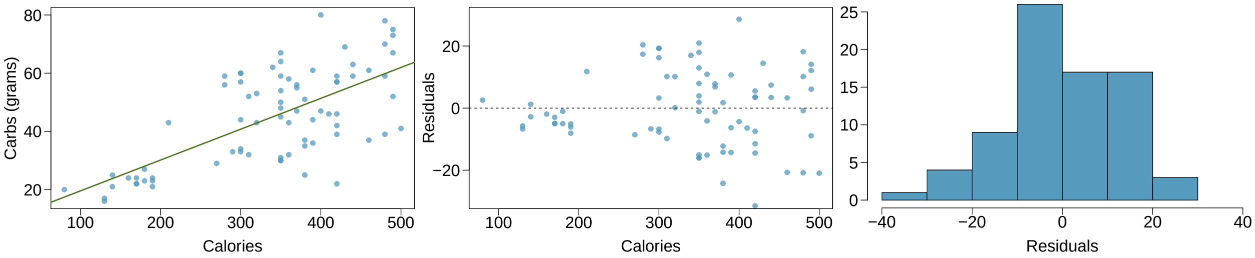
Chapter 8 - Introduction to Linear Regression

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.

There is a strong positive relationship betwen the number of calories and amount of carbohydrates in the Starbucks food menu.

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

In this case, the response variable is calories and the explanatory variable is carbohydrates.

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

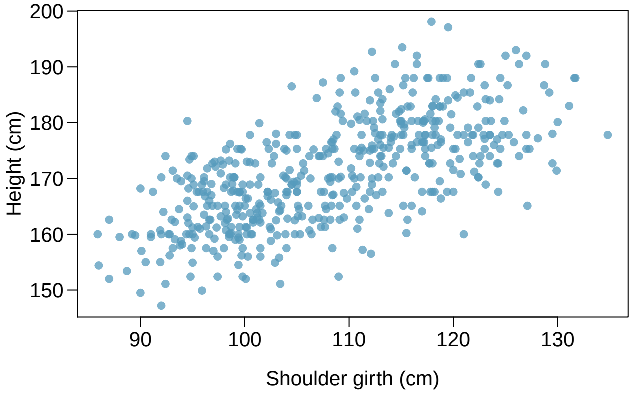
We may want to fit a regression line to extrapolate the calorie count of new item added to the menu if we knew its carbohydrates (in grams).

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

The 3 conditions for fitting a least squares line are met.

1. Linearity: Residuals are evenly spaced above the zeroth line.
2. Nearly normal residuals: The residuals look normally distributed.
3. Constant variability

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.19 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 positive and linear. Increasing shoulder girth correlates with increasing 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. The x,y value pairs will change as we use a different system of units, but the trend will still be the same.

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 slope of the regresion can be calculated using the relationship:

$$b_1 = \frac{s_y}{s_x} R$$

```r
S.y <- 9.41
S.x <- 10.37
R <- 0.67
b.1 <- (R * S.y/S.x) %>% round(3)
cat("The slope of the regresion is:",b.1)
```

```
## The slope of the regresion is: 0.608
```

The intercept of the regresion can be calculated using the relationship:

$$b_0 = \bar{y} - b_1\bar{x}$$

```r
y.mean <- 171.14
x.mean <- 107.20
b.0 <- (y.mean - b.1 * x.mean) %>% round(3)
cat("The intercept of the regresion is:",b.0)
```

```
## The intercept of the regresion is: 105.962
```

Finally, the equation for the regression line is:

$$y(Height) = 105.962 + 0.608 * x(Girth)$$

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

The minimum height even for the lowest shoulder girth is around 106 cms.

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

We can calculate R^2 by squaring the correlation between height and shoulder girth of 0.67. Using this value R^2 is 0.449

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

```r
given.girth <- 100
predicted.h <- 105.962 + 0.608 * given.girth
cat("The predicted height of this student using the model is", predicted.h, "centimeters.")
```

```
## The predicted height of this student using the model is 166.762 centimeters.
```

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

Residual is the difference between the observed (y_i) and predicted
$\hat{y}_i$.
$e_i = y_i - \hat{y}_i$

The residual in this case is ei. This residual just means that the real observed value is less than that of the prediction.

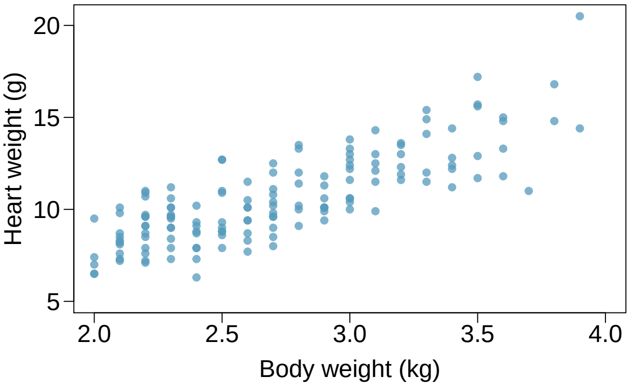
(f) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

It will not be appropiate because the morphology of a one year old has different ratios of height and width as compared to those of an adult.

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

$s = 1.452$ $\quad R^2 = 64.66\%$ $\quad R^2_{adj} = 64.41\%$



(a) Write out the linear model.

$$y(HeartWeight) = -0.357 + 1.452 * x(BodyWeight)$$

(b) Interpret the intercept.

The negative intercept is the estimate of the expected mean Heart Weight when all the explanatory predictors are at zero. Which in this case is only the Body Weight. IN this case the intercept is a mathematical abstraction that adjust predicted values of Y relative to the plot.

(c) Interpret the slope.

The slope indicates a positive relationship between Body Weight and Heart Weight. Increasing body weight correlates with an increasing heart weight.

(d) Interpret R^2.

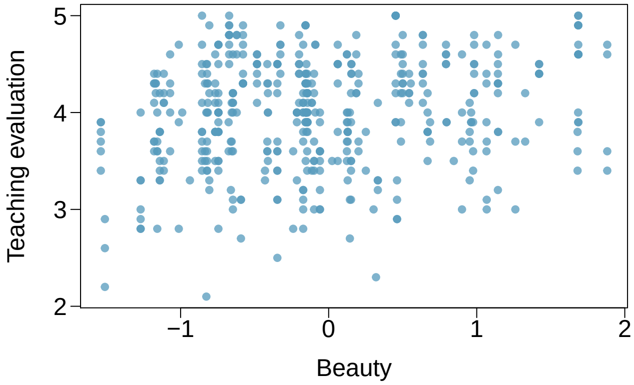
The R^2 value indicates that 64.66% of the variability of the Heart Weight data is explained by the Body Weight parameter.

(e) Calculate the correlation coefficient.

The correlation coefficient is the square root of the Rˆ2 value. In this case the correlation coefficient is 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.

The slope of the regresion can be calculated using the relationship:

$$b_0 = \bar{y} - b_1\bar{x}$$

```r
b.0 <- 4.010
y.mean <- 3.9983
x.mean <- -0.0883
b.1 <- ((y.mean - b.0)/x.mean) %>% round(3)
cat("The slope of the regresion is:",b.1)
```

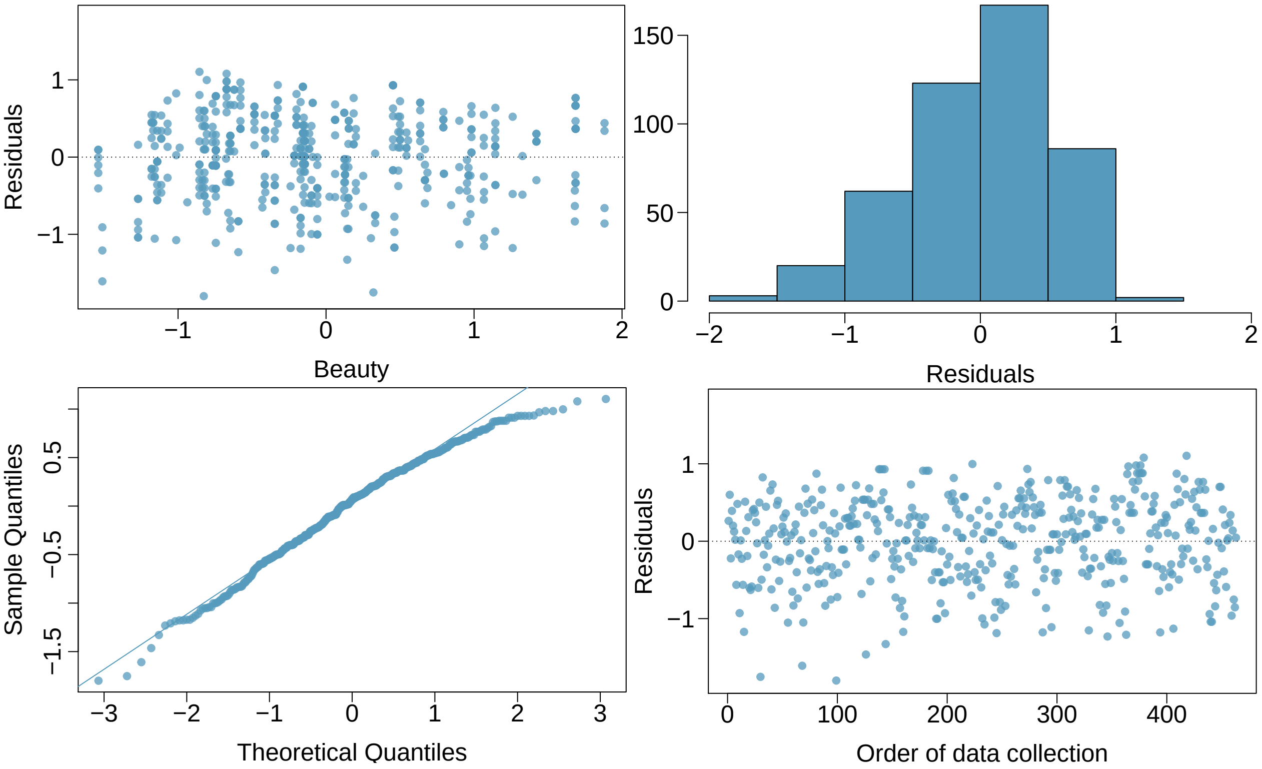
```
## The slope of the regresion is: 0.133
```

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

There is a weak positive response between teaching evaluation and beauty. Is it convincing? Even when we don't have a correlation coefficient or an R^2 value, the t-value of 4.13 indicates that the probability that beauty does not explain teaching evaluation is very low.

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

1. Linearity: The relation between teaching evaluation and beauty is linear and positive.
2. Nearly normal residuals: The residuals appear nearly normal even if slightly negative or left skew.
3. Constant variability: The residuals are evenly space around the zeroth horizontal line.


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