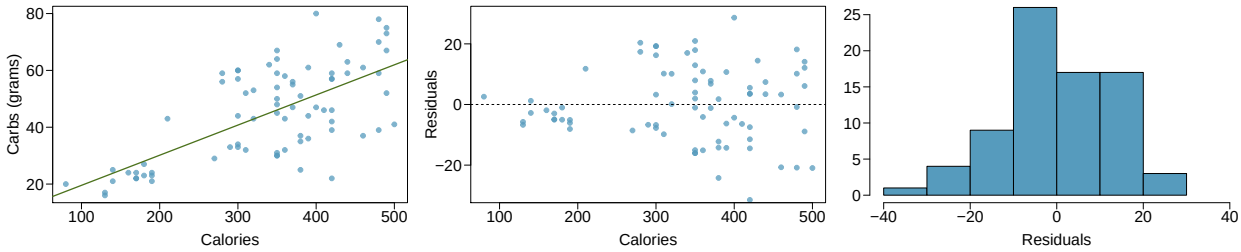
Chapter 8 - Introduction to Linear Regression

Shana Green

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 direction is positive because the pattern in the scatterplot slopes upward. The form is linear and the strength is weak because the data is not very clear. The points are spread far apart.

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

The explanatory variable is the calories and the response variable is the carbohydrates (gram).

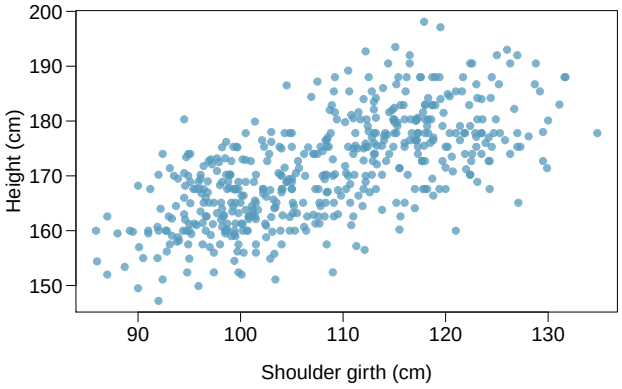
(c) Why might we want to fit a regression line to these data?

We might want to fit a regression line to these data so we can predict the amount of carbohydrates of a particular food item with a known number of calories.

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

The conditions required are: Independent observations, linear relationship, equal variance, and normal residuals. The equal variance is not satisfied because the vertical spread in the residual plot is much higher to the right in the residual plot. Since that condition is not met, it is not appropriate to fit 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.

The direction is positive because the pattern in the scatterplot slopes upward. The form is linear and the strength is moderate because the data is clear. The points are not close together.

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

If we change the unit of measure, it would not affect the relationship.

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.

$$\beta_1 = r\frac{s_y}{s_x}$$
$$\beta_1 = r\frac{s_h}{s_g}$$

```
girth_mean <- 107.20
girth_sd <- 10.37
height_mean <- 171.14
height_sd <- 9.41
r <- 0.67

slope <- round(r * (height_sd / girth_sd),5)
slope
```

```
## [1] 0.60797
```

```
intercept <- height_mean - (slope * girth_mean)
intercept
```

```
## [1] 105.9656
```

$\beta_0 = 105.9656$ **and** $\beta_1 = 0.60797$**. The equation of the regression line is** $y = 105.9656 + 0.60797x$

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

The height increases on average by 0.60797 cm per cm of shoulder girth. The average height at a shoulder girth at 0 cm is 105.9656 cm.

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

```
rr<-r^2

rr
```

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

```
rr*100
```

```
## [1] 44.89
```

There is 44.89% variability in height can be explained by the 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.

```r
x<-100

student_100<-(slope * x) + intercept
student_100
```

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

The predicted height for the random student with a shoulder girth of 100 cm is 166.7626.

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

```r
y<-160

residual<-y-student_100
residual
```

```
## [1] -6.762616
```

The observed value is 160 cm and the predicted value is 166.7626. We overestimated the height by -6.762616 cm.

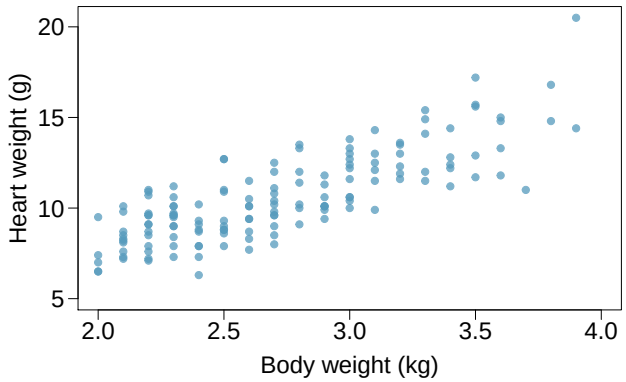
(f) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

Based on the previous scatterplot, the range of the shoulder girth data set is between 80 and 140 cm. 56 cm is not within the range, so it is not appropriate to use this linear model.

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.

$$y = -0.357 + 4.034x$$

(b) Interpret the intercept.

The intercept is 0.357 grams. This is not realistic because a weight of 0 kg should correspond with a heart weight of 0 gram.

(c) Interpret the slope.

The heart weight increases on average by 4.034 grams per kilograms of body weight.

(d) Interpret R^2.

$R^2 = 64.66 = 0.6466$

64.66% of the variation in the heart weight is explained by the straight-line relationship between the body weight and the heart weight.

(e) Calculate the correlation coefficient.

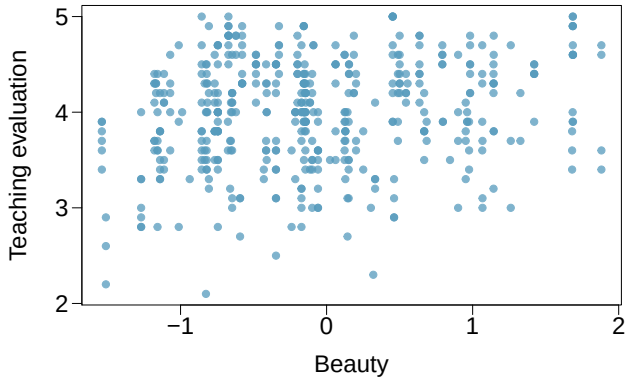
```
correlation<-0.6466
cc<-sqrt(correlation)

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.

```
x<--0.0883
y<-3.9983
b<-4.010

slope<-(y-b)/x
slope
```

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

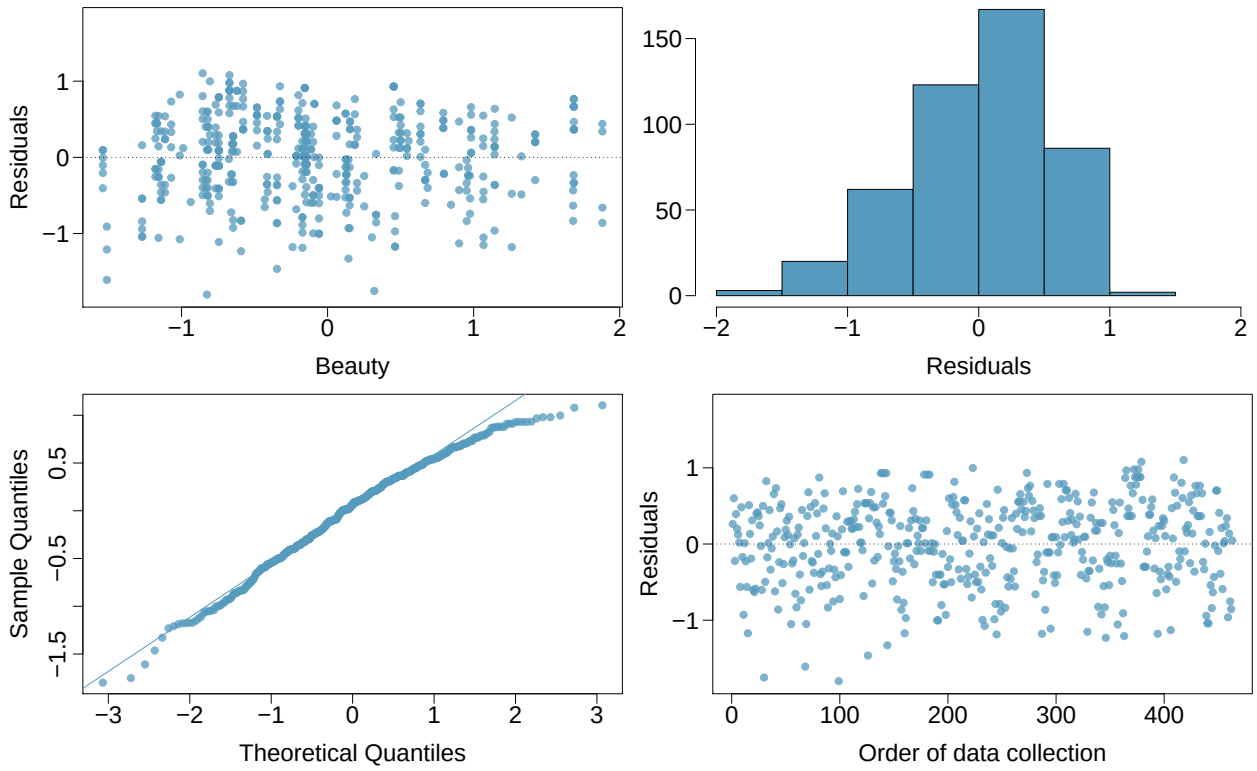
(b) Do these data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive? Explain your reasoning.

$$**SE_b = 0.0322**$$

Hypothesis:

$$**H_0 : \beta = 0**$$

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


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