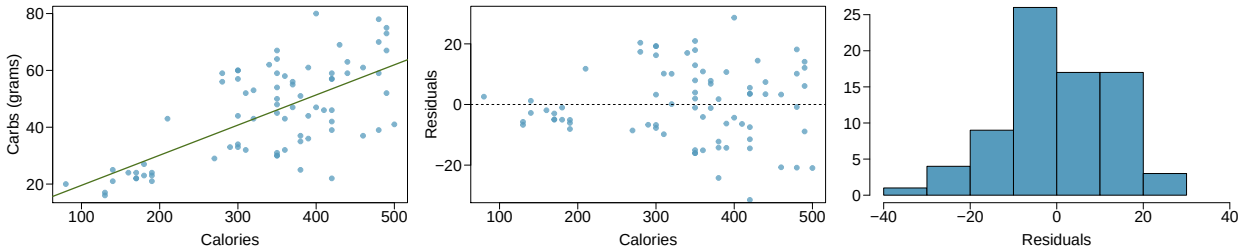
Chapter 8 - Introduction to Linear Regression

Zachary Palmore

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 linear relationship between calories and carbohydrates.

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

In this scenario the number of calories is the explanatory variable and the number of carbohydrates measured in grams is the response.

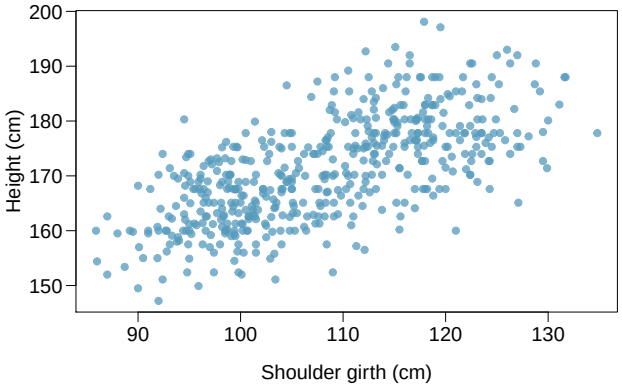
(c) Why might we want to fit a regression line to these data?

We might want to fit a regression line to these data to see if there is a correlating trend and determine how well they are correlated. In this scenario, we fit a regression line to find the relationship between calories and carbohydrates. As the number of calories increase, so too do the carbohydrates. Depending on the results of the relationship, the explanatory variable, in this case the number of calories, may be a good predictor of the number of carbohydrates in an item.

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

Yes, the data displays a linear trend. The data is also independent and originates from a random sample. The residuals are nearly normal and variability of points around the least squares line remains roughly constant.

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 is a relatively strong, positive, linear relationship between shoulder girth and height.

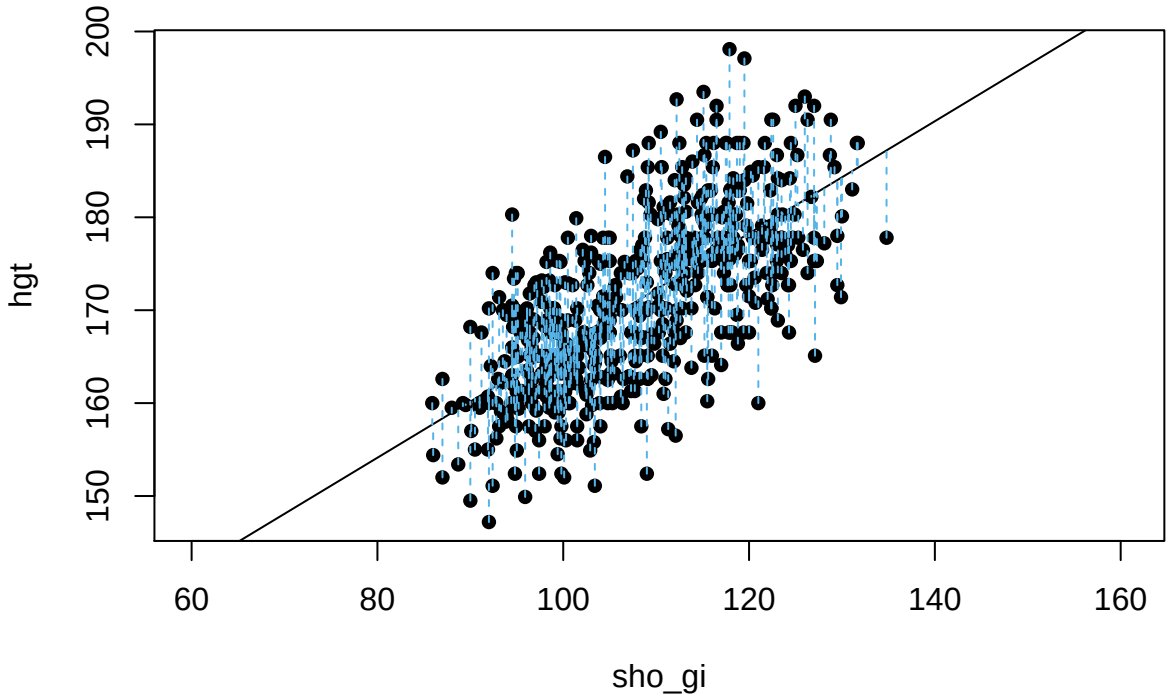
(b) How would the relationship change if shoulder girth was measured in inches while the units of height remained in centimeters?

If girth was measured in inches while height remained in centimeters it would appear as though the graph was squished on the x-axis and the positive linear trend would look stronger and more vertical. However, this would only be due to the change in x-axis values. The range of girth in centimeters spans 40 cm while the range in inches spans only 16 because inches are a larger measure than centimeters.

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.

```
m_gir<-107.2
s_gir<-10.37
m_hgt<-171.14
s_hgt<-9.41
gir.hgt.cor<- 0.67
plot_ss(x = sho_gi, y = hgt, data = bdims, showSquares = FALSE)
```




```
## Click two points to make a line.
## Call:
## lm(formula = y ~ x, data = pts)
##
## Coefficients:
## (Intercept)            x
##    105.8325       0.6036
##
## Sum of Squares:  24932.64
```

The equation to predict height is $y = 105.8325 + 0.6036(x)$ where y is representative of height and x is the shoulder girth.

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

The slope is small positive number indicating that for every increase in shoulder girth there is a small incremental increase in height. It also indicates a positive direction of the data.

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

```
hgt_shogi_lm <- lm(hgt ~ sho_gi, data = bdims)
summary(hgt_shogi_lm)$r.squared
```

```
## [1] 0.4432035
```

The r-squared value is about 0.4432.

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

Using the equation $y = 105.8325 + 0.6036(x)$ we can compute a height (y) for a shoulder girth of 100 (x = 100).

```
x <- 100
105.8325 + 0.6036*(x)
```

```
## [1] 166.1925
```

At a shoulder girth of 100 cm, the predicted height is about 166.2 cm.

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

```
160-166.2
```

```
## [1] -6.2
```

The residual for a height of 160 is -6.2. This means that our linear model overestimated this height by 6.2 cm. If we had used only the trend line itself to estimate the height we would have estimated 6.2 cm above the actual height.

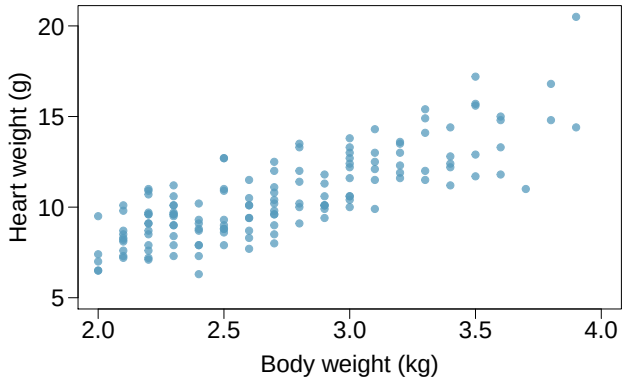
(f) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

No, there is no data available at a shoulder girth lower than about 90. Estimating the height of a child with a shoulder girth that small would be a gross extrapolation.

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

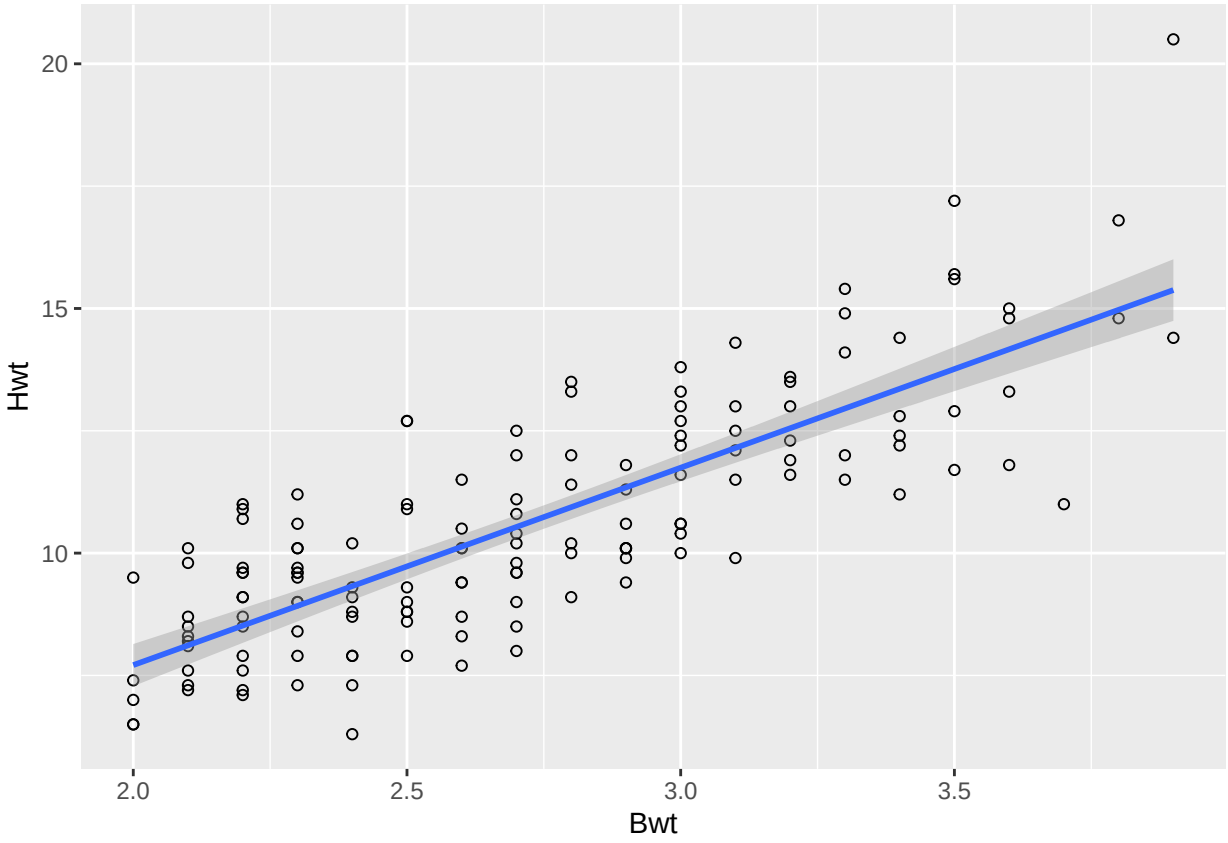
$s = 1.452$ $R^2 = 64.66\%$ $R^2_{adj} = 64.41\%$



(a) Write out the linear model.

```
ggplot(cats, aes(x = Bwt, y = Hwt)) + geom_point(shape=1) + geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```r
cats %>%
  summarise(cor(Bwt, Hwt, use = "complete.obs"))
```

```
##   cor(Bwt, Hwt, use = "complete.obs")
## 1                           0.8041274
```

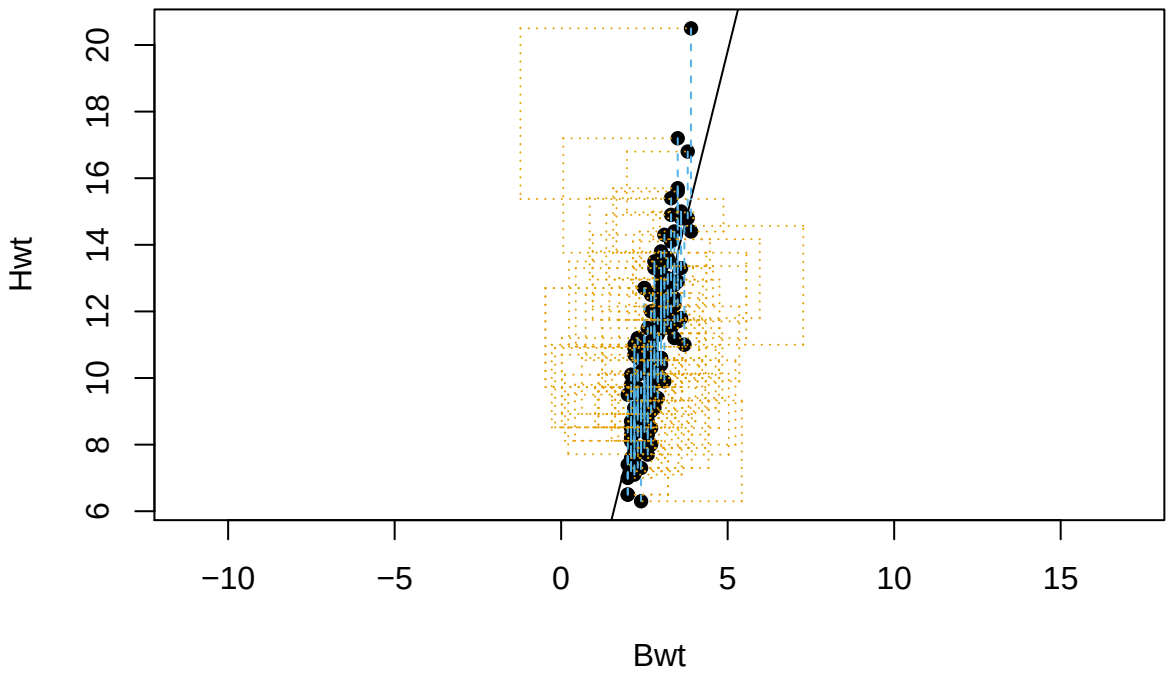
```r
cor.test(cats$Bwt, cats$Hwt)
```

```
##
##  Pearson's product-moment correlation
##
## data:  cats$Bwt and cats$Hwt
## t = 16.119, df = 142, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7375682 0.8552122
## sample estimates:
##       cor
## 0.8041274
```

```r
cts.bhwt <- lm(Bwt ~ Hwt, data = cats)
summary(cts.bhwt)
```

```
##
## Call:
## lm(formula = Bwt ~ Hwt, data = cats)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58283 -0.22140 -0.00879  0.20825  0.91717
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.019637   0.108428   9.404   <2e-16 ***
## Hwt         0.160290   0.009944  16.119   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2895 on 142 degrees of freedom
## Multiple R-squared:  0.6466, Adjusted R-squared:  0.6441
## F-statistic: 259.8 on 1 and 142 DF,  p-value: < 2.2e-16
```

```r
plot_ss(x = Bwt, y = Hwt, data = cats, showSquares = TRUE)
```




```
## Click two points to make a line.
## Call:
## lm(formula = y ~ x, data = pts)
##
## Coefficients:
## (Intercept)            x
##     -0.3567       4.0341
##
## Sum of Squares:  299.533
```

The linear model for this data is represented by the equation y = -0.3567 + 4.0341(x)

(b) Interpret the intercept.

The intercept is just below zero indicating that when Bwt is equal to zero the cats Hwt is negative. This of course defies the idea of matter having mass, so in context we can interpret this to mean, there are no points of Hwt when their Bwt is zero. Furthermore, we can say from the intercept that the data only begins to show Hwt after Bwt is above zero.

(c) Interpret the slope.

The slope is 4.0341 indicating a steep, upward positive relationship between the Bwt and Hwt of cats.

(d) Interpret R^2.

The adjusted R^2 is 0.6441. About 64% of the variance in the dependent variable (Hwt) can be explained by variation in the independent variable (Bwt).

(e) Calculate the correlation coefficient.

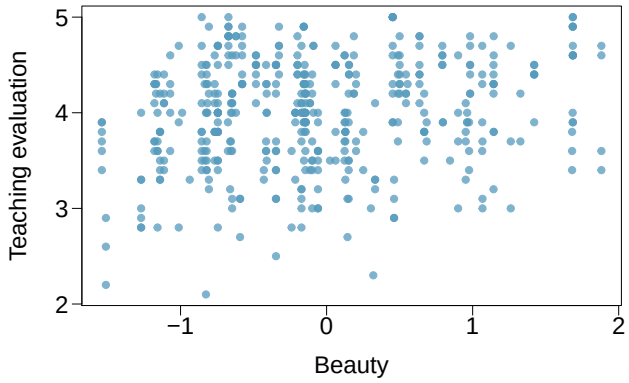
The correlation coefficient is 0.8041.

```
cats %>%
  summarise(cor(Bwt, Hwt, use = "complete.obs"))
```

```
##   cor(Bwt, Hwt, use = "complete.obs")
## 1                           0.8041274
```

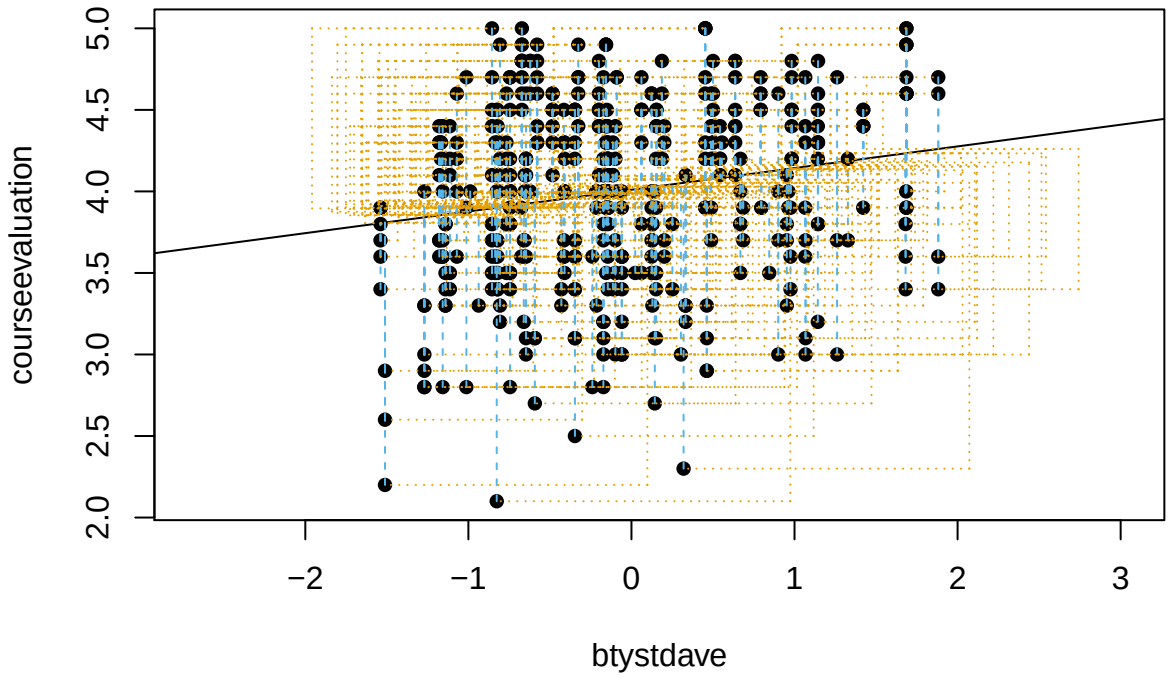
Rate my professor. (8.44, p. 340) Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures 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.

```
plot_ss(x = btystdave, y = courseevaluation, data = prof_evals_beauty, showSquares = TRUE)
```




```
## Click two points to make a line.
## Call:
## lm(formula = y ~ x, data = pts)
##
## Coefficients:
## (Intercept)            x
##       4.010        0.133
##
## Sum of Squares:  137.156
```

The slope 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.

No, the slope is positive in this instance however, this could be due to random chance variation. It is only exhibiting a slightly positive trend.

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

The conditions are:

* Linearity

```
* Nearly Normal Residuals
* Constant Variability
* Independent Observations
```

Based on the plots below, linearity is met as the data does show a linear trend. Residuals are also nearly normal as there are no outliers or overly influential points. Variability in the residuals is also relatively constant and observations are independent of one another which means all conditions are satisfied.