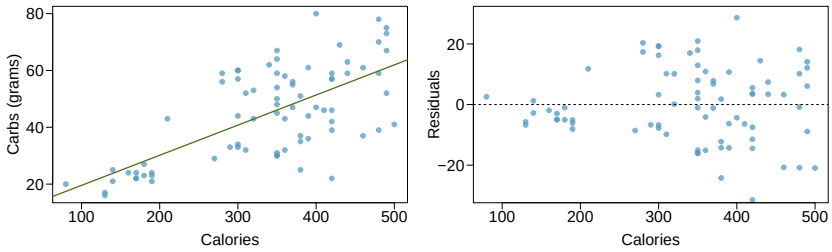
Chapter 8 - Introduction to Linear Regression

Jack Wright

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.

```
## numeric(0)
```

```
## numeric(0)
```



(a) Describe the relationship between number of calories and amount of carbohydrates (in grams) that Starbucks food menu items contain.

There is a positive linear relationship between the number of calories and the number of carbohydrates in a menu item.

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

The calories is the explanatory variable and the carbs are the response variable.

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

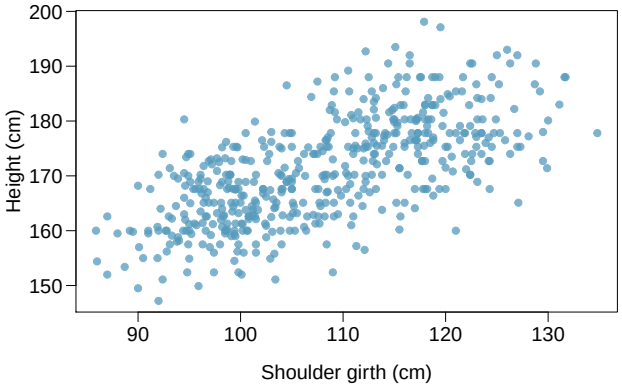
We want to fit a regression line to the data because we want to see the correlation between the calories and carbs. The slope is how corrolated they are

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

conditions: -linearity -nearly normal residuals -constant variability

The conditions are NOT MET, because the variability increases as the calories increase. There is not 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. 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.

Height and shoulde girth appear to have a linear relationship

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

The fundamental relationship would not change, but the line would have a higher slope, because the change in height would happen over a much smaller interval.

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.

```
mean_girth<-107.2
sd_girth<-10.37
mean_height<-171.14
sd_height<-9.41
cor<-.67

(m<-cor*(sd_height/sd_girth))
```

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

```
(y_intercept<-mean_height-m*mean_girth)
```

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

```
cat("the equation for the line is y=",m,"x+",y_intercept)
```

```
## the equation for the line is y= 0.6079749 x+ 105.9651
```

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

the slope is the amount of increase in height per unit increase in girth. The y intercept is theoretically what a person's height would be if their shoulder girth was zero... an impossibility. The mean squares regression line is not to be interpreted outside of the scope of the data, but the intercept is crucial for creating a linear equation.

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

```
cat("the R^2 is correlation^2, or",round(cor^2,2))
```

```
## the R^2 is correlation^2, or 0.45
```

The R^2 value is the percentage of the variability in height that is predicted by the girth of the shoulders.

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

```
student_girth<-100

student_height<-m*student_girth+y_intercept

cat("the predicted student height is ", round(student_height,2),"cm.")
```

```
## the predicted student height is  166.76 cm.
```

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

```
student_residual<- 160 - student_height

cat("the residual is ",round(student_residual,2),"cm.")
```

```
## the residual is  -6.76 cm.
```

This is the difference between the actual data and the predicted value of the data by the least squares line.

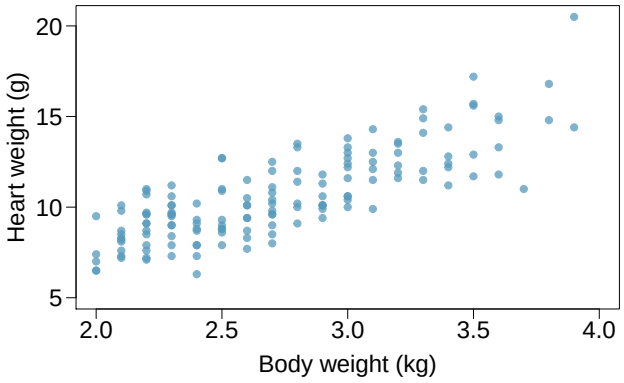
(f) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

No, because the shoulder girth is outside of the scope of our 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.

\begin{center} {

```
## numeric(0)
```



\end{center}

(a) Write out the linear model.

```
m<-4.034
y_int<--.357

cat("y=",m,"*x",y_int)
```

```
## y= 4.034 *x -0.357
```

(b) Interpret the intercept.

The intercept is below zero, which would seem to be impossible given a weight cannot be negative, but the y intercept is just a component of a linear equation, it is not meant as a data point.

(c) Interpret the slope.

for each unit increase in weight, the heart weight inceases by the value of the slope (d) Interpret R^2.

The R^2 is .64 which means that 64% of the variance in the weight of a cat's heart is determined by its body weight (e) Calculate the correlation coefficient.

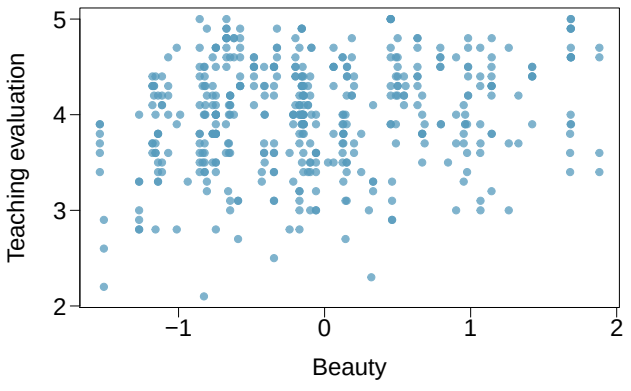
```
cor<-sqrt(.64)

cat("the correlation coefficient is the square root of r^2, or ",cor)
```

```
## the correlation coefficient is the square root of r^2, or  0.8
```

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.

```
## numeric(0)
```



\end{center}

(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.

I have the y intercept and the mean of y, so I know the delta of y over that interval. I also know that the y intercept is defined as the value of y when x=0, so i know the x values over that same interval, so I can calculate the slope

```
m<-(4.010-3.9983)/(0--.0883)
cat("the slope is ",round(m,2))
```

```
## the slope is  0.13
```

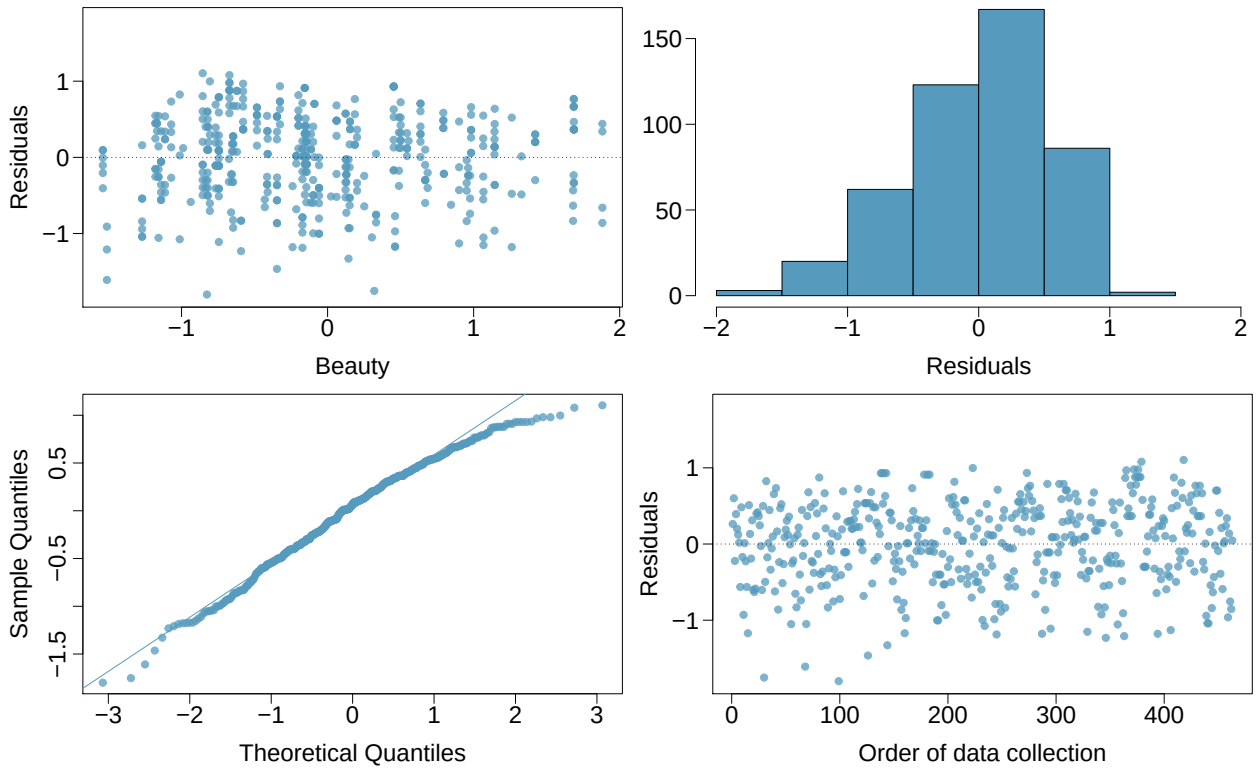
(b) Do these data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive? Explain your reasoning.

No, because the slope is so close to zero, outliers at high leverage points could easily change the slope to negative.

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

```
## numeric(0)
```

```
## numeric(0)
```



Judging by the plots provided, the data does meet the conditions for linear regression.

-linearity: the data looks roughly linear -nearly normal residuals: the residuals are around zero and have no pattern -constant variability: there is no trend in the variation and it is spread evenly across the data.