data<-read.table("http://www.yorku.ca/nuri/econ2500/bps6e/data/regression.txt",header=T)  
attach(data) 
names(data) 

y<-Fat  
x<-NEA  # nonexercise activity  

plot(x,y,pch=16)  
r<-cor(x,y)  
b<-r*(sd(y)/sd(x))  
a<-mean(y)-b*mean(x)  

b    # -0.003441487  
a    # 3.505123  

5.1 City mileage, highway mileage. We expect a car’s highway gas mileage to be related to its city gas mileage. Data for all 1040 vehicles in the government’s 2010 Fuel Economy Guide give the regression line  

\[
\text{highway mpg} = 6.554 + (1.016 \times \text{city mpg})
\]

for predicting highway mileage from city mileage.  

(a) What is the slope of this line? Say in words what the numerical value of the slope tells you.  

(b) What is the intercept? Explain why the value of the intercept is not statistically meaningful.  

(c) Find the predicted highway mileage for a car that gets 16 miles per gallon in the city. Do the same for a car with city mileage 28 mpg.  

(d) Draw a graph of the regression line for city mileages between 10 and 50 mpg. (Be sure to show the scales for the x and y axes.)
(a) Slope = 1.016. On average, highway mileage increases by 1.016 mpg for each additional 1 mpg change in city mileage.

(b) Intercept = 6.554 mpg. This is the highway mileage for a nonexistent car that gets 0 mpg in the city.

c) For a car that gets 16 mpg in the city, \(6.554 + (1.016)(16) = 22.81\) mpg. For a car that gets 28 mpg in the city, \(6.554 + (1.016)(28) = 35.002\) mpg.

(d)

```r
n <- 100
x <- sort(runif(n,10,60))
y <- 6.554+1.016*x + rnorm(n,sd=.25)
## Plot the fitted values from the regression model
model <- lm(y~x)
plot(x,fitted(model),
ylab="city mileage",
main="Direct Relationship (Y = 6.554 + 1.016 X)",
type="l",

```
5.17 To earn more, get married? Data show that men who are married, and also divorced or widowed men, earn quite a bit more than men the same age who have never been married. This does not mean that a man can raise his income by getting married, because men who have never been married are different from married men in many ways other than marital status. Suggest several lurking variables that might help explain the association between marital status and income.
Age is probably the most important lurking variable: married men would generally be older than single men, so they would have been in the workforce longer and therefore had more time to advance in their careers.

5.33 What’s my grade? In Professor Krugman’s economics course the correlation between the students’ total scores prior to the final examination and their final-examination scores is $r = 0.5$. The pre-exam totals for all students in the course have mean 280 and standard deviation 40. The final-exam scores have mean 75 and standard deviation 8. Professor Krugman has lost Julie’s final exam but knows that her total before the exam was 300. He decides to predict her final-exam score from her pre-exam total.

(a) What is the slope of the least-squares regression line of final-exam scores on pre-exam total scores in this course? What is the intercept?

(b) Use the regression line to predict Julie’s final-exam score.

(c) Julie doesn’t think this method accurately predicts how well she did on the final exam. Use $r^2$ to argue that her actual score could have been much higher (or much lower) than the predicted value.

Answer

The $x$-values will be pre-exam scores, and the $y$-values will be final exam scores. This is probably not the Princeton University, Nobel prize-winning economist Paul Krugman.

(a) $b = 0.5 \times \frac{8}{40} = 0.1$, and $a = 75 - (0.1)(280) = 47$. The regression equation is $\hat{y} = 47 + 0.1x$.

(b) $\hat{y} = 47 + (0.1)(300) = 77$.

(c) Julie is right. With a correlation of $r = 0.5$, $r^2 = (0.5)^2 = 0.25$, so the regression line accounts for only 25% of the variability in student final exam scores.

5.38 Our brains don’t like losses. Exercise 4.29 describes an experiment that showed a linear relationship between how sensitive people are to monetary losses (“behavioral loss aversion”) and activity in one part of their brains (“neural loss aversion”).
(a) Make a scatterplot with neural loss aversion as $x$ and behavioral loss aversion as $y$. One point is a high outlier in both the $x$ and $y$ directions.

(b) Find the least-squares line for predicting $y$ from $x$, *leaving out the outlier*, and add the line to your plot.

(c) The outlier lies very close to your regression line. Looking at the plot, you now expect that adding the outlier will increase the correlation but will have little effect on the least-squares line. Explain why.

(d) Find the correlation and the equation of the least-squares line with and without the outlier. Your results verify the expectations from (c).

Answer

(a) The outlier in the upper-right corner is circled because it is hard to see it with the two regression lines.

(b) With the outlier omitted, the regression line is $\hat{y}(\hat{y}^-) = 0.586 + 0.00891x$.

(c) The line does not change much because the outlier fits the pattern of the other points; $r$ changes because the scatter (relative to the line) is greater with the outlier removed, and
the outlier is located consistently with the linear pattern of the rest of the points.

(d) The correlation changes from 0.8486 (with all points) to 0.7015 (without the outlier). With all points included, the regression line is $\hat{y} = 0.585 + 0.0879x$ (nearly indistinguishable from the other regression line).

```
attach(data)
names(data)
y<-data$Behave
x<-data$Neural
x<-x[-16]
y<-y[-16]
lm(y~x)
```

Call:
```
  lm(formula = y ~ x)
```

Coefficients:
```
(Intercept)            x
 0.585811     0.008909
```

**5.44 Do artificial sweeteners cause weight gain?** People who use artificial sweeteners in place of sugar tend to be heavier than people who use sugar. Does this mean that artificial sweeteners cause weight gain? Give a more plausible explanation for this association.

**Answer**

In this case, there may be a causative effect, but in the direction opposite to the one suggested: People who are overweight are more likely to be on diets, and so choose artificial sweeteners over sugar. (Also, heavier people are at a higher risk to develop Type 2 diabetes; if they do, they are likely to switch to artificial sweeteners.)

**5.48 Some regression math.** Use the equation of the least-squares regression line (box on page 130) to show that the regression line for predicting $y$ from $x$ always passes through the point $(\bar{x}, \bar{y})$. That is, when $x = \bar{x}$, the equation gives $\hat{y} = \bar{y}$.

**EQUATION OF THE LEAST-SQUARES REGRESSION LINE**

We have data on an explanatory variable $x$ and a response variable $y$ for $n$ individuals. From the data, calculate the means $\bar{x}$ and $\bar{y}$ and the
standard deviations $s_x$ and $s_y$ of the two variables, and their correlation $r$. The least-squares regression line is the line 
\[ \hat{y} = a + bx \]
with **slope**

\[ b = \frac{s_y}{s_x} \]

and **intercept**

\[ a = \bar{y} - b\bar{x} \]

**Answer**

We have slope $b = r \frac{s_y}{s_x}$, and intercept $a = \bar{y} - b\bar{x}$, and $\hat{y} = a + bx$. Hence, when $x = \bar{x}$, 
\[ \hat{y} = a + b\bar{x} = (\bar{y} - b\bar{x}) + b\bar{x} = \bar{y} \]