**Regression, Correlation, and Scatterplots**

This laboratory is a bit different than the previous two. Today you’ll work through a warm-up analysis of our own class data, learning how to generate scatterplots, regression lines, and compute correlation. Then the Assignment will ask you to apply those skills to a dataset focused on the impact of poverty.

**Warm-up: Height and Speed**

On our Moodle page you can find a dataset called HeightSpeed.csv. This is a csv file, and since most people seem to prefer to use “file.choose”, I’ll mention that you can load this into R using

>HS=read.csv(file.choose(),header=TRUE)

You can verify that you have the correct file with reasonable data by entering

>head(HS)

Recall that the Height column records the reported heights in inches, and the Speed column is the fastest the individual has driven.

**a)** Produce a scatterplot where the y-variable (response) is “speed” and the x-variable (explanatory) is height. You can use:

>plot(HS$Height,HS$Speed)

You could also add in a title

>plot(HS$Height,HS$Speed, main='Relationship between Fastest Speed Driven \n and Height of the Individual')

Did you notice how I was able to split the title of the scatterplot into two lines?

**b)** In general, outliers should not be thrown out unless there’s a good reason, but there are several reasons why it may be legitimate to conduct an analysis without them. In class we noticed observations where individuals report being zero inches tall, and one person appears to not have much driving experience. As these observations are outside of the experiences of the large majority of respondents, we will ignore any observation that is less than 10 inches tall or where the individual has never driven 30 mph. To do this we need to create a new matrix with those observations removed.

>HS2=subset(HS,HS$Height>10 & HS$Speed>30)

We can then

>plot(HS2$Height,HS2$Speed)

You can ask R to give you the **l**inear **m**odel—the least squares regression line—using the command

>lm(HS2$Speed~HS2$Height)

R should return the slope and y-intercept of the least squares regression line. In order to add the linear model to the scatterplot you need to tell R the values of the y intercept (the **a** value) and slope (the **b** value) and it plots a **line**.

>abline(lm(HS2$Speed~HS2$Height))

**b)** A prediction error, also called a residual, is calculated as

“actual y-value” – “predicted y-value.”

Suppose that a person (not in the dataset) is 68 inches tall and has driven 80 mph. The linear model would predict a fastest speed of , which you can have R compute for you

>13.65+1.157\*68

You should get an answer of 92.326, hence the residual is -12.326. If you are going to make a number of these computations, it will be helpful to give a name to your regression model.

>MyLine=function(x){return(1.157\*x + 13.65)}

You can then compute the predicted speed for any given height, such as 68, using

>MyLine(68)

Question/Task: What is the prediction error for a person who is 64 inches tall and who reports having driven 85 mpg?

**c)** As we noted in class it appears that there is a positive association between an individual’s height and the fastest speed they have every driven. We were skeptical, though, and thought perhaps that sex (female/male) was a confounding variable. If you want to apply different colors to the different sexes, add in the “col” option differentiated by the sex variable.

>plot(HS2$Height,HS2$Speed, col=HS2$Sex)

Question/Task: Does it appear likely that sex is a confounding variable that provides an explanation for the positive association between an individual’s height and her or his fastest speed?

**d)** You can isolate the observations that are male and the observations that are female.

>HSM=subset(HS2,HS2$Sex=='Male')

>HSF=subset(HS2,HS2$Sex=='Female')

Using these subsets we can compute the correlation for the entire group, and for the males and females. First, the correlation for all the respondents is

>cor(HS2$Height,HS2$Speed, method="pearson")

And we can compute the correlation for our two subgroups

>cor(HSM$Height,HSM$Speed, method="pearson")

>cor(HSF$Height,HSF$Speed, method="pearson")

Question/Task: Explain why these three correlations indicate that sex is confounding the correlation between height and speed.

**e)** We can also create two plots side-by-side:

>par(mfrow=c(1,2))

>plot(HSM$Height,HSM$Speed,main='Male Drivers'))

>plot(HSF$Height,HSF$Speed,main='Female Drivers'))

Question/Task: On the basis of these two graphs, explain how we can see that sex is a confounding variable in this problem.

Note for Future Reference: If all that R gave you was the intercept and slope of your regression line, that wouldn’t be anything special. (You can get that with Excel.) There’s a lot more hidden beneath the surface, and you can use the **summary** command to access this hidden information.

First, let’s name our linear model

>LM=lm(HS2$Speed~HS2$Height)

and then ask for the summary

>summary(LM)

In response R should show you:

Call:

lm(formula = HS2$Speed ~ HS2$Height)

Residuals:

Min 1Q Median 3Q Max

-38.841 -8.879 -1.154 6.197 48.846

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) **13.6501** 34.3814 0.397 0.6930

HS2$Height **1.1568** 0.5061 2.285 0.0265 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

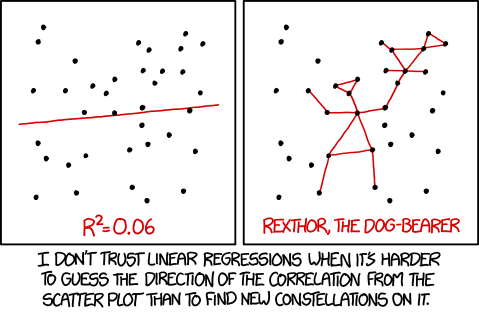
Residual standard error: 19.06 on 51 degrees of freedom

**Multiple R-squared: 0.0929**, Adjusted R-squared: 0.07512

F-statistic: 5.223 on 1 and 51 DF, p-value: 0.02647

While all of this information is important, most of it won’t be covered in this course. However a few items are of direct relevance to us, and some of you may encounter items in this summary in a class such as econometrics. For now, let’s focus on the most important values, which I’ve bolded for you.

First, note that the intercept and slope are present in this summary in the “Estimate” column. The second important piece of information is the “Multiple R-squared” value. This is the value that shows us how well our regression line describes our data. It is also the square of the correlation that you previously computed.



©XKCD

**Assignment! The Impact of Poverty** Name:

I have placed a dataset on poverty on our Moodle page. Load Poverty.R or Poverty.txt onto your computer, whichever format you prefer. It seems most people like to use the “file.choose” command, so you would enter either

>Poverty=load(file.choose())

for the .R file or

>Poverty=read.table(file.choose(), header=TRUE)

for the .txt file.

The dataset includes teenage mother birth rates and poverty rates for the 50 states of the U.S. and the District of Columbia. You can see the top entries in the file using

>head(Poverty)

The variable ***PovPct***is the percent of a state’s population in 2000 living in households with incomes below the federally defined poverty level. The variable ***Brth15to17***is the birth rate for females 15 to 17 years old in 2002, calculated as births per 1000 persons in this age group. ***Brth18to19*** is the birthrate for females aged 18 to 19; ***TeenBrth*** is the birth rate for all teenage females; and ***ViolCrime*** is the violent crime rate.

Your assignment is to provide short answers to the following ten questions. Ben and John will be happy to help you with your code, but we won’t verify the accuracy of your answers. You can collaborate with others in the class.

**1.** What is the five number summary for poverty rates? What is the five number summary for birth rates among 15 to 17 year olds?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Min | Q1 | Median | Q3 | Max |
| PovPct |  |  |  |  |  |
| Brth15to17 |  |  |  |  |  |

**2.** Plot ***Brth15to17***(as y-variable) versus ***PovPct*** (as x-variable). Is there a positive or negative association? Does the pattern appear to be linear?

**3.** What is the correlation between ***PovPct*** and ***Brth15to17***?

**4.**  Recall that you can access a single row of data using a command like

>Poverty[39,]

which will display the 39th row in the table. You can use the command

>Poverty[which.min(Poverty$PovPct), ]

to display the row the contains the minimal poverty rate. Which state has the minimal poverty rate, and what is that rate?

**5.** You can order your list by focusing on any chosen column. For example,

> Poverty[order(Poverty$PovPct),]

will list all fifty-one locations in increasing order of poverty rates. What are the three locations with the highest poverty rates and how do the 15 to 17 year old birth rates in these three locations compare to the 15 to 17 year old birth rates of the other states?

**6.** Determine the regression line for these data with ***Brth15to17***as the y-variable and ***PovPct*** as the x-variable. Write the equation and give a one sentence description of what this slope says about the relationship between ***PovPct*** and ***Brth15to17*** .

**7.** In Pennsylvania, ***Brth15to17*** = 17.2 and ***PovPct*** = 12.2. How does the actual 15 to 17 year old birth rate compare to what the regression equation would predict for a state with a poverty percentage = 12.2?

**8.** The variable ***Brth18to19*** is the rate of giving birth for females in the 18 and 19 year old age group. Plot ***Brth18to19*** versus ***PovPct***. Does the relationship in this plot appears to be weaker, stronger, or about the same strength as the relationship between ***Brth15to17*** and ***PovPct*** ? Verify your instinct (or rectify your instinct) by computing the correlation.

**9**. The variable ***ViolCrime*** is a measure of the rate of violent crimes. Plot ***ViolCrime*** versus ***PovPct***. There is an obvious potential outlier. Determine which location produces this observation. Would the crime rate be flagged as a potential outlier by the “1.5xIQR” guideline?

**10.**  Compute the regression line for these data with ***ViolCrime***as the y-variable and ***PovPct*** as the x-variable. What is the difference between the actual violent crime rate and what the regression equation would predict for the potential outlier?