



1

CASE $Q \rightarrow Q$ REGRESSION

Two Variables: Examining Relationships
Unit 1: Exploratory Data Analysis



UF UNIVERSITY of FLORIDA

{weblink:

http://content.bfwpub.com/webroot_pubcontent/Content/BCS_4/IPS7e/Student/Statistics/Applets/twovar.html}

Regression. Specifically simple linear regression, which is all we will cover in this course.

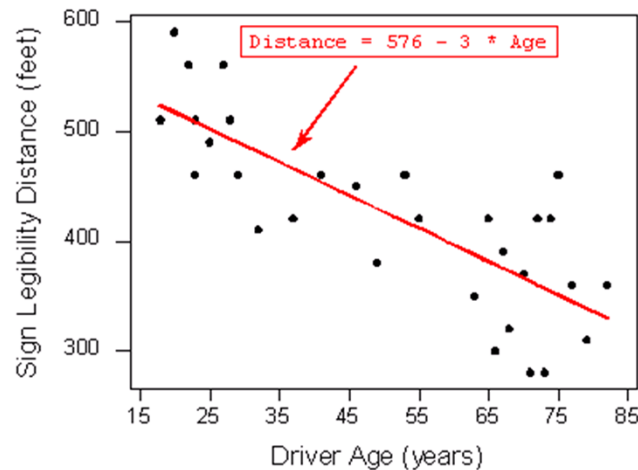
Right now, in exploratory data analysis, we simply want to understand how the regression line summarizes our data, how it can be used to make predications, and how to interpret the slope of the regression line in context.

When we discuss inferential statistics, we will come back and discuss simple linear regression in more detail.

Note that you will not be calculating the regression equation by hand in this course, you will be using software and interpreting the output.

Linear Regression

- Use “least squares” to find “line of best fit”



Here is our example for driver age and sign legibility distance.

Our regression equation is $\text{Distance} = 576 - 3 * \text{Age}$. The simple linear regression equation is found by determining the estimated slope and y-intercept (sometimes just called “the intercept”) that minimize the squared deviations from the regression line. For this reason, the line is also called the “least squares regression equation.”

The 576 is our y-intercept and the -3 is our slope.

The y-intercept is where the regression line crosses the y-axis, where x would equal zero.

The y-intercept is only interpretable if it both makes sense to talk about $x=0$ in your scenario AND $x=0$ is within the range of the data you measured.

Neither of these is true in this case.

If we were to attempt to interpret the y-intercept it would say “the average sign legibility distance for drivers who are zero years old, is 576 feet” which clearly is not a reasonable statement.

The slope is -3 and tells us: “for an increase in age of 1 year, the average sign legibility distance decreases by 3 feet.”

What if we wanted to talk about an increase in age of 5 years? Or 10 years?

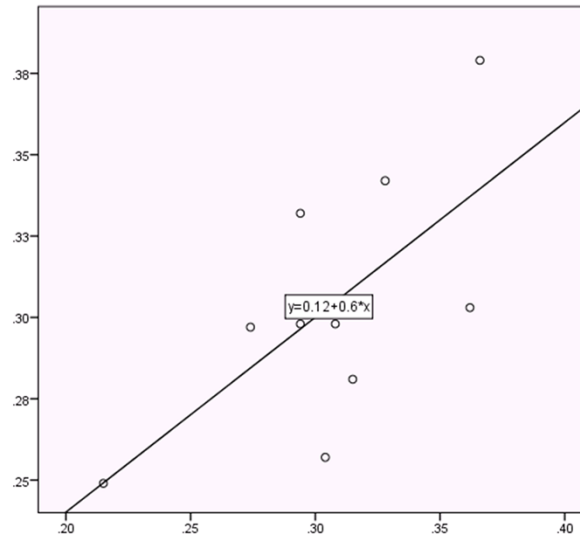
In that case, we can simply multiple the slope by 5 or 10.

So, “for an increase in age of 5 years, the average sign legibility distance decreases by 15 feet.”

And “for an increase in age of 10 years, the average sign legibility distance decreases by 30 feet.”

It can even be the case that we want some fraction of a whole unit.

Linear Regression: $Y = 0.12 + 0.6 X$



For example, if our x-variable is 1997 batting average and our y-variable is 1999 batting average, it would make not sense to interpret a 1-unit increase in batting average as it is constrained between 0 and 1 and usually less than 0.40 or 40%.

The slope does say “for a 1-unit increase in batting average in 1997, 1999 batting average increases by 0.6”. However, neither of these changes is possible in reality!

It would make much more sense to say, “for a 0.10 increase in 1997 batting average, 1999 batting average increases by 0.06” as this would be a possible comparison you could make between two players.

Any value recorded as a proportion or constrained to be a decimal would have this issue.

Linear Regression

- **In context**, slope (-3) means:
 - For an increase in age of 1 year, the maximum sign legibility distance decreases, **on average**, by 3 feet
- Predicted average distance for
 $\text{Age } 60 = 576 + (-3 * 60) = 396$
- Avoid extrapolating beyond your data range
- How do different kinds of outliers effect slope?
- How do different kinds of outliers effect intercept?



Back to our sign legibility example, let's review:

In context, the slope (-3) means: For an increase in age of 1 year, the maximum sign legibility distance decreases, **on average**, by 3 feet

The "on average" is important as our regression equation is predicting the AVERAGE y-value for a given x-value, not an individual y-value, which we expect to vary around the line.

We can use the regression equation to predict the average sign legibility distance. For example, for drivers of age 60, the mean or average sign legibility distance is predicted to be around 396 feet.

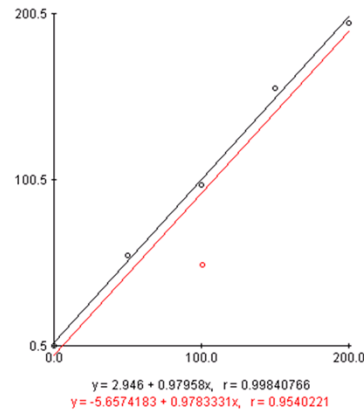
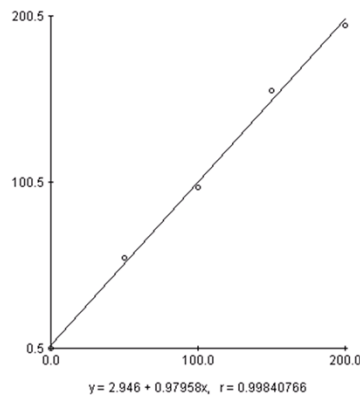
A few additional comments:

We should avoid extrapolating beyond the range of our data, in other words, you should predict only for x-values that are in the range of the data you measured.

It may be ok to extrapolate slightly beyond your data range but the farther away from your observed x-value range, the less reliable the prediction will be.

Now let's look at how outliers alter the slope.

Regression and Outliers



[Applet](#)

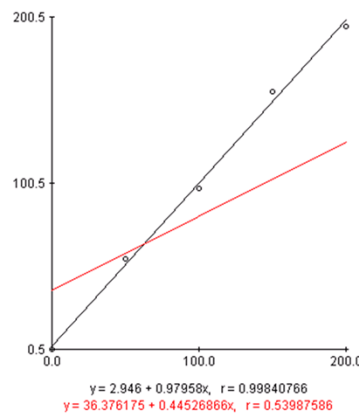
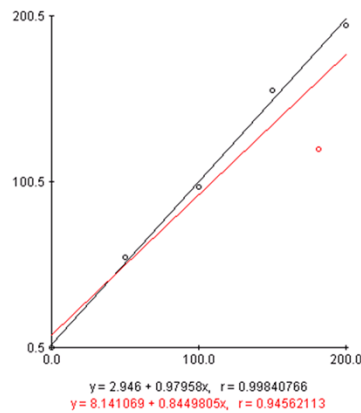
UF UNIVERSITY of FLORIDA

{weblink: <http://www.stat.sc.edu/~west/javahtml/Regression.html>}

All outliers will pull the regression line toward them.

You can “balance” the outlier so that it doesn’t change the slope but shifts the line up or down. In this case the outlier has shifted the line down resulting in a smaller y-intercept but basically the same value for the slope.

Regression and Outliers



[Applet](#)

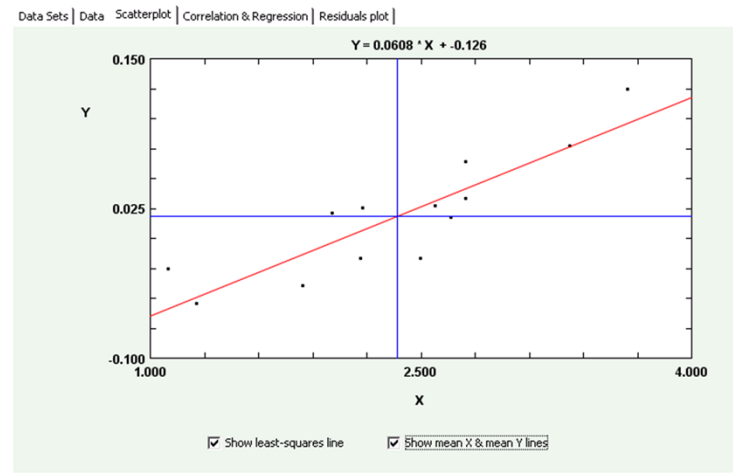
UF UNIVERSITY of FLORIDA

{weblink: <http://www.stat.sc.edu/~west/javahtml/Regression.html>}

Here, we see that the line is pulled toward the outliers in such a way as to rotate the line. In both cases, the y-intercept would increase and the slope would decrease, just to a greater degree the farther away from the line the point is moved.

If we placed the point on the other side, we would see the opposite effect. Try it for yourself! The link is provided in the course materials as well as the transcript for this video.

Case Q-Q Applet – Scatterplot




Returning again to the data from the two-variable calculator applet with x = social distress and y = brain activity.

The regression equation is $Y = -0.126 + 0.0608(x)$.

Thus for each 1 unit increase in social distress (which is plausible based upon the data), on average, brain activity increases by 0.0608.


That may not seem like much but notice the units of brain activity range from a little above zero to about 0.15.



CASE $Q \rightarrow Q$ REGRESSION

Two Variables: Examining Relationships

Unit 1: Exploratory Data Analysis



In summary, for a linear relationship between two quantitative variables, we can summarize the relationship with the linear regression equation also known as the “least squares regression line or equation.”

Using the slope of this equation, we can estimate, on average, how much the y-variable changes for each 1-unit increase in the x-variable and make predictions about the average y-value for a specific value of x.

Remember whenever you interpret the slope to be specific about the variables under study and the units of measurement of the variables.