Bivariate Quantitative Data

So—we've done quite a bit with univariate data (quantitative and qualitative), and we've also looked at bivariate qualitative data (Chi-Square for Independence). Now it's time to look at bivariate quantitative data.

Exploring the Relationship

As was the case with univariate quantitative data, we begin by looking at the distribution graphically. Since we've now got two variables, our old methods of visualization won't work.

A New Graphic Display

Enter the scatterplot. Plot one variable horizontally, and the other vertically. Each measurement pair becomes a coordinate pair in the plot.

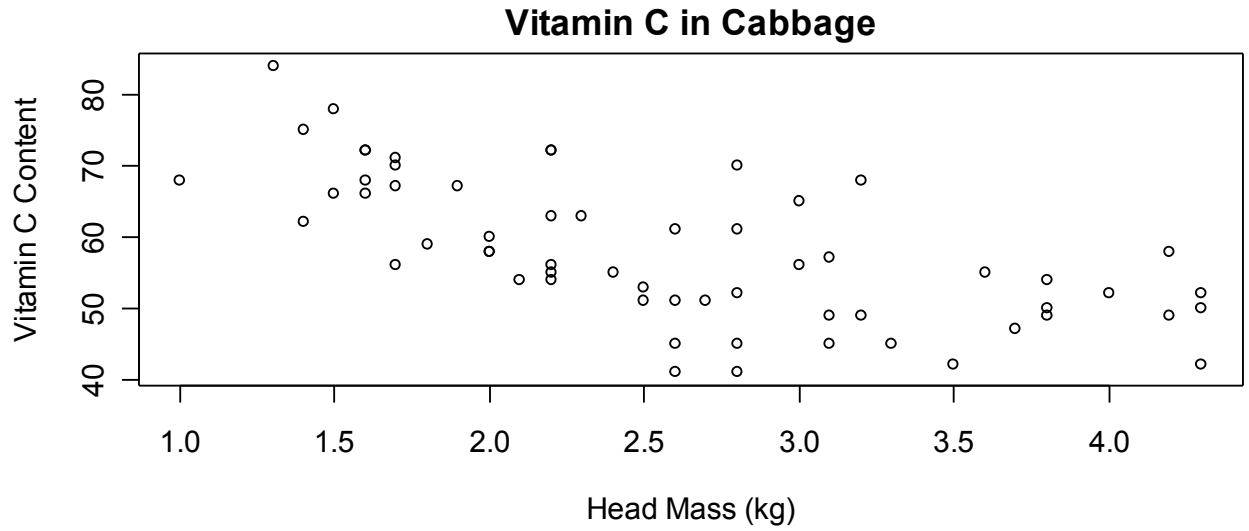

Figure 1 - A Scatterplot

This is "a plot of Vitamin C content versus Head Mass." Note that the *y*-variable comes first.

Deciding which variable to plot horizontally is important. Once again, if we were the ones actually doing the work, we'd know—but we as students will have to work harder to figure it out.

Every bivariate study includes an **explanatory variable** (*independent*) and a **response variable** (*dependent*). We are going to look and see if changes in one of these (the explanatory) causes a change in the other (the response). So, in our mind—and only in our mind—do we wonder "does a change in *this* variable cause a change in *that* variable?" Or—and it is OK to say this one aloud—"can I predict the value of *that* variable from *this* variable?"

The variable that you predict is the response. The one that you use as a basis for the prediction is the explanatory variable.

Visual Items of Interest

With univariate data, we looked for center, shape and spread. Not so for bivariate data…

Linearity/Association

The primary quality for which we will look is **linearity**—that's the simplest type of function that can relate two variables. Naturally, we don't expect the points to fall exactly on a line—the line is going to be fuzzy. The cabbage example above is fairly fuzzy, but generally linear. Also notice that the "line" seems to be falling—it has a negative slope. We say that the data have **negative association**.

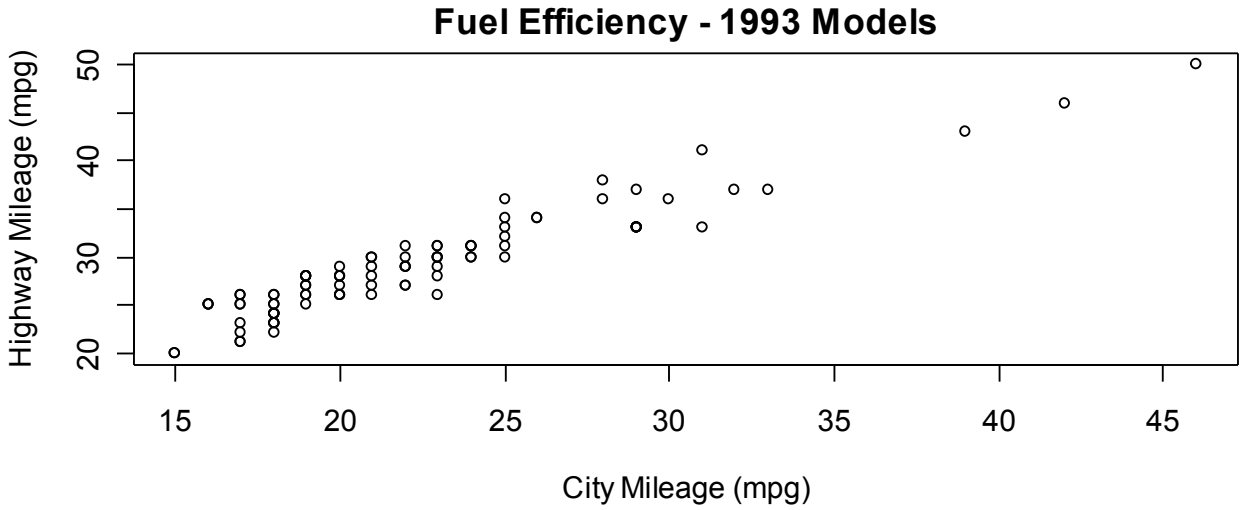

Figure 2 - Car Mileage Scatterplot

The scatterplot above is also generally linear, with **positive association**.

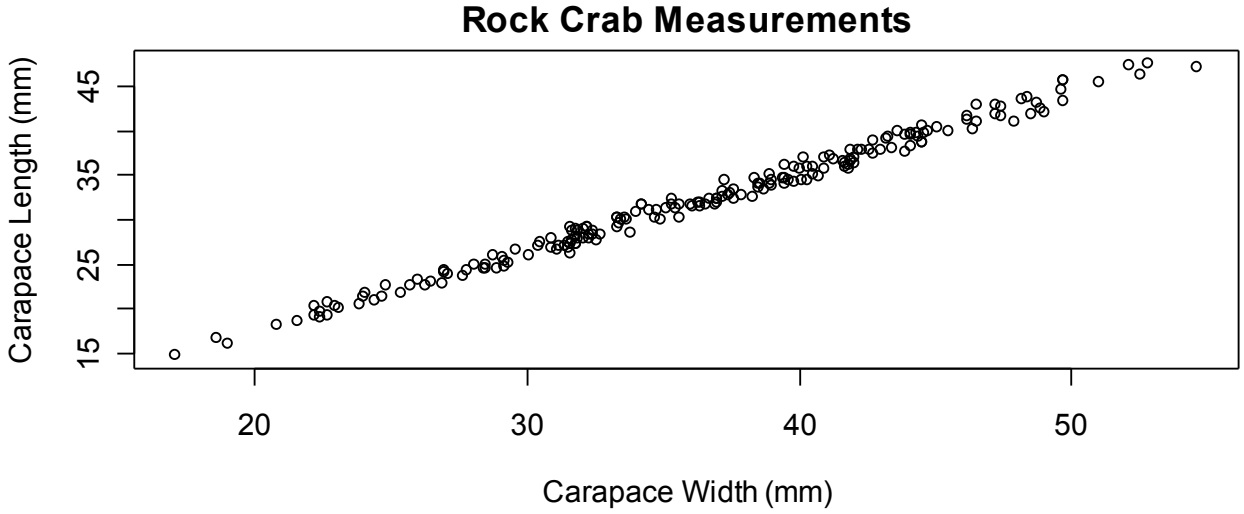

Figure 3 - Rock Crab Scatterplot

The scatterplot above is very linear with positive association.

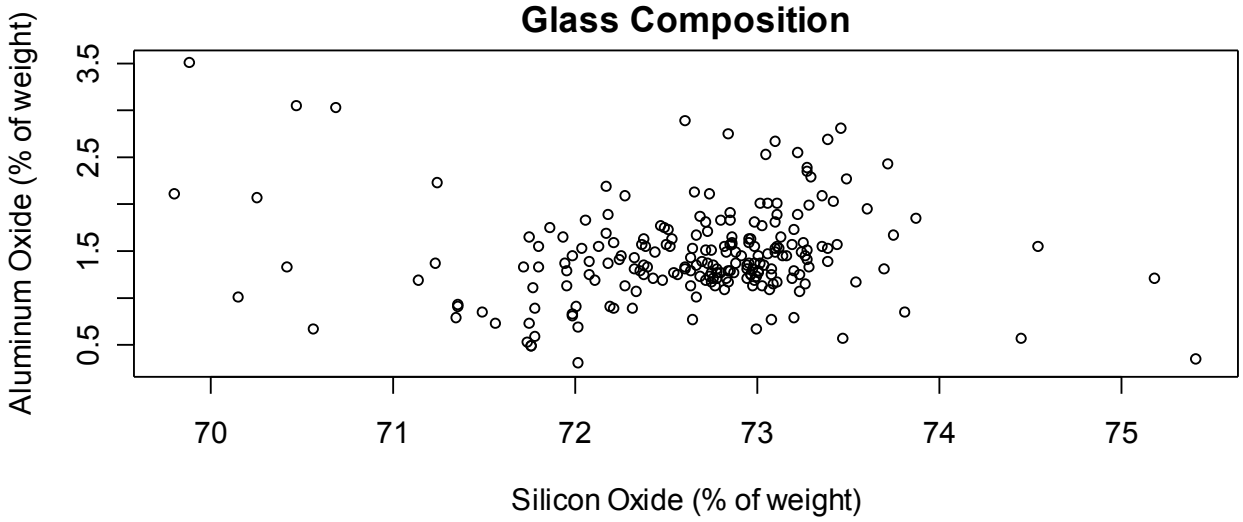

Figure 4 - Glass Fragment Scatterplot

The scatterplot above is not very linear—it has almost no association.

Clusters/Gaps

Clusters of points within the plot can indicate the presence of another variable. Gaps are regions (values) of the explanatory variable that have no associated response measurements.

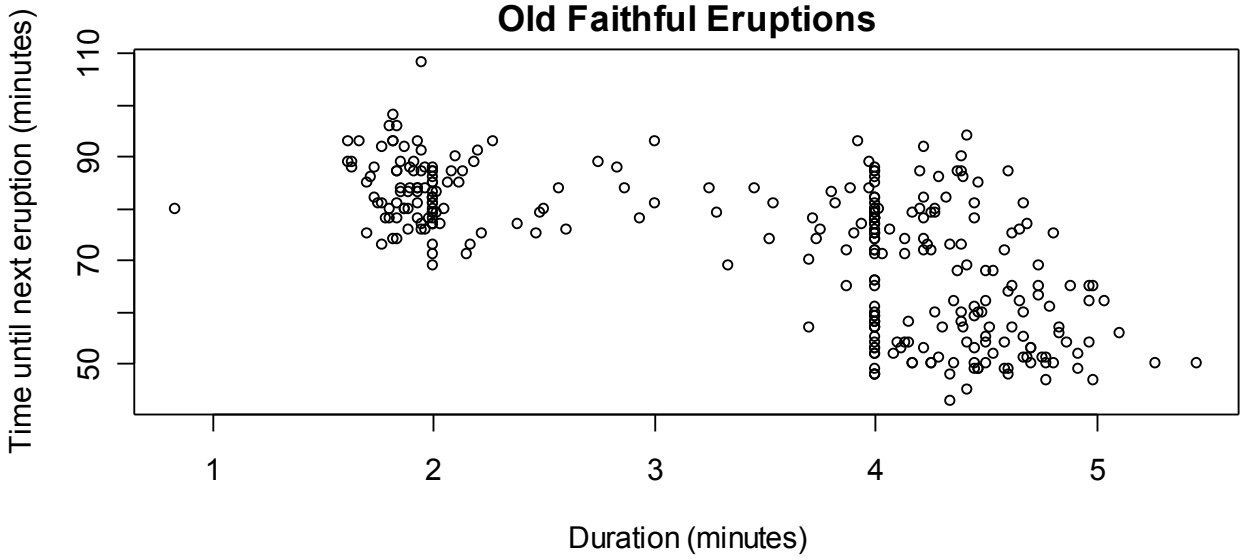

Figure 5 - Old Faithful Scatterplot

The scatterplot above shows two clear clusters—one near 2 minutes; the other between 4 – 5 minutes.

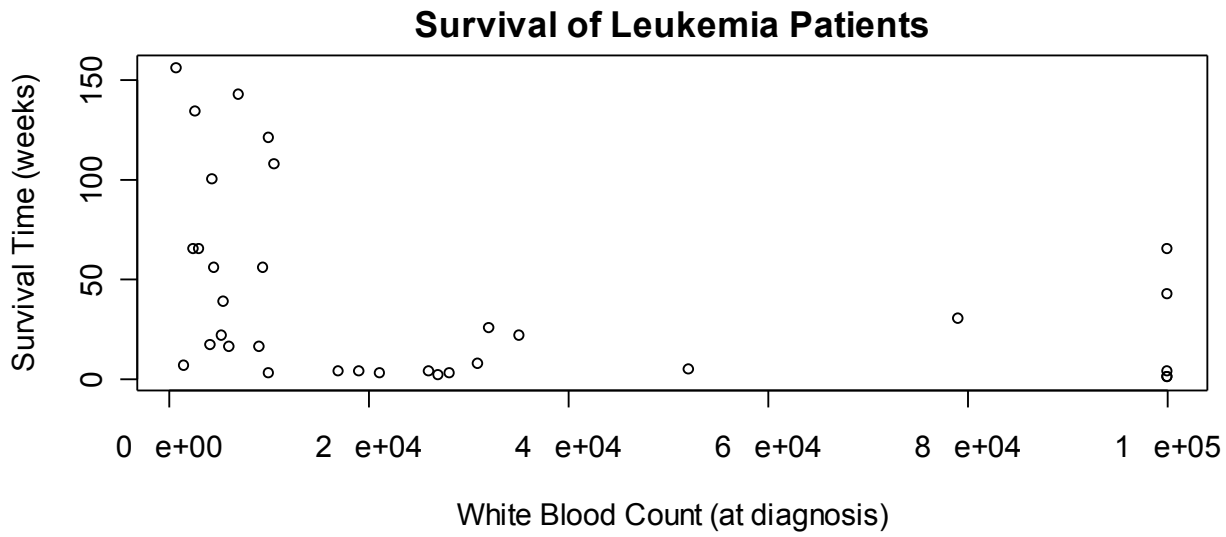

Figure 6 - Leukemia Scatterplot

The scatterplot above shows a gap between 600,00 and 80,000 white blood cells (and probably another between 80,000 and 100,000).

Outliers

Any point which does not seem to be in accord with the others is an **outlier**. There is no numerical rule for determining if a point is an outlier.

From the Old Faithful example—the eruption that lasted less than a minute appears to be an outlier.

Measuring the Strength of a Linear Relationship

Not all linear relationships are equal. Some a very nearly perfect lines—look again at the Crab example. Others are hardly linear at all—look at the Glass or Leukemia examples.

This is quite subjective, though—it would be nice to firm up these observations with some numbers…

Pearson's Product-Moment Correlation Coefficient

Yikes! There's a name destined to send 'em screaming…and something of a misnomer, too. It was Sir Francis Galton who first developed **Correlation** as a method of measuring association between variables. Karl Pearson took the idea and extended it.

Anyway—here's the formula.

Equation 1 - The Correlation Coefficient

$$r = \frac{\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n-1}$$

Also known as the **Correlation Coefficient**, it measures the *strength* and *direction* of a *linear* relationship. Facts:

[a] r has a value between -1 and 1 (inclusive). The farther the value is from zero, the closer the points are to falling on a line. The nearer to zero, the less the points look like a line (more like a blob). This is the strength.

[b] the sign of r is also the type of association the data exhibit (positive or negative). This is the direction.

[c] r has no units, and is unaffected by changing units in the original data. Notice in the formula that the numerator looks a lot like the standardizing formula—in fact, that's exactly what it is. Standardized scores are the same no matter what transformations are applied to the data; so r stays the same, also!

[d] I suppose this is more of a reminder—r is only useful (appropriate) for linear relationships. If the scatterplot does not appear linear, then don't compute a value for r!

The Coefficient of Determination/Variation

Another measure of association. Its symbol is r^2. You find it by…you guessed it! You square the value of r. Thus, r^2 can have values between 0 and 1.

r and r^2 were developed separately as two different ways of measuring association—there are ways to get r^2 without first having r—but, as it turns out, they are related. As for its meaning—tune in later.

Example

[1.] Tax efficiency is a measure of the tax burden of various investment schemes. An investigation of tax efficiencies of several mutual funds resulted in the following data:

Table 1 - Investments and Taxes

Investment in Energy (%)	Tax Efficiency
3.1	98.1
3.2	94.7
3.7	92.0
4.3	89.8
4.0	87.5
5.5	85.0
6.7	82.0
7.4	77.8
7.4	72.1
10.6	53.5

Describe the relationship between the variables.

I don't suppose it would be too weird to suggest that a scatterplot might be in order?

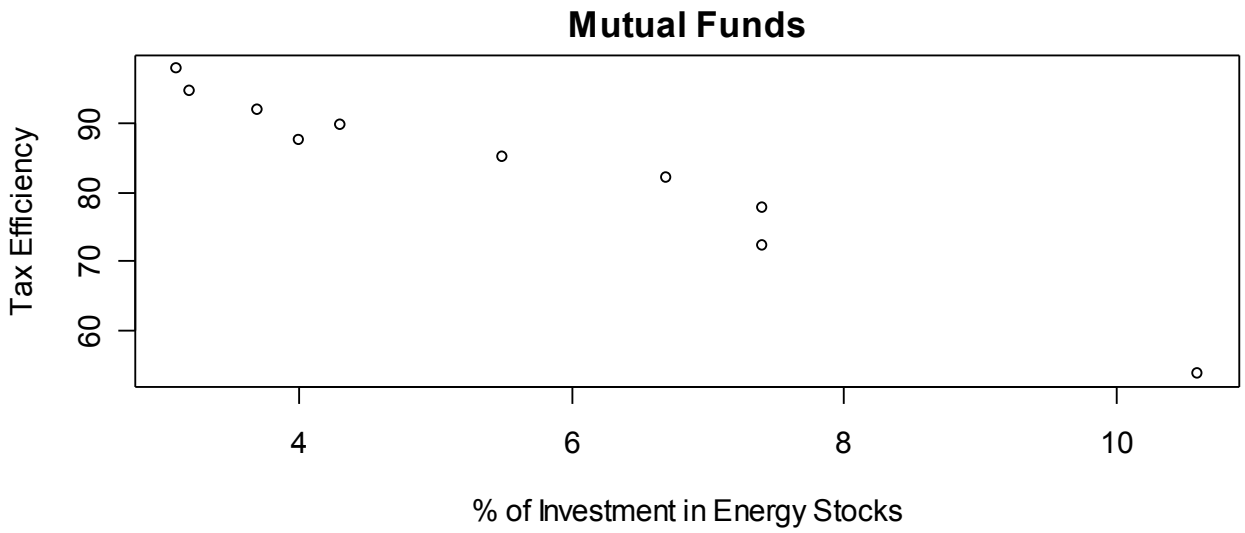

Figure 7 - Scatterplot for Example 1

There appears to be a linear relationship—fairly strong—with negative association. There seems to be a gap in the 8-10 range, and an outlier at the end.

The correlation coefficient is -0.9748, which confirms my observation of strong, negative, linear association. $r^2 = 0.9503$, which tells me that over 95% of the variation in tax efficiency can be explained by the least squares regression of tax efficiency on percent of investments in energy.

Establishing a Model

First of all, a quote: "All models are wrong. Some models are useful." (George P.E. Box; b. 1919 and still alive as of 2004)

The Idea

There are lots of ways to model the relationships between variables. It is important that you not think that what we do is *the* way. There are many paths to the summit…

We are concerned with linear relationships; the simplest type of function that we can use. So, we'll be trying to fit a line to the data. As we've seen, it's quite rare that all the data actually fall on a line. There's going to be some error in any model that we construct.

Let's consider the example of Acorn Sizes and Tree Heights for some Oak species…

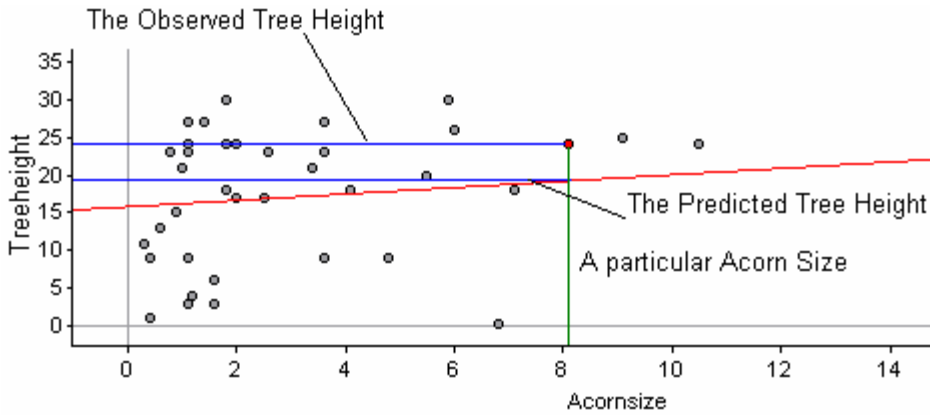

Figure 8 - Illustration of a Residual

Once we find some line, then for every point, there's a difference between the observed response (the top blue line) and the response predicted by our model (the bottom blue line). The difference between these is called the **residual** for that point. Specifically, a residual is the observed response minus the predicted response. You can think of this as *error* in the model. Of course, error is bad. We'd want to find a line that minimizes the total error of the model.

The Method of Least Squares

The **Method of Least Squares** (or Least *Squares Regression*) finds a line that minimizes the sum of the squares of the residuals. Some people refer to it as the "line of best fit," but I don't think that's a good way to look at it / talk about it.

Note that this line will pass through the point with coordinates $(\bar{x}, \bar{y})$.

The slope of the LSR line is $r\frac{s_y}{s_x}$. Notice that this makes sense—slope is *y* over *x*, isn't it? The correlation modifies that, making it flatter when the correlation is close to zero.

The intercept of the line can now be found—if the line is of the form $y = mx + b$, and we know a point and the slope, then we can solve for the intercept! In particular, it is at $\bar{y} - r\frac{s_y}{s_x} \cdot \bar{x}$.

The form of the equation of a line needs addressing here. You will probably have an understandable fetish for the old $y = mx + b$ form. However, this is not the only way. What can't it be $y = b + mx$? Why do we have to use *m* and *b*? Other people don't…in particular, statisticians don't. In fact, they prefer $y = b_0 + b_1x$. We're not going to go that far—although, if you look at the formula sheet that is provided on the AP exam, you'll notice that exact form is on there!

No; we're going to use the form that is also widely accepted, and (perhaps most important to you) the form that is used in the textbook. In particular, we'll use $y = a + bx$.

Well, almost that. You see, the variable *y* is reserved for the observed response. By writing $y = a + bx$, we're saying that our model hits all the points, and has no error. Of course, that's nonsense. We need to adjust the notation to reflect the fact that the model produces a predicted response. Here's how we do that: $\hat{y} = a + bx$. The "hat" is read "predicted," and is vitally important! Don't forget it!

Example

[2.] Back to the tax example—the least squares regression line is $\hat{y} = 112.6759 - 5.2640x$, where x is percentage of the investment in energy stocks, and $\hat{y}$ is the predicted tax efficiency.

Model Diagnostics

The Fit of the Line

Once you've computed the equation of the least squares line, then plot it along with the data to check for fit. The data points should surround and envelop the line—about half above and half below.

More on r^2

r^2 measures the proportion of variation in the response variable that can be explained by the least squares regression of (the linear relationship between) the explanatory and response variables.

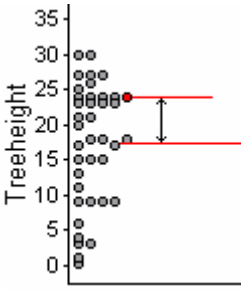

Figure 9 - Variation in the Response Variable

Here we have some data on the heights of some oak tree species. The indicated distance is the deviation for one species—the difference between its height and the mean height. This is just one part of the variation in (what will momentarily become) the response variable. Now, let's add an explanatory variable.

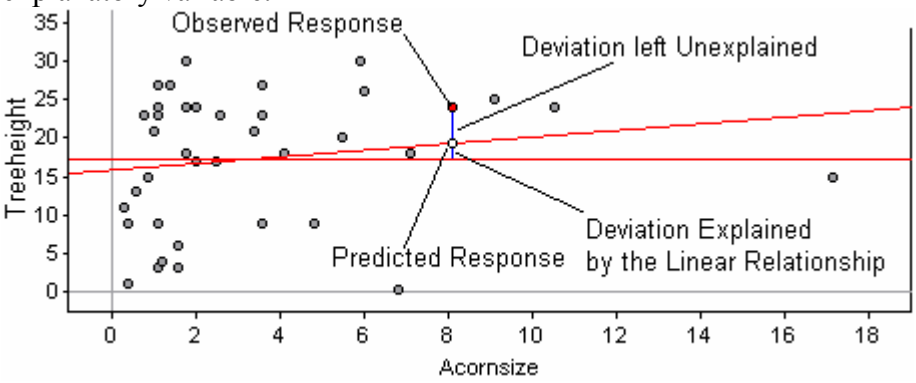

Figure 10 - Explained and Unexplained Variation

So here we have the scatterplot with two lines—the horizontal line at the mean tree height (mean response), and an oblique line, which is the least squares regression line. Our selected species still has a deviation, but now we can explain some of the deviation. In particular, we can

explain the deviation between the predicted height (the LSR line) and the mean height. What we can't explain is the remaining deviation—from the LSR line to the observed response.

r^2 measures the proportion of all deviation (the sum of the deviations) that can be accounted for by the least squares regression line.

Perhaps you should memorize the sentence**: r^2 measures the proportion of variation in the response variable that can be explained by the least squares regression of (the linear relationship between) the explanatory and response variables.**

Influential Points

We've already discussed outliers—data that don't seem to belong. But some outliers are more important than others—they have more influence over the position of the least squares regression line. An **influential point** is one that, when removed, causes a significant change in the position and/or direction of the regression line.

Typically, a point that is an outlier in the *x*-direction will exert influence on the line. Here are a few words on the matter from Dave Bock, a High School teacher and textbook author:

"Points tug at the regression line in an attempt to make their residuals smaller. But the regression line pivots around the mean-mean point. Points close to that fulcrum (left to right) can't make their residuals much smaller, hence they do not change the slope of the line much. Points far away (in the *x*-direction) can exert a lot of leverage - changes in the slope can make their residuals much smaller."

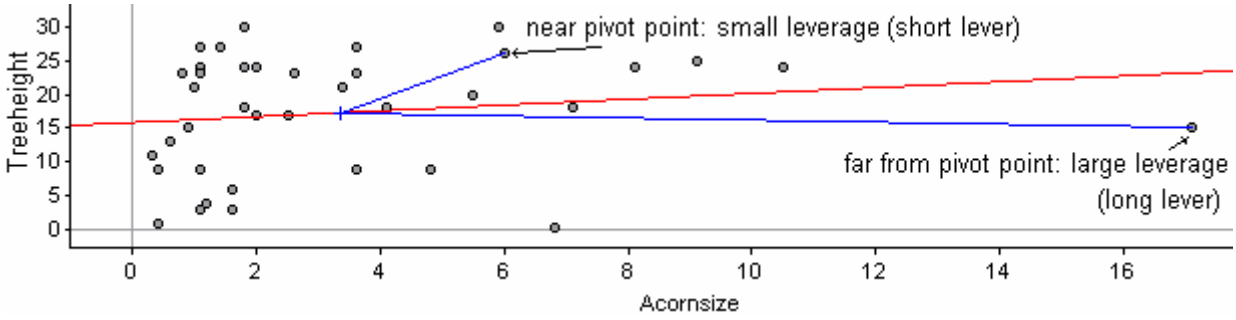

Figure 11 - Influential Points

Analyzing the Residuals

Analyzing the residuals—the error in the model—can tell us a lot about whether or not our model is useful. There are many thing to look for.

Random Distribution about the Line

First of all, there should be no pattern to the error—to the residuals. If there's a pattern to the residuals, then we can predict the error, and we could construct a better model!

To check for random distribution, we use a **residual plot**—a plot of the residuals against the predicted response. Some textbooks define a residual plot as residuals against explanatory variable—which is fine, if you only want to do simple linear regression (the type we're doing). However, if you plan to (have to?) take any more statistics in your life, you'll almost certainly deal with multiple linear regression, where there can be many explanatory variables, but only one response.

So—plot the residuals against the predicted response, and check to see that there are no patterns.

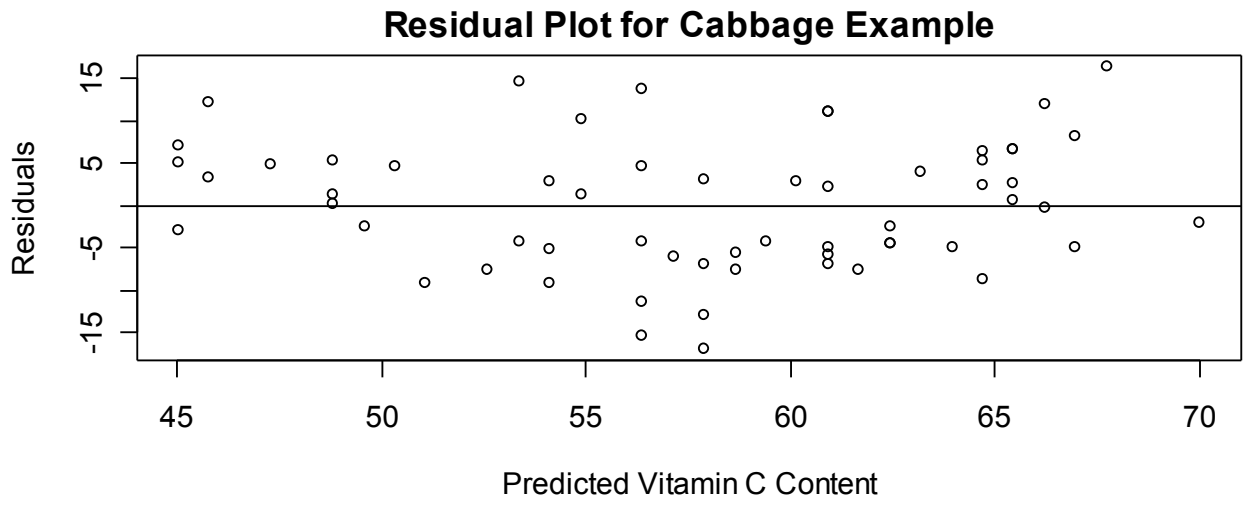

Figure 12 - Residual Plot for Cabbage Data

Here is a good residual plot—there is no apparent pattern.

Here is the residual plot from a study of specific humidity against temperature:

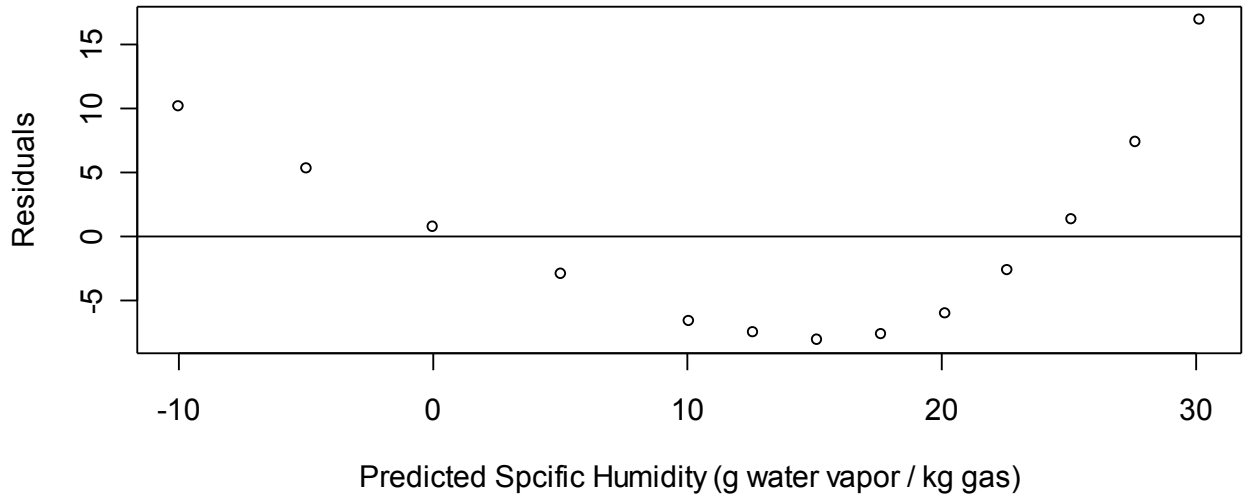

Figure 13 - A curved residual plot

This is a bad residual plot—it has a clear pattern to it (a curve).

A curve in the residual plot almost always means that the original data have a curved relationship. Sometimes, the curve is so slight that you can't see it in the original scatterplot. If the residuals are sufficiently small, then you can probably still use the linear model to make predictions—for us, though, if we see a curve, we'll go and look for a different model.

Here's another bad residual plot:

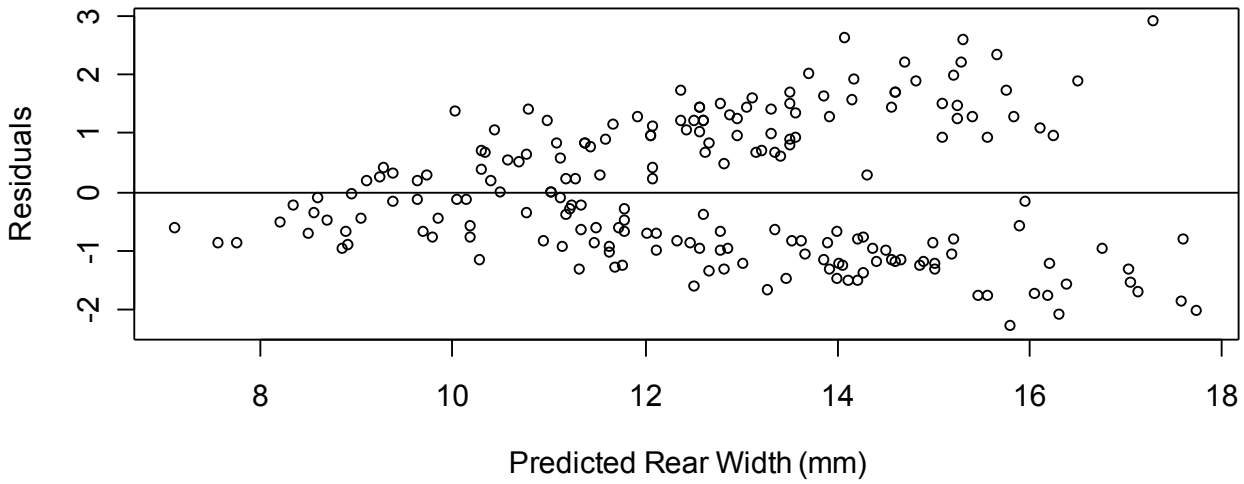

Figure 14 - Non-constant variation

This is bad, because it indicates…

Constant Variation about the Line

…that the size of the error depends on the size of the prediction. In other words, the error is smaller on one end of the regression line than on the other end. This is bad; if we've got a good, useful model, then the variation about the line (the size of the error) should be constant.

The "cone" or "fan" shape in the previous residual plot indicates a non-constant variation. It would be just as bad (but really unusual) to see a "dumbbell" shape).

Out in reality, there are ways to transform the data to alleviate this problem; for us, it's the end of the road.

Normality

For reasons that are difficult to explain (I do have a 31 page document that attempts to explain it!), if error is random (unpredictable), then the errors will have a normal distribution.

The easiest way for us to check this is with some graph (a histogram, probably) of the residuals.

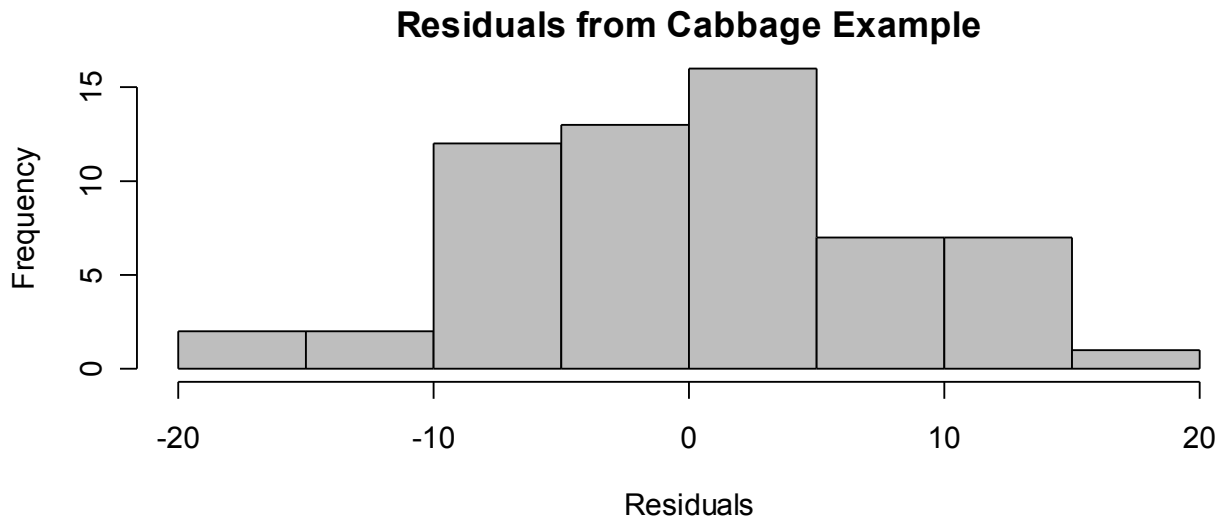

Figure 15 - Histogram of Residuals for Cabbage Data

You'll be looking for any evidence of non-normality. It has to be pretty clear—here's a histogram of a random sample from a normal distribution.

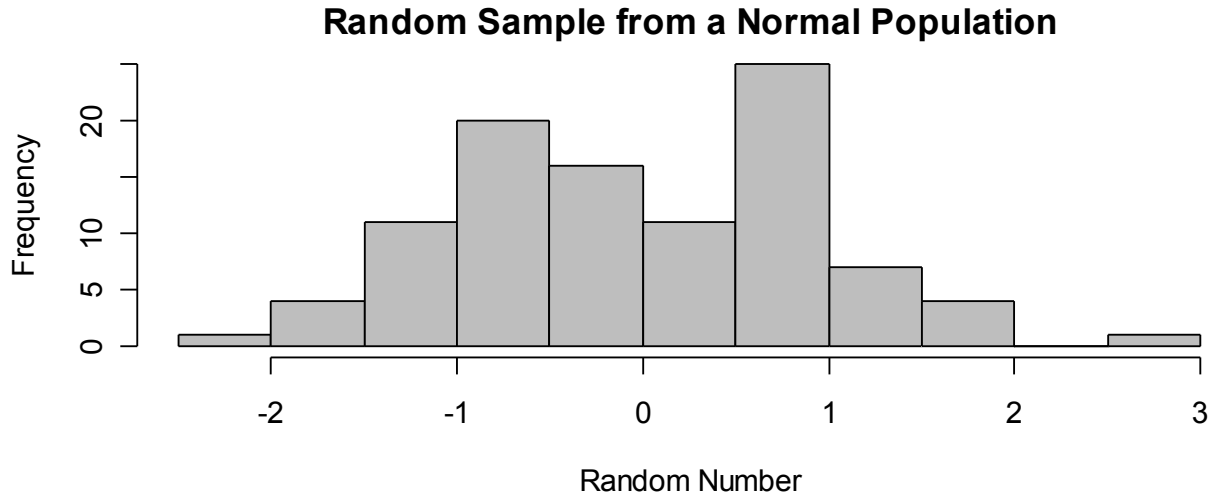

Figure 16 - Histogram of Normal Data

That doesn't look all that normal—but there's not a clear NON-NORMAL shape there.

There is, naturally, another way. It is called a normal probability plot (or a normal quantile plot; or q-q plot). This plots the predicted *z*-score against the actual *z*-score (or vice-versa; the TI calculators actually let you choose this) for each datum. If the distribution is normal, then these values ought to be the same for each datum, and you ought to get a straight line.

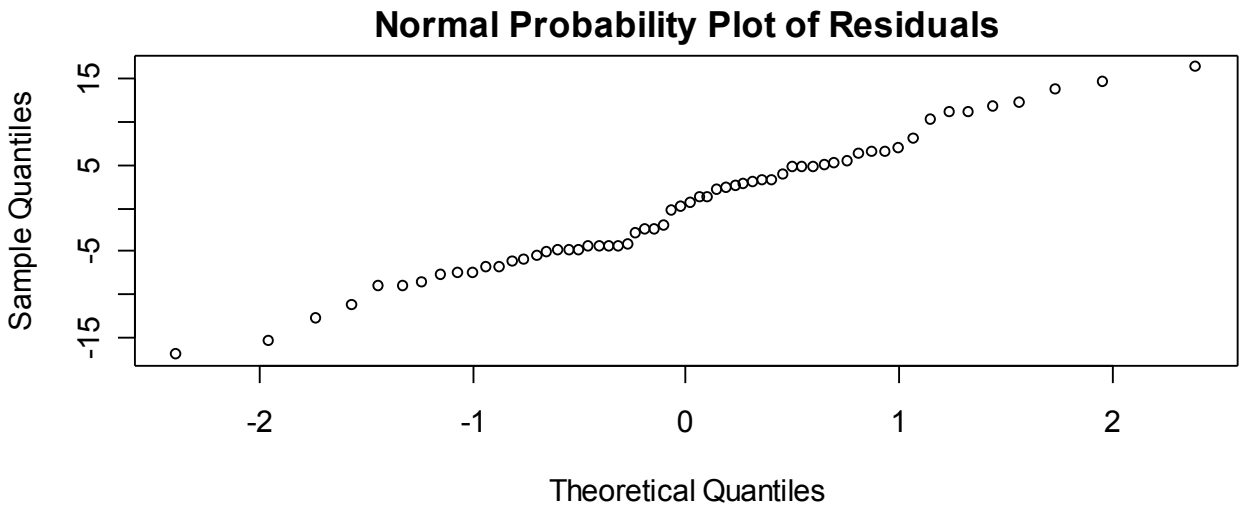

Figure 17 - Normal Probability Plot

You're looking for a straight line…again, really we're looking for anything that is clearly non-normal. Here's the NPP version of the most recent histogram above:

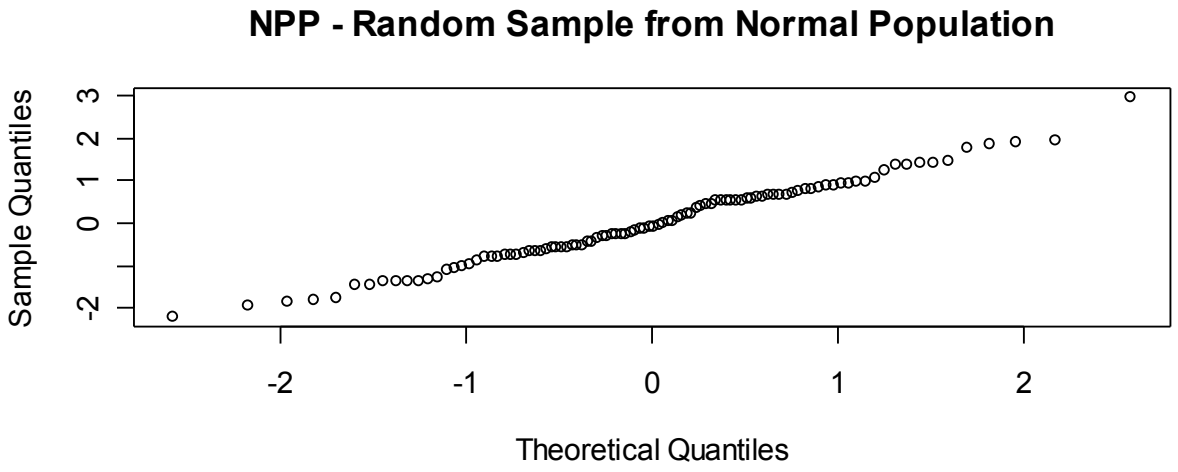

Figure 18 - Normal Probability Plot for Normal Data

Example

[3.] Back to the tax example…

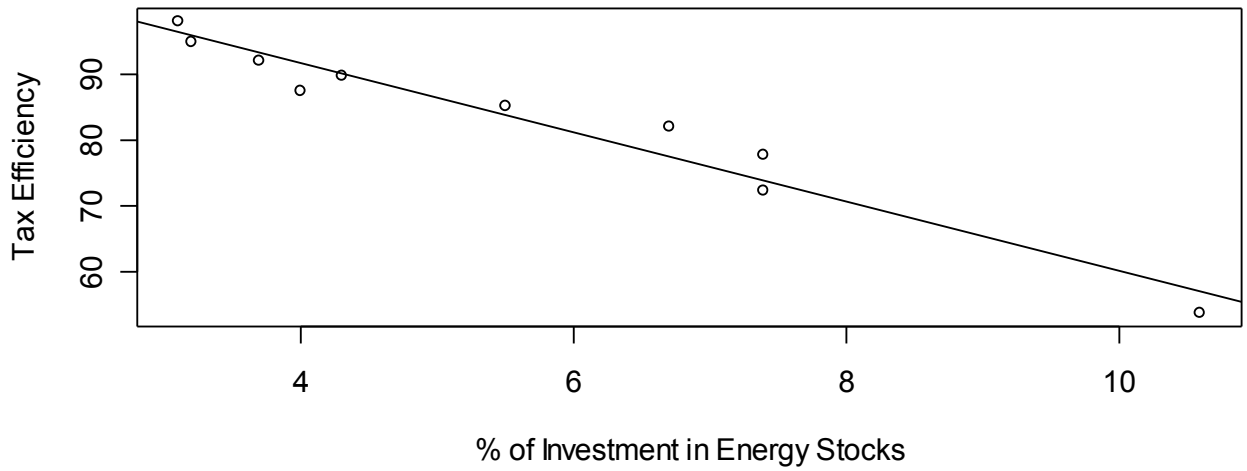

Figure 19 - Least Squares Line for Tax Data

The LSR line seems to fit well. There do not appear to be any obvious influential points.

Here's the residual plot.

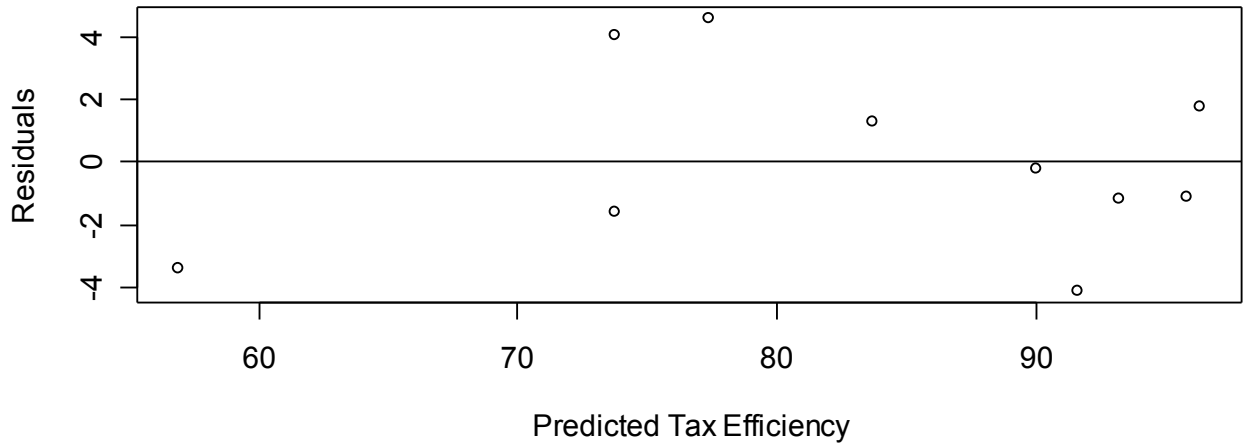

Figure 20 - Residual Plot for Tax Data

Looks OK—no discernable pattern.

Finally, checking the normality of the residuals:

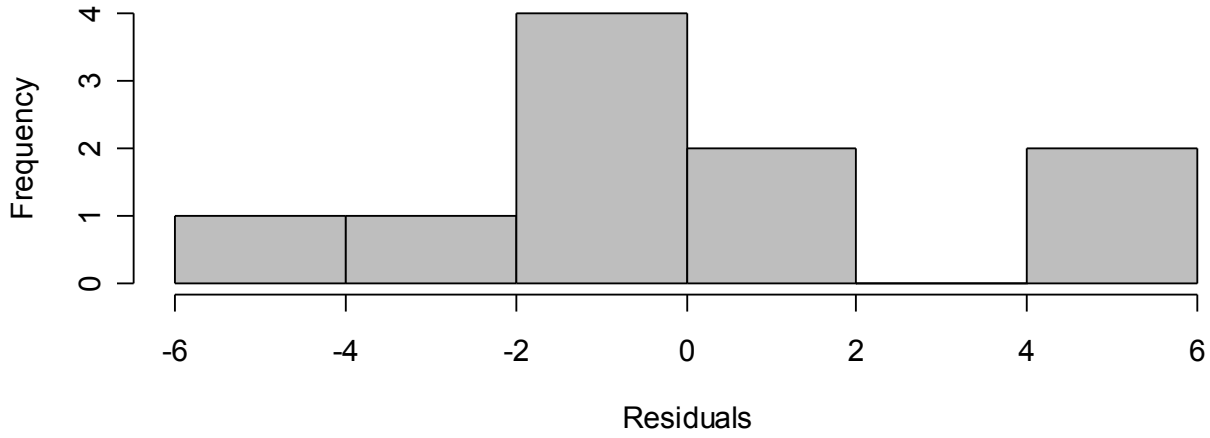

Figure 21 - Histogram of Residuals for Tax Data

The residuals do not appear to be non-normal. Here's the Normal Probability Plot for comparison:

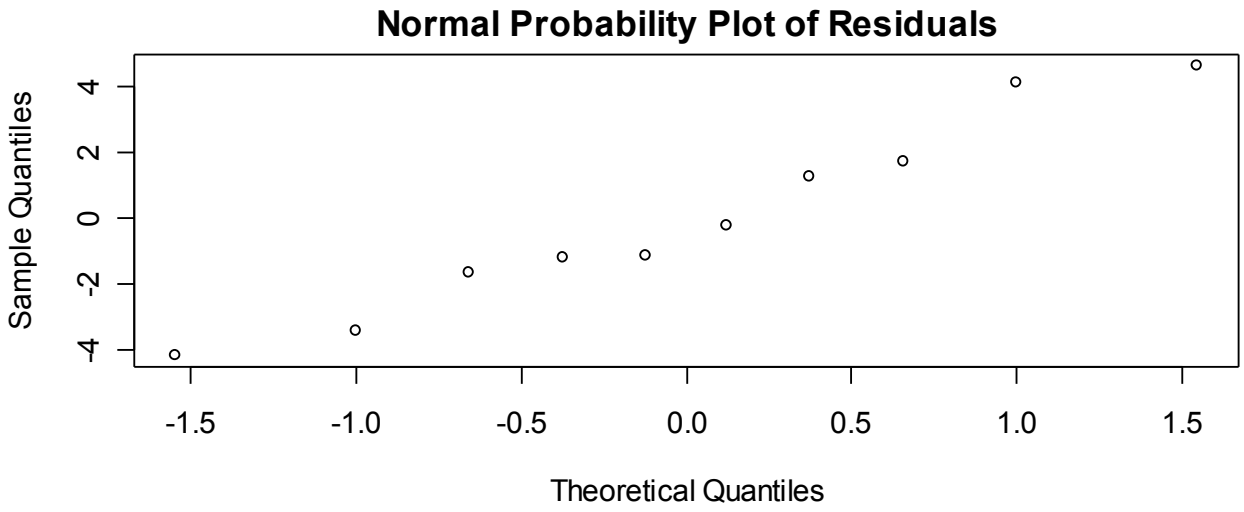

Figure 22 - Normal Probability Plot for Tax Data