2/7/2023

Testing Model Assumptions

So, we've built our model. The null hypothesis was rejected: the model is better than none, but does that mean it's an appropriate model? So, we want to test our model assumptions to see if they hold. We want to look for any potential outliers, perhaps, and begin discussing potential ways of repairing our models to resolve some of these problems.

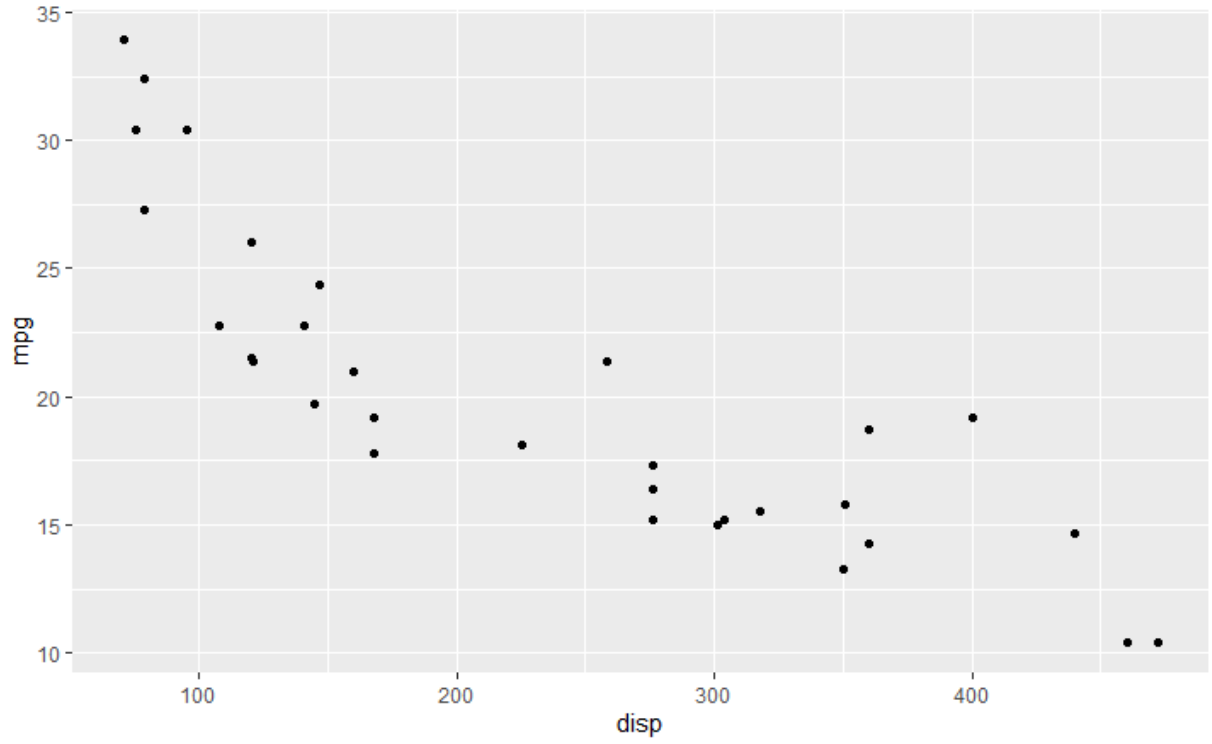
Let's recall our model assumptions.

- The errors (residuals) are random
- The errors are normally distributed with a mean of zero and constant variance
- The model is linear

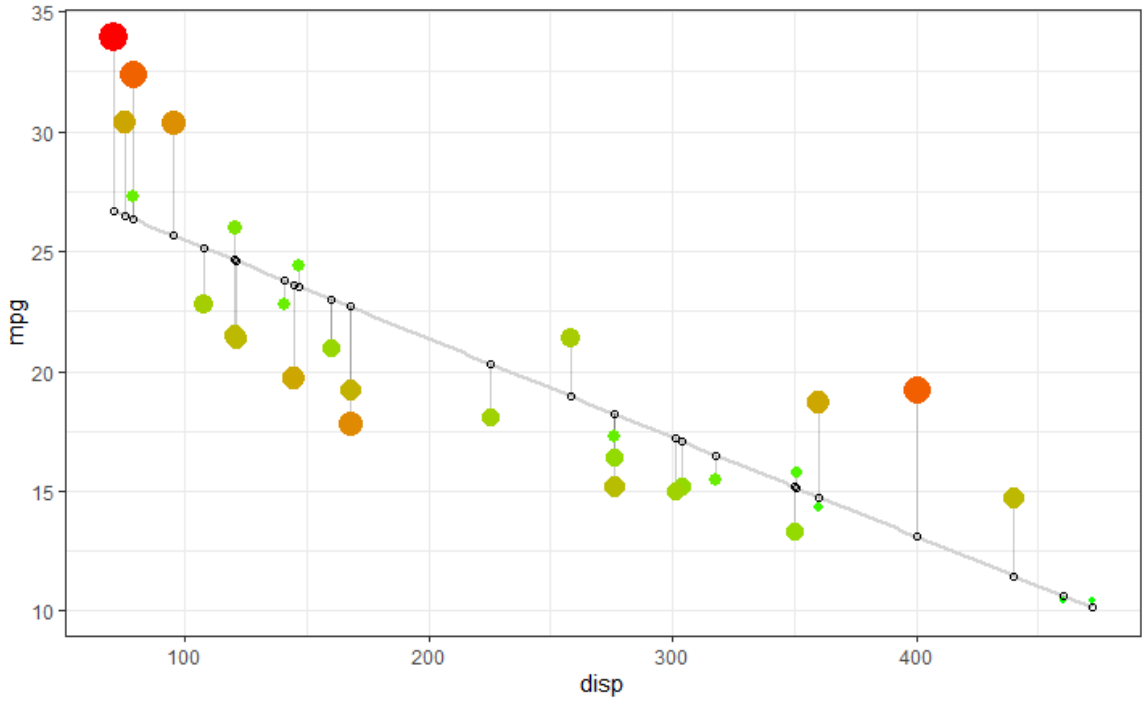
We will look at outliers in greater detail in a later lecture, but we'll begin to discuss them here because they can have a large impact on our model depending on its location in the data set. They are much more significant when we have only a few datapoints, but can still be problematic when we have a lot of data.

To test these assumptions on the errors, we are going to construct a residual plot. A residual plot is essentially a scatter plot. In a simple linear regression model like this, it is typical to plot the residuals against the independent variable. However, sometimes it is useful to plot the residuals against the observed dependent variables. We'll look at both.

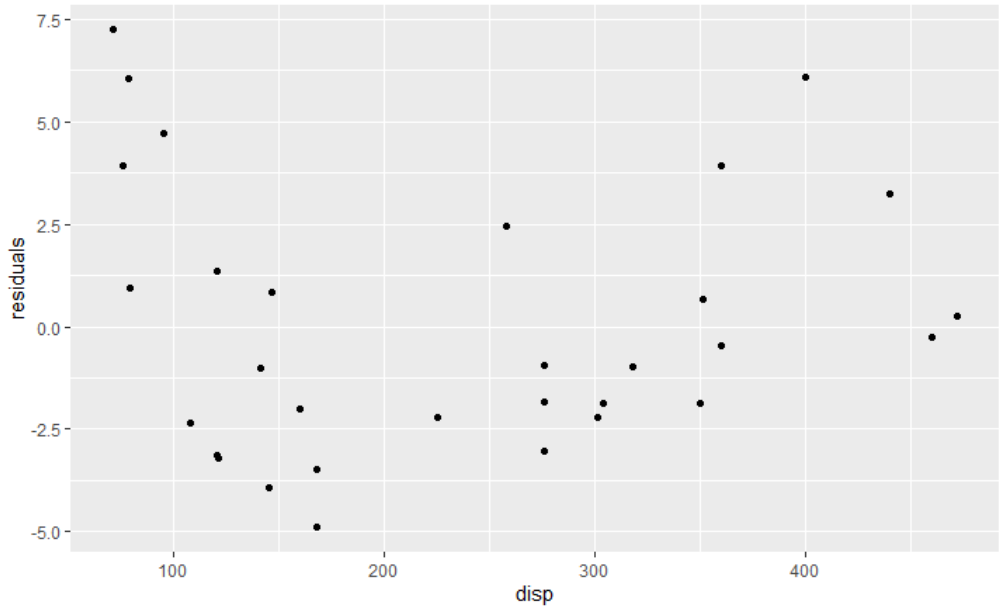
Recall that in a previous lecture we had looked at the relationship of the disp(lacement) variable in mtcars to the mpg variable. Let's replot that with displacement on the horizontal axis and mpg on the vertical axis.



The model doesn't appear entirely linear to me, but the curve is somewhat shallow so perhaps we could make a linear model work well enough. Let's impose a line on the data and see how the assumptions about the errors stack up.

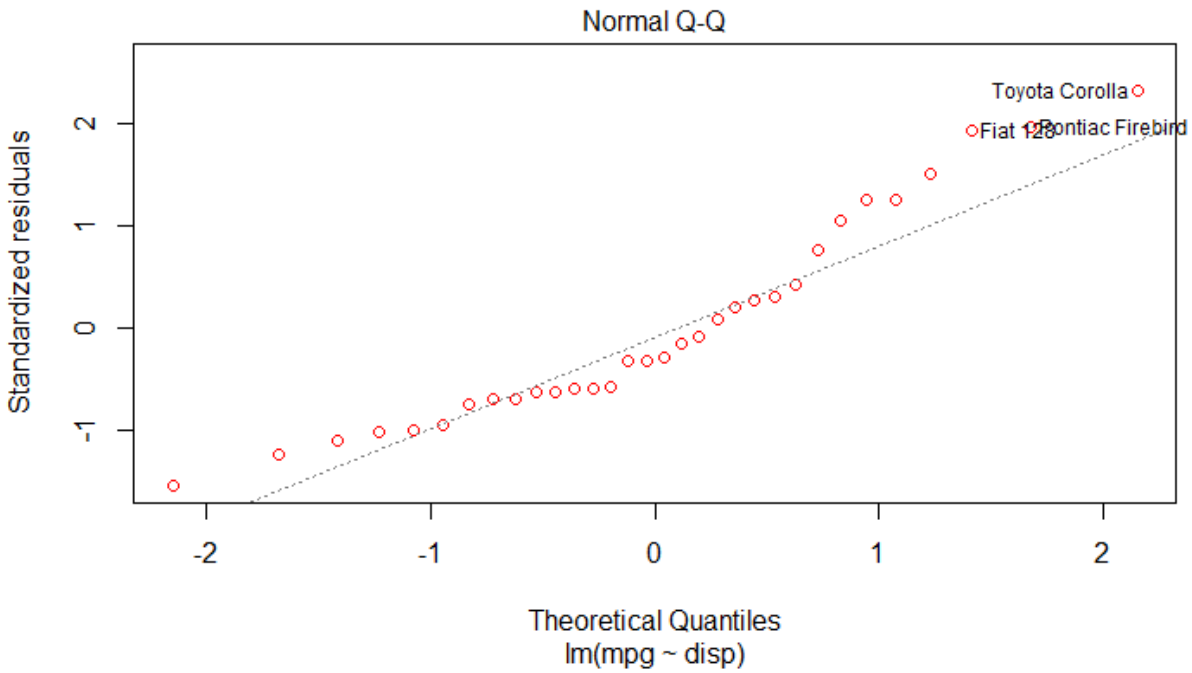


This graph imposes a line on the data and color-codes the true observations based on their distance from the regression line. A potential weakness of our linear model here is that the residuals appear to be mostly below the line in the middle, and above the line on either end. This is typical of a curved relationship when you impose a line on it. Let's look at the residuals alone plotted against the disp variable.



What we want is to see a random scatter here with no pattern. (imagine a horizontal line through y=0.) That does not appear to be what we have. The data appears to curve rather strongly in a kind of U shape on the graph. This means that a linear function is the best fit for this data.

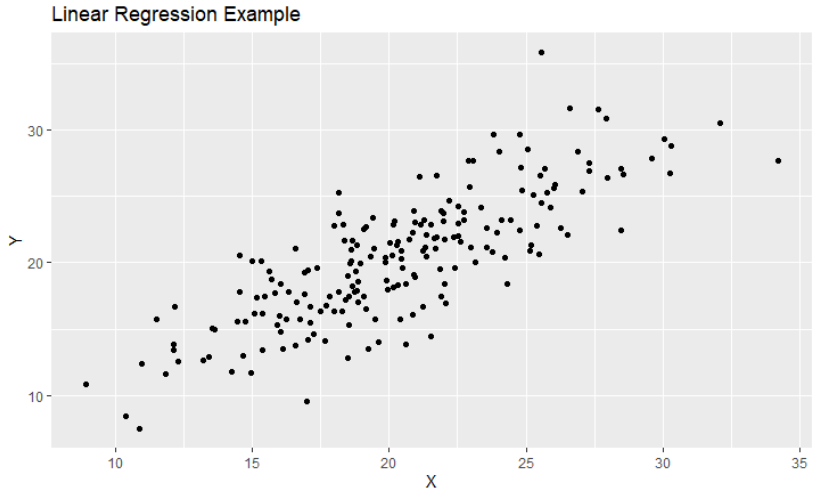
We can confirm another violation we are likely to have by looking at normality plot.

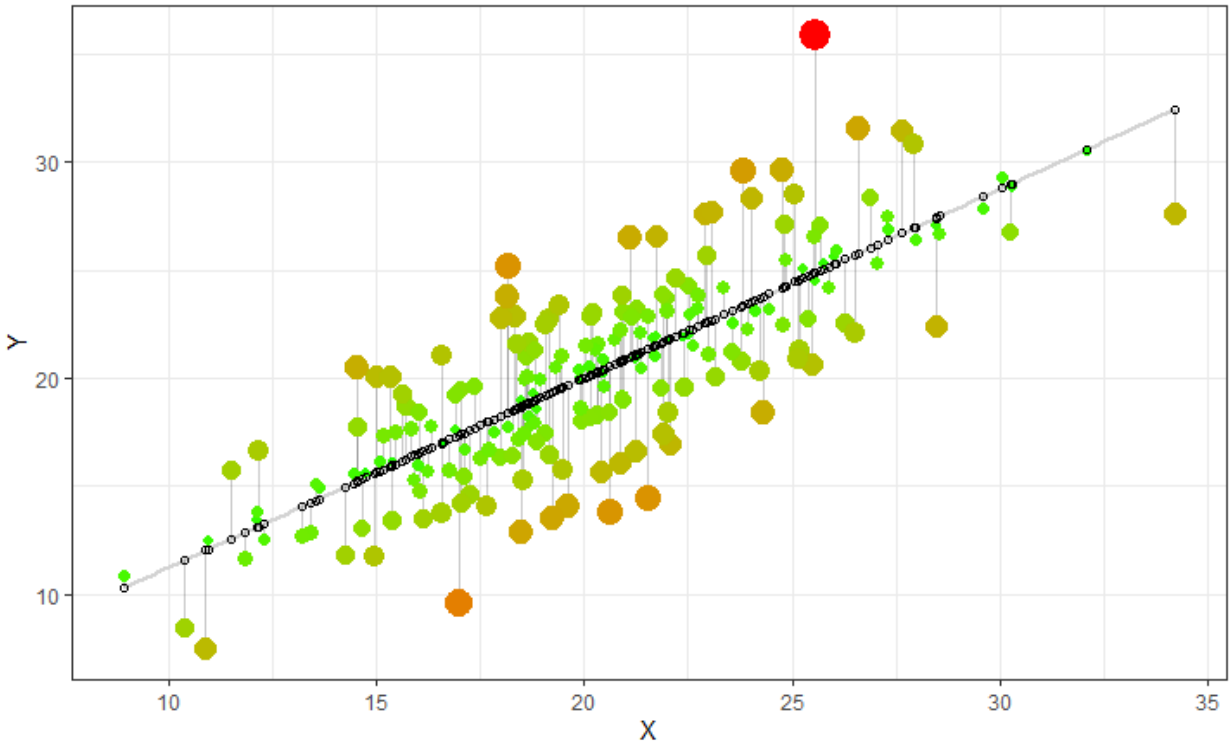


This isn't a great fit to the straight line, suggesting clearly that the residuals are not normally distributed.

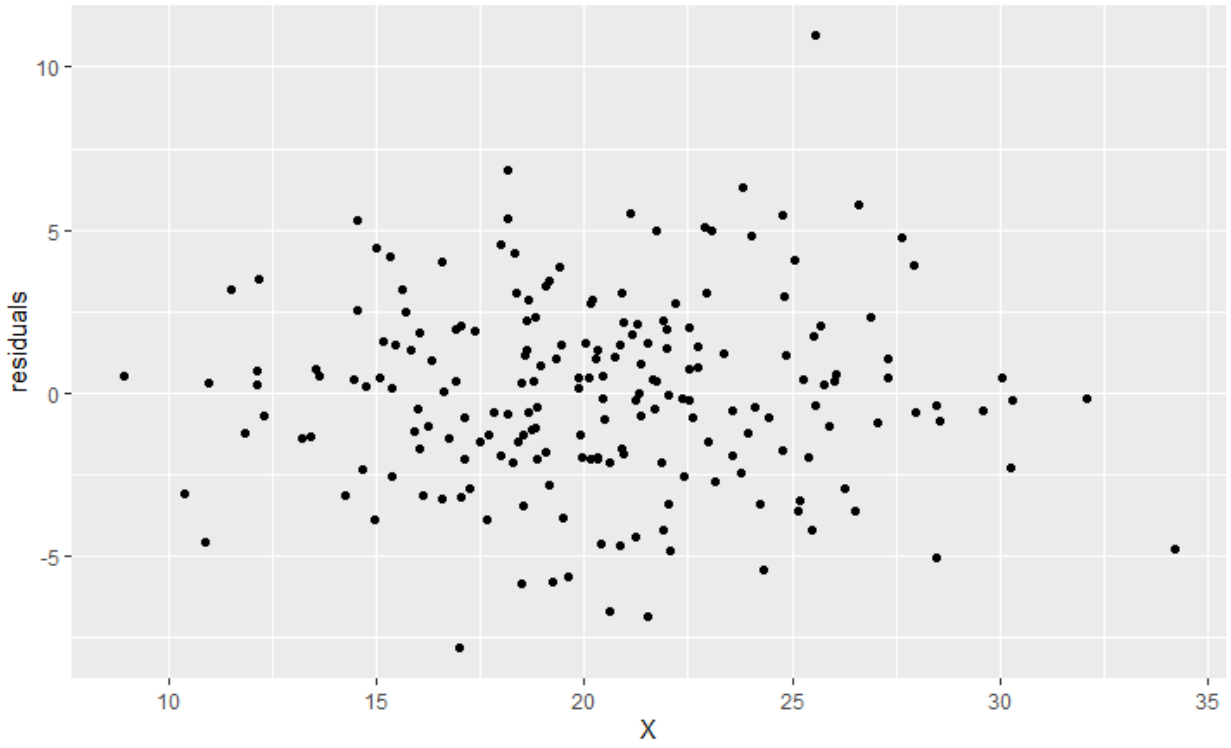
What should these features look like? Let's look at a simulated data set where we are certain that the relationship is linear.

Recall an earlier example.

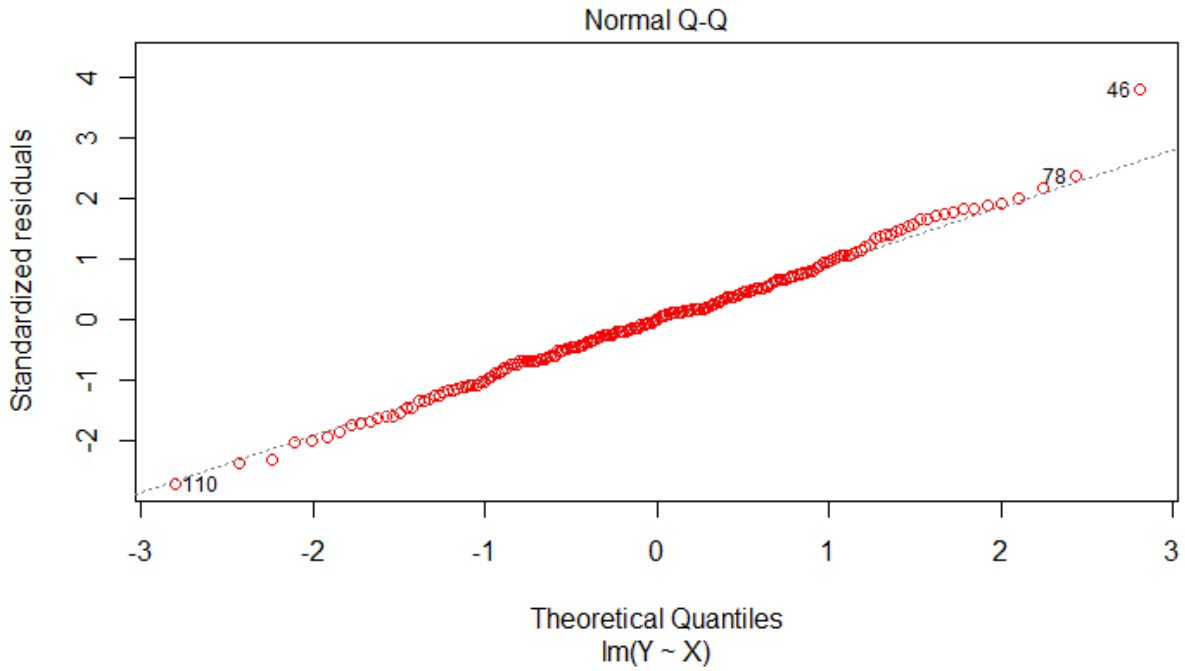




If we impose a line on the data and plot the original observations, we can see a very different pattern in this graph. The residuals fall on both sides of the graph and relatively evenly. Let's look then at a residual plot.



This is the kind of random scatter with zero features that we want from a residual plot. Let's check the normal probability plot.



This is the kind of tight fit to the line we want from normally distributed errors. This is a good sign that a linear model is the correct model for this data, and we've met the model assumptions.

We can examine the summary of the fitting procedure to test our model further.

Call:
lm(formula = Y ~ X, data = dat)

Residuals:

Min	1Q	Median	3Q	Max
-7.8167	-1.9052	0.0075	1.7489	10.9515

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.48770	0.95444	2.606	0.00984 **
X	0.87604	0.04581	19.123	< 2e-16 ***

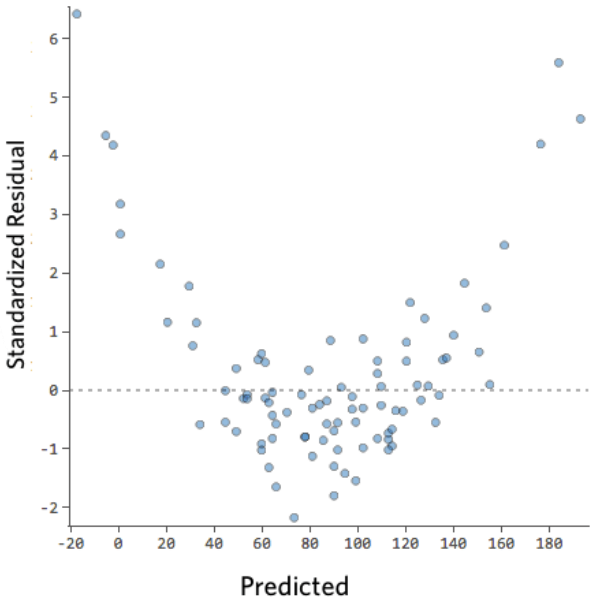
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.894 on 198 degrees of freedom
Multiple R-squared: 0.6487, Adjusted R-squared: 0.647
F-statistic: 365.7 on 1 and 198 DF, p-value: < 2.2e-16

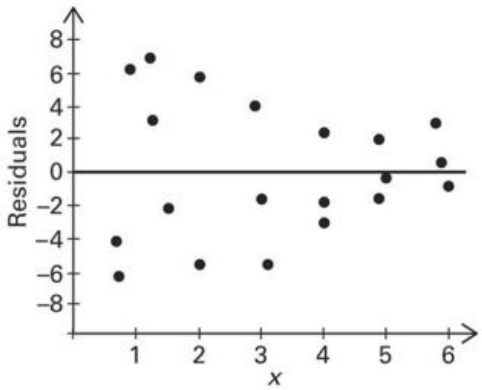
Note what is included in this analysis. The five-number summary on the residuals is here. There is data on the coefficients of our model including the estimate, the standard error for the coefficient, the t-value and the corresponding P-value. We talked in the previous lecture about conducting hypothesis tests on the slope coefficient, but we can conduct a test on any of the coefficients in the model, including the intercept. At the bottom, we also have a model test (for a simple linear model like this, one that is equivalent of the slope test). Recall that this test is not a test of the linearity of the model (as we were using the residual plot for), it's just a test of whether the relationship is more useful than none at all. So, even if the data doesn't strictly meet the assumptions of linearity and such, the model still may be better than the mean. As we develop more models, we'll be able to select models that fit better than the simple linear model and thus, have more predictive power.

We can also use the information in the table on the standard error for each coefficient to construct confidence intervals for each coefficient, using t-critical values for our confidence levels and $n - 2$ degrees of freedom.

We saw an example earlier of a residual plot that told us that the model was not linear (violating an assumption of our linear model). Another, even clearer example is below.



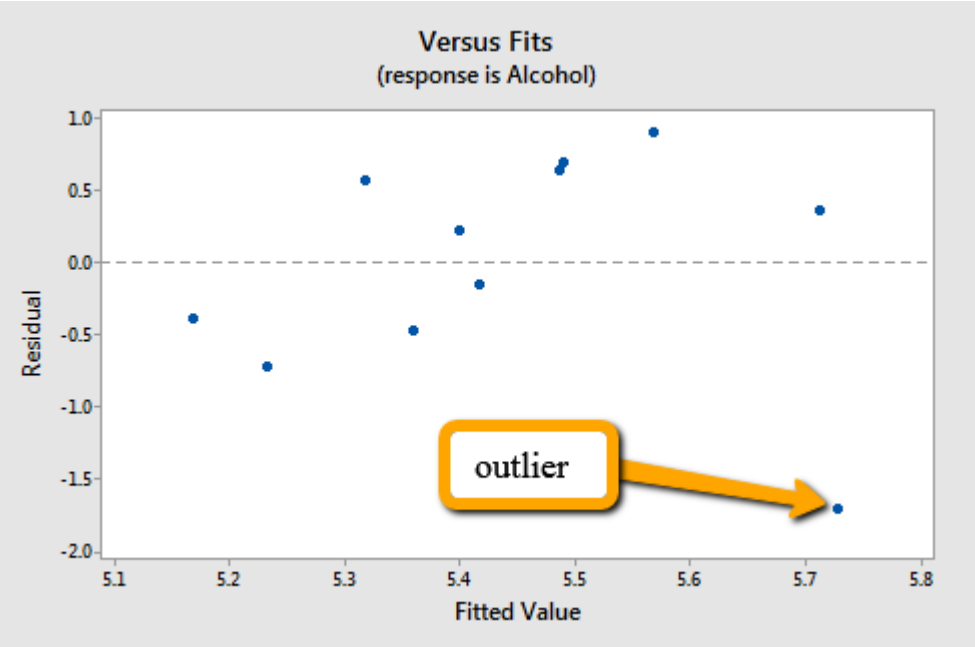
But this is not the only problem we can spot in a residual graph. Another problem is shown in the plot below.



In this graph, there is a kind of funnel or fluting effect where the residuals on one end of the graph are more spread out than the other end. This problem is called heteroscedasticity. It means that the spread or variance of the data is not constant. It's larger on one end than the other. Remember that we assumed that the variance was constant.

This particular problem we may be able to repair by performing a transformation of the variables (such as taking the log or the square root of one or more variables). We will look at this in greater detail when we examine nonlinear models later in the course. Some transformations are considered intrinsically linear, and these can maintain many features of traditional linear models after applying the transformation, including error assumptions and correlation calculations.

The last thing we want to look for in a residual plot is for any potential outliers.



If you have one (or a small number of) residual(s) that is much larger than all the others, that may be an outlier. We'll look at these more closely in future lectures, but these could be problematic observations that we will want to look at more closely. They are potential problems for our model. As the number of observations increases, we will have more of them but they may be less problematic unless they are very extreme. If you look back at our simulated model, there does appear to be one outlier on that residual plot, but there are 200 data pairs in that plot, and not a dozen or so in the one just above.

Some statisticians prefer to consider standardized residuals.

$$e_i^* = \frac{y_i - \hat{y}_i}{s\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}}} \quad i = 1, \ldots, n$$

Essentially, this formula converts all the residuals to their corresponding z-score. This may make it a little easier to identify outliers outcomes larger than 2 for unusual values, or extreme outliers if they are bigger than 3. This formula also accounts for the position in the data: values closer to the center will have a larger denominator and thus a smaller standardized value, compared to values near the edges of the data. However, the features we are looking for in the data are not significantly affected by the standardization.

The last topic I want to briefly mention is an extension of linear regression called weighted least-squares. In a traditional least-squares model, all the points in the data set have an equal impact on the resulting model. In a weighted least-squares model, we can give more weight to some values than to others, so that they have more impact on the resulting regression line than others. For instance, you may have data collected from different sources, but think that data from one source is more reliable than the others. You can weight the regression model so that those points will dominant the model, without having to throw out the other data sources.

We aren't going to discuss this kind of model in depth, but I've linked a couple of sources below if you want to learn more about this technique.

In the next lecture, we'll start looking at multiple linear regression, where we can use more than one independent variable to make our predictions.

References:

1. https://faculty.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf
2. https://assets.openstax.org/oscms-prodcms/media/documents/IntroductoryStatistics-OP_i6tAI7e.pdf
3. https://rpubs.com/iabrady/residual-analysis
4. https://quizplus.com/quiz/138379-quiz-10-correlation-and-regression/questions/11024705-the-following-residual-plot-is-obtained-after-a-regression-e
5. https://www.fuelcellstore.com/blog-section/model-validation-using-residuals
6. https://online.stat.psu.edu/stat462/node/120/
7. https://towardsdatascience.com/weighted-linear-regression-2ef23b12a6d7
8. https://online.stat.psu.edu/stat501/lesson/13/13.1