Lecture 4

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.

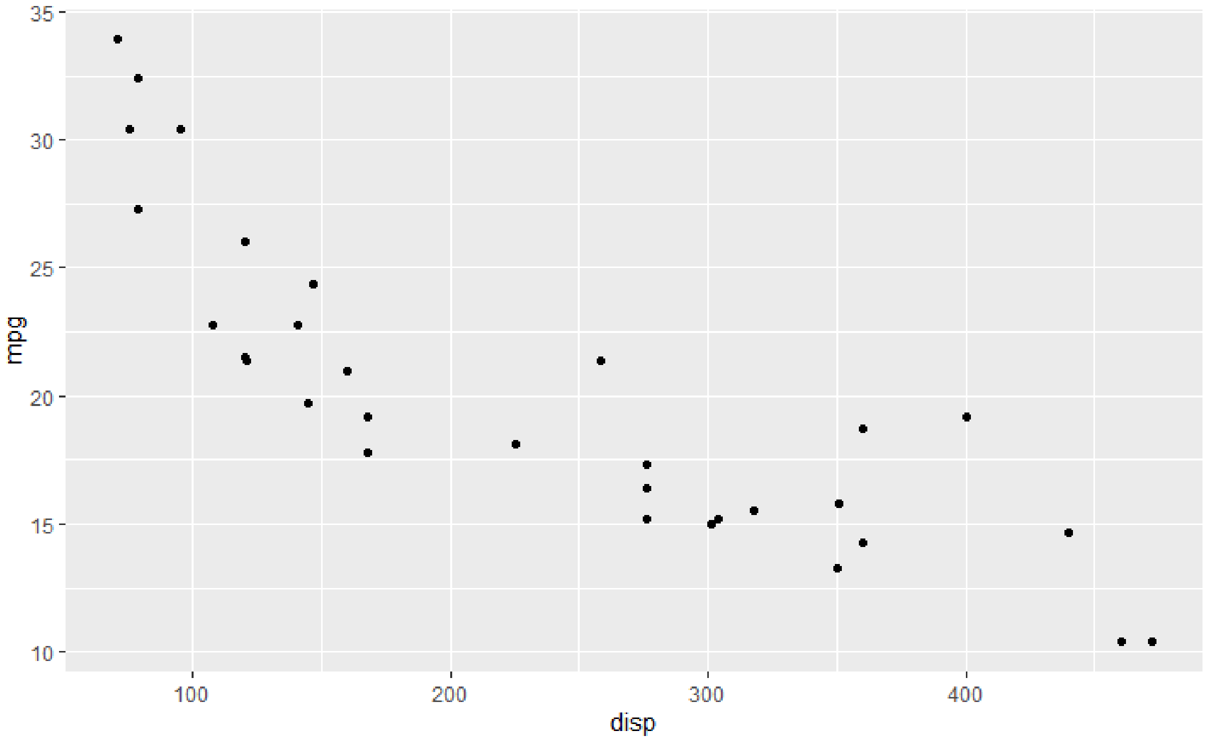
Let's recall our model assumptions.

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

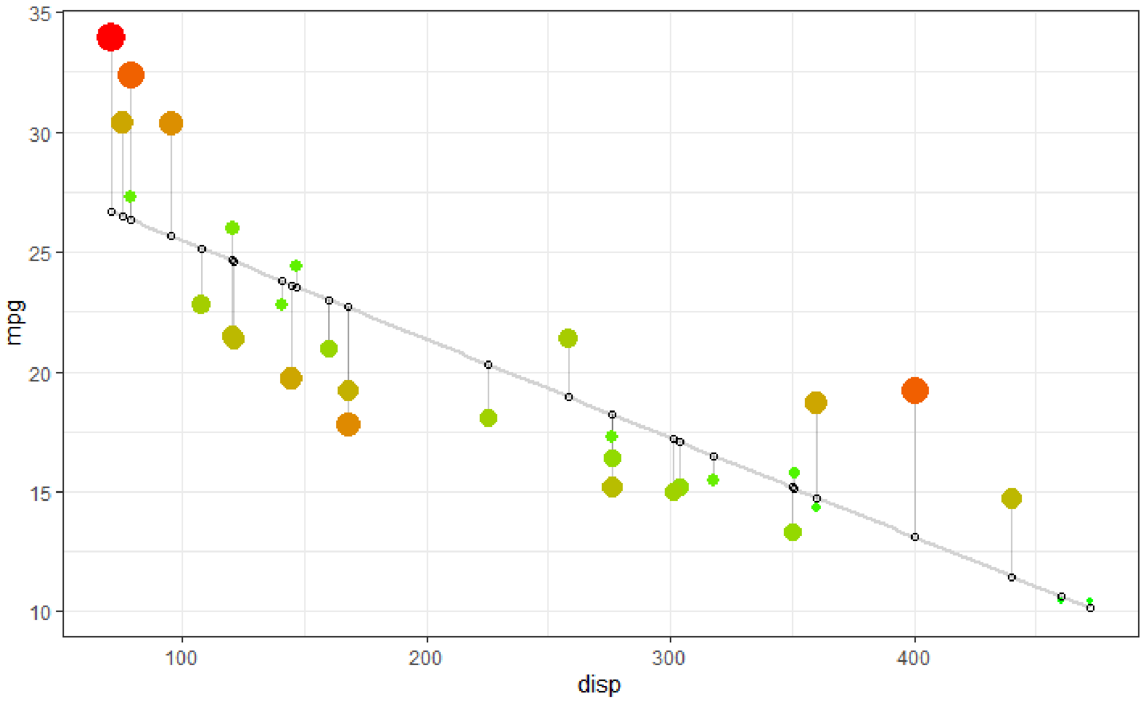
We will look at outliers in greater detail in a later lecture, but we'll begin to discuss them here because they can have a large impact on our model depending on its location in the data set. They are much more significant when we have only a few datapoints, but can still be problematic when we have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

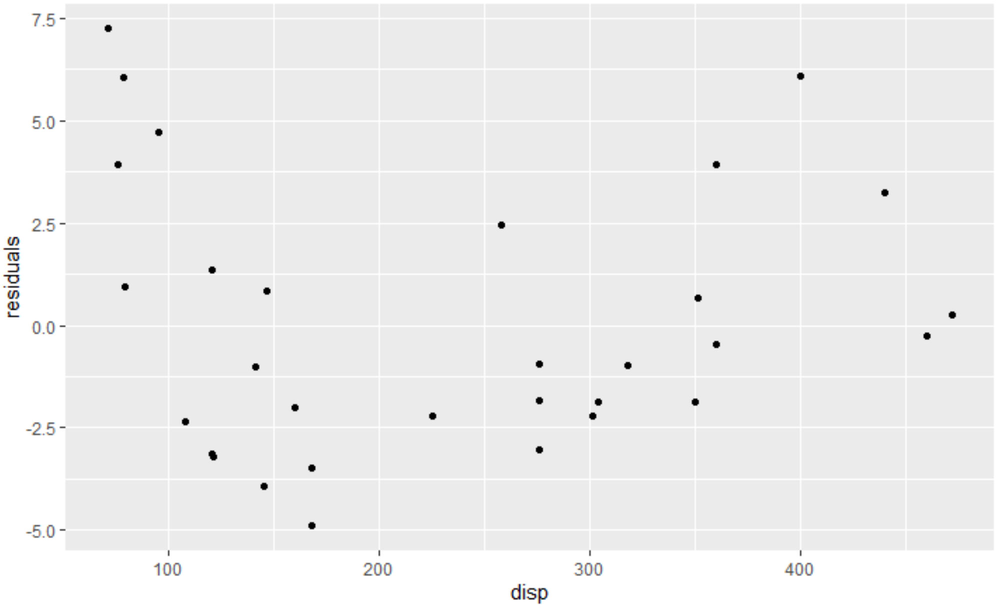
Recall that in a previous lecture we had looked at the relationship of the disp(lacement) variable in mtcars to the mpg variable. Let's replot that with displacement on the horizontal axis and mpg on the vertical axis.



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.

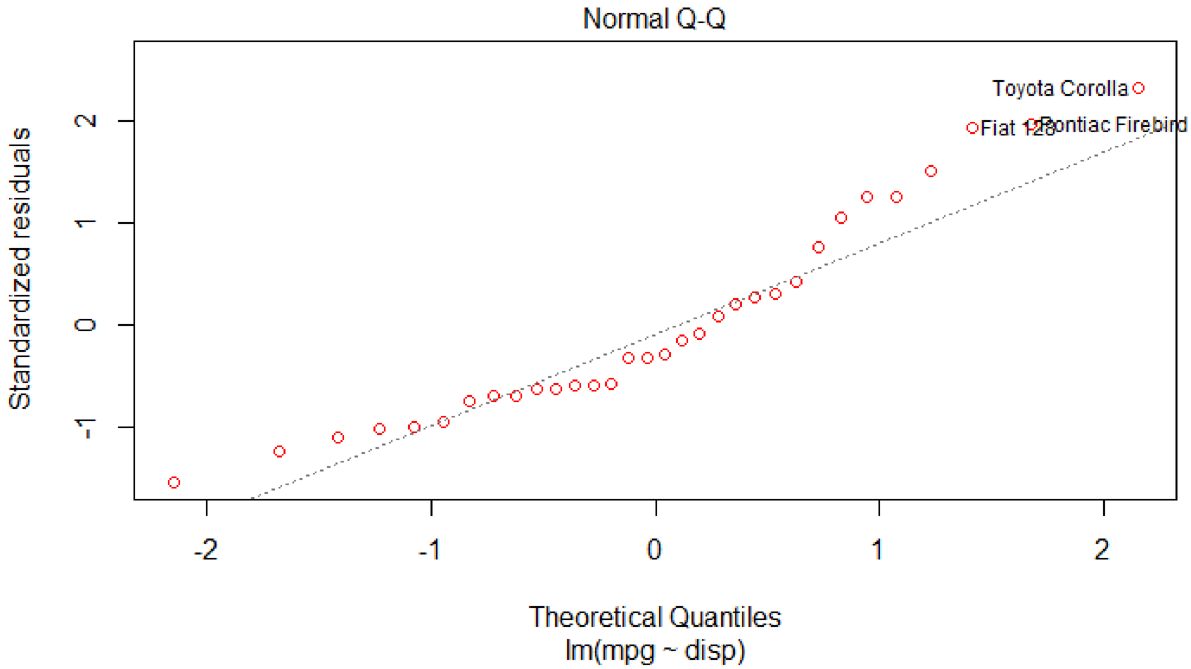


This graph imposes a line on the data and color-codes the true observations based on their distance from the regression line. A potential weakness of our linear model here is that the residuals appear to be mostly below the line in the middle, and above the line on either end. This is typical of a curved 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.

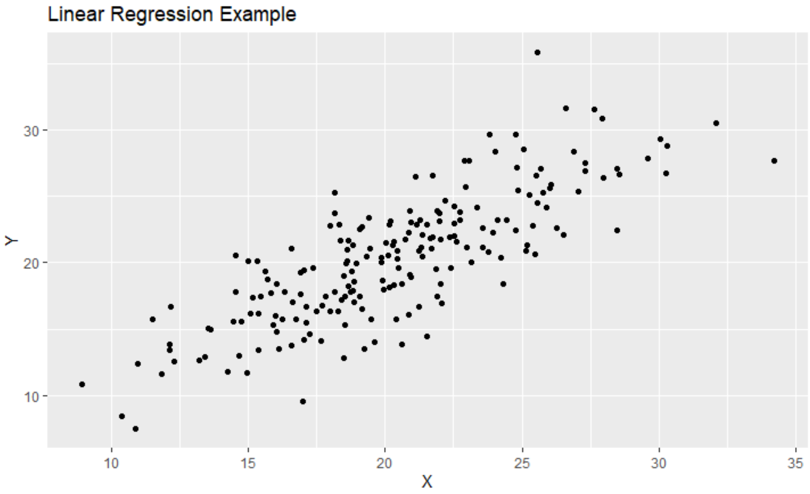
We can confirm another violation we are likely to have by looking at normality plot.

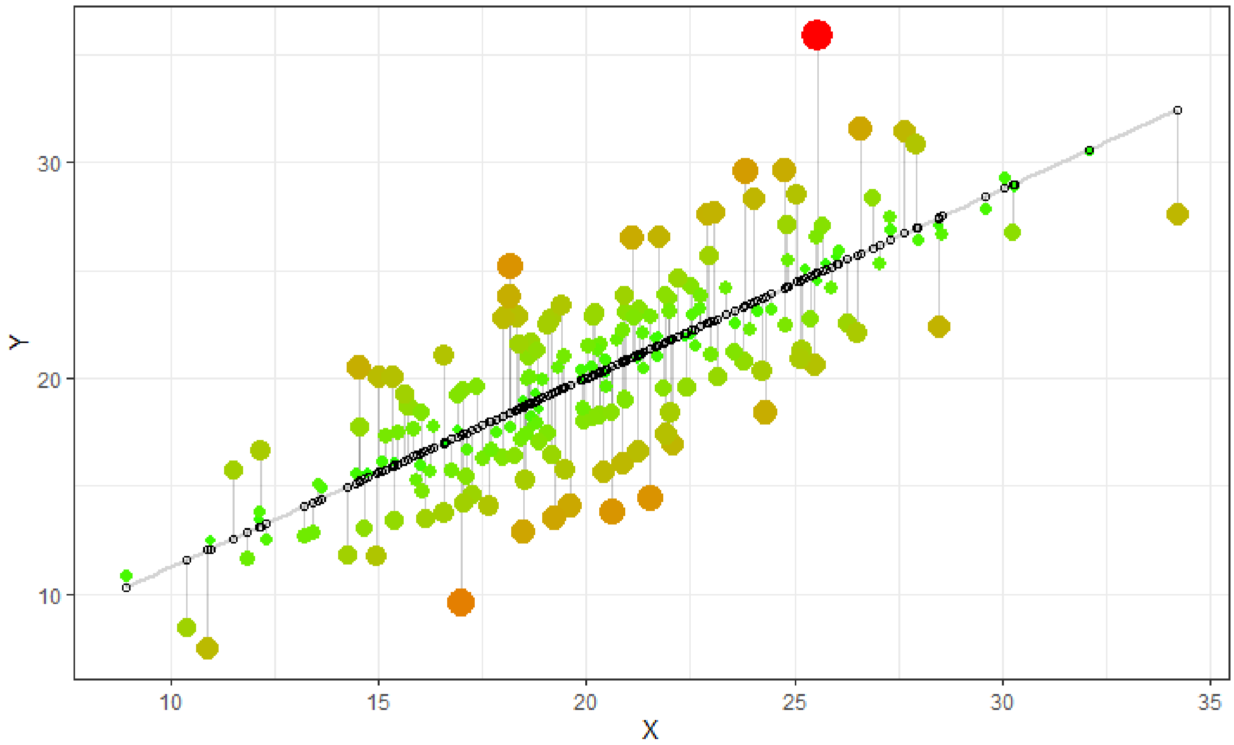


This isn't a great fit to the straight line, suggesting clearly that the residuals are not normally distributed.

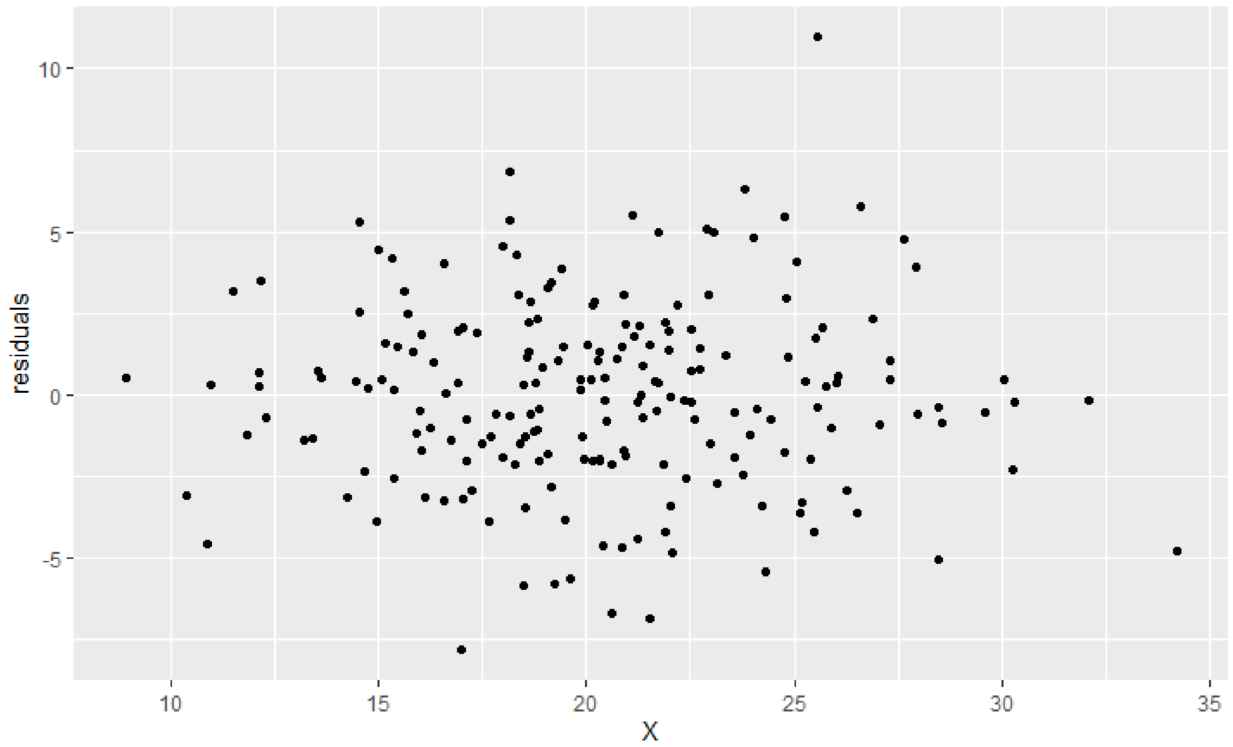
What should these features look like? Let's look at a simulated data set where we are certain that the relationship is linear.

Recall an earlier example.

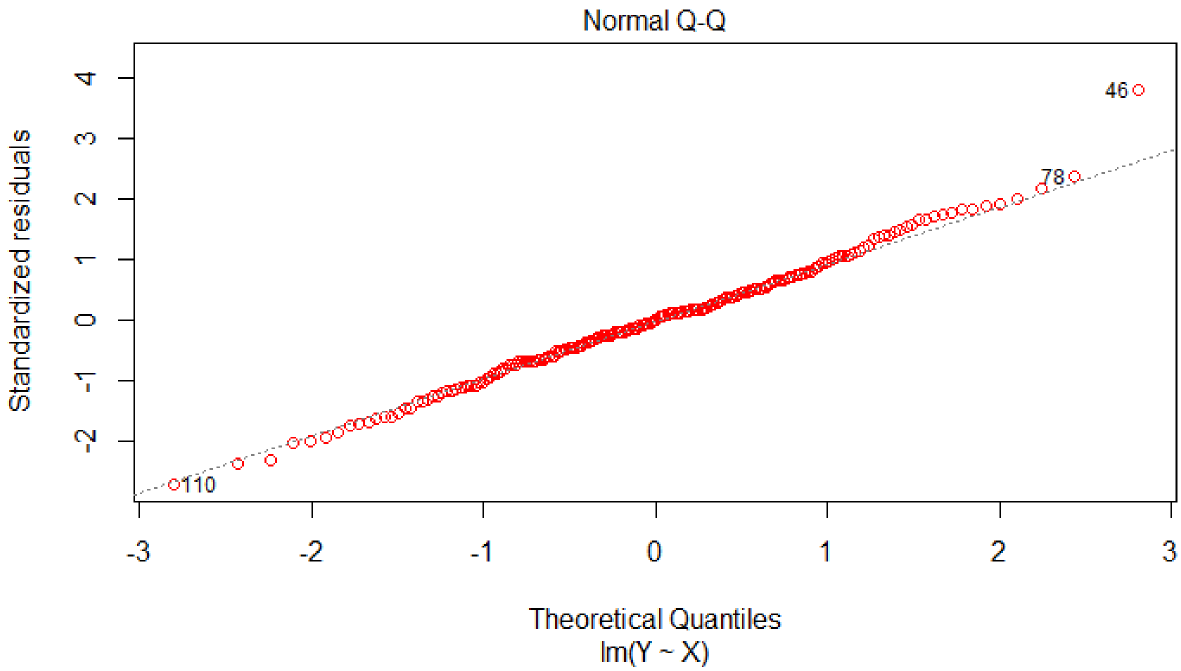




If we impose a line on the data and plot the original observations, we can see a very different pattern in this graph. The residuals fall on both sides of the graph and relatively evenly. Let’s look then at a residual plot.



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 ***

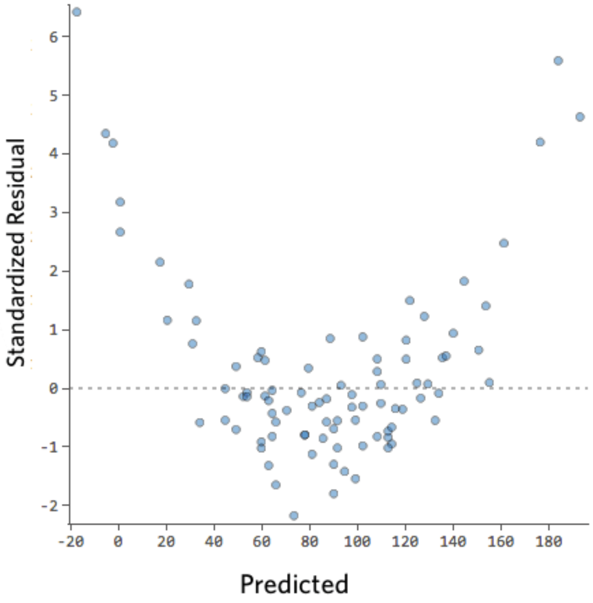
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

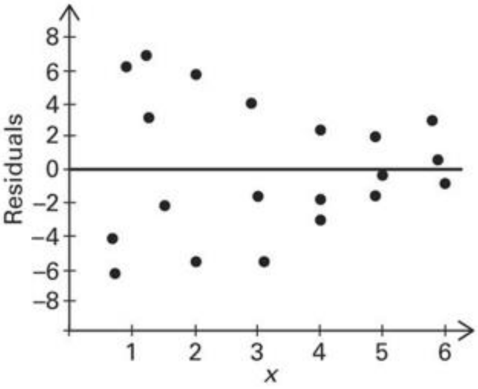
Note what is included in this analysis. The five-number summary on the residuals is here. There is data on the coefficients of our model including the estimate, the standard error for the coefficient, the t-value and the corresponding P-value. We talked in the previous lecture about conducting hypothesis 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.



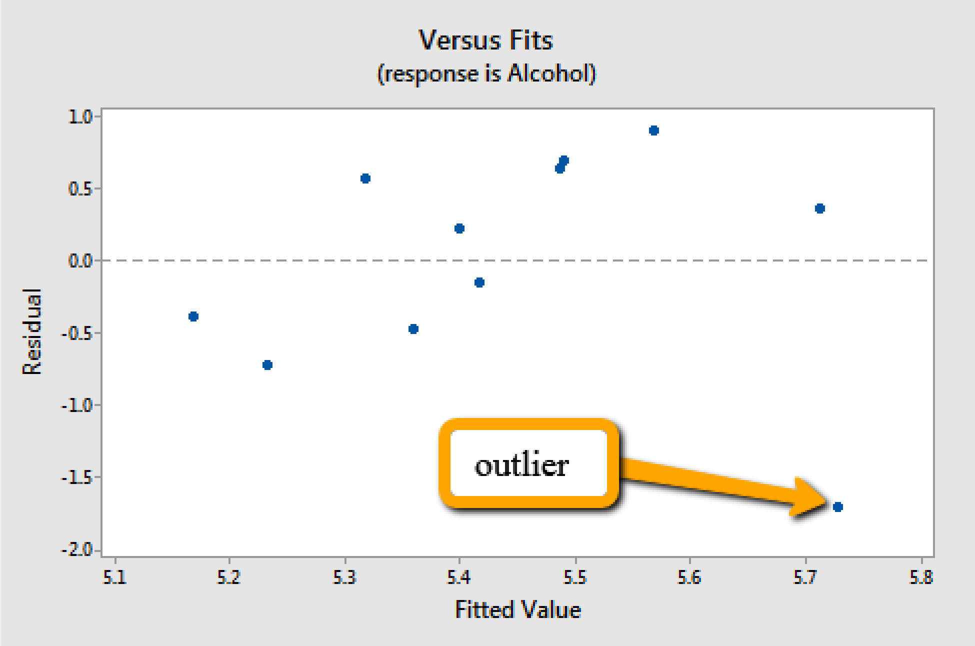
But this is not the only problem we can spot in a residual graph. Another problem is shown in the plot below.



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.

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

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.

Multiple linear regression

Multiple linear regression is an extension of simple linear regression, but here we are using more than one independent variable to predict one output variable. Typically, our model equation takes the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \epsilon$$

Like the simple linear model, the β's represent the true regression coefficients of the population which we estimate with b's. The variables in the model are all linear, and the error term at the end has a mean of zero and a constant standard deviation, which we can estimate from the residuals. The model in multiple dimensions is a (hyper)plane – when using two independent variables, it is a regular plane which we can graph and examine explicitly. When there are more variables, we can no longer look at the graphs, but we can still make predictions and do our other types of analysis. The linear equation here is sometimes referred to as a first-order model.

When we have more than one slope variable the ANOVA analysis for the model is no longer equivalent to the slope test since we have to look at more than one slope. We should now interpret the ANOVA, full-model analysis, as testing whether any coefficient for the variables in the model are non-zero, similar to a more traditional ANOVA. If the P-value is too high (greater than the significance level), then we conclude that none of the coefficients are non-zero. If the P-value is below the significance level, then we can conclude that at least one coefficient is non-zero.

From that point, we conduct tests on the individual coefficients. Testing information (test-statistic, standard error, P-value) is included in the summary results in R. We must then consider which if the variables is statistically significant. It may be that they all are, or only some of them. Later we will develop model selection strategies for building models with coefficients that are all statistically significant. For our initial discussion, we are going to consider the steps for building our initial model, testing parameters, and other new things we need to consider in the multi-variable case.

Each of our multiple-variable models may have different numbers of variables, and each case will require slightly different summation equations. However, the equation will use for the linear algebra approach does not change. We are, therefore, going to use that method with a small example to illustrate how the setup changes as the number of variables increases. We'll use technology to solve the models for us in any real world context.

Let's consider a situation where we have two variables that are independent which are predicting a third variable. Our data will then come in ordered triples, and our linear model will have the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Let's suppose our data is $\{(1,2,10),(2,4,15),(2,6,18),(3,7,24),(4,7,27)\}$. The first coordinate is x_1, the second coordinate is x_2, and the third coordinate is y. Replace these into our model equation.

$$\begin{gathered}\beta_0 + \beta_1(1) + \beta_2(2) = 10\\ \beta_0 + \beta_1(2) + \beta_2(4) = 15\\ \beta_0 + \beta_1(2) + \beta_2(6) = 18\\ \beta_0 + \beta_1(3) + \beta_2(7) = 24\\ \beta_0 + \beta_1(4) + \beta_2(7) = 27\end{gathered}$$

The coefficients of our β's go into our A matrix, the β's go into the B matrix, and the constants go into the Y matrix.

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 2 & 6 \\ 1 & 3 & 7 \\ 1 & 4 & 7 \end{bmatrix}, B = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, Y = \begin{bmatrix} 10 \\ 15 \\ 18 \\ 24 \\ 27 \end{bmatrix}$$

We solve this using the normal equation

$$A^T A B = A^T Y$$

Or the solution matrix

$$B = (A^T A)^{-1} A^T Y$$

We find that $B \approx \begin{bmatrix} 3.176 \\ 3.706 \\ 1.294 \end{bmatrix}$, which gives us the equation $\hat{y} = 3.176 + 3.706x_1 + 1.294x_2$.

Let's think about interpretation a moment. The constant is the predicted value of y when both x_1 and x_2 are equal to zero. As with the simple linear model, this may not be possible. If zero is far outside the domain of even one variable, then the intercept may be meaningless. So, interpret with care. The coefficient of each variable can be treated as regular slopes if all other variables in the equation are held constant. For instance, the value of y increases by 3.706 for each one unit increase in x_1 if x_2 remains the same. Likewise, the y value increases by 1.294 for each one unit increase in x_2 if x_1 remains constant.

One way we analyzed the relationship between variables in the simple linear model was using correlation. But correlation on works on pairs of variables, not three or more. There is another way to analyze models with multiple variables, using R^2. In other contexts, it's referred to as the coefficient of determination. In multiple variable models, it may be referred to as the coefficient of multiple determination. It really is the same thing. In the simple linear case, we can square the correlation to obtain this value, but in this case, we need another method to arrive at this number.

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

The numerator is the difference between predicted values and the mean, squared. The denominator is the difference between the original observations and the mean, squared. The denominator can be thought of as the variability in the original observed values, and the numerator in the variability in the predicted values. The ratio here can be interpreted as the amount of the variability in the y value that can be accounted for by the model relationship. That means that a high R^2 means that the model is a good fit for the data because the relationship account for most of the changes in y-values. If there is a low R^2 values, then the model accounts for very little of the variability and does little to improve our predictions over just using the mean. Because the value must be between 0 and 1, it is often expressed as a percentage. When we do a model summary in R, R^2 is one of the reported values.

The value $1 - R^2$ is sometimes thought of as the fraction of the original variability left in the residuals.

Our example model included only two variables, but we can include many more. How do we know if we've included too many? One way is to adjust the R^2 value for the number of variables included in the model. In statistics, we generally want to follow the principle of parsimony, which says that we want the simplest model possible that produces the best predictions. We can overfit models if we use too many variables and so the adjusted R^2 is a way of considering this. $adj\text{-}R^2 =$

$$R_a^2 = 1 - \frac{n-1}{n-(k+1)}\left(\frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}\right)$$

Here, k is the number of variables used in the model, and n is the sample size. The adjusted R^2 values will generally be smaller than R^2, but as we add variables, there will come a point when adding variables stops adding much to R^2, and may even result in decreasing the adjusted R^2. This is a sign that adding the extra variable is not adding enough new information or power to the model to justify using the extra variables.

The adjusted R^2 is also often included in our model summaries when we do regression in R.

We can use the R^2 to obtain what is called R the multiple correlation coefficient. It is the positive square root of the coefficient of determination. You can use the same range of values we used for correlation to assess the strength of the model.

As with the simple linear model, we can construct confidence intervals, and prediction intervals. We can conduct hypothesis testing on each variable in a model, and construct confidence intervals on each coefficient. These procedures follow the same general methods we used in the simple linear regression model.

One additional concern that we have with multiple variable models that we did not have before is that our “independent” variables may not be truly independent. We would prefer that our independent variables not be highly correlated with each other. This issue is sometimes referred to as collinearity. This will be one of the tests we’ll want to conduct when we assess our models.

Let’s look at a model using our mtcars dataset. Let’s look at a model of mpg using the disp(lacement) variable and the wt(weight) variable. The summary output looks like this.

Call:
lm(formula = mpg ~ disp + wt, data = mtcars)

Residuals:

Min	1Q	Median	3Q	Max
-3.4087	-2.3243	-0.7683	1.7721	6.3484

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	34.96055	2.16454	16.151	4.91e-16 ***
disp	-0.01773	0.00919	-1.929	0.06362 .
wt	-3.35082	1.16413	-2.878	0.00743 **

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

Residual standard error: 2.917 on 29 degrees of freedom
Multiple R-squared: 0.7809, Adjusted R-squared: 0.7658
F-statistic: 51.69 on 2 and 29 DF, p-value: 2.744e-10

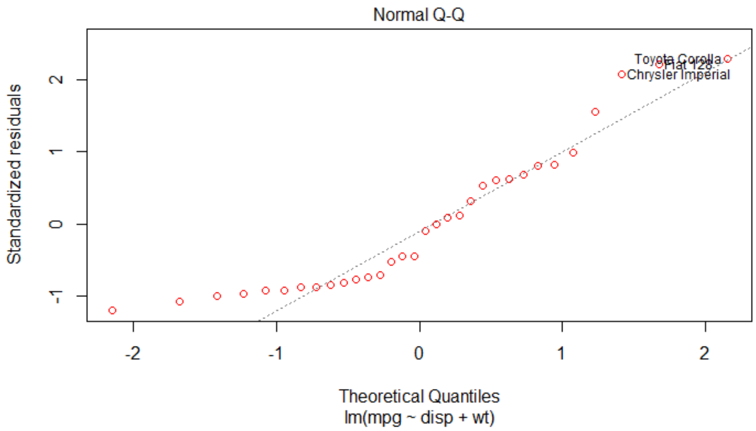
Let’s first consider our full model. Look at the last line of the summary output. This is the result of the ANOVA test, and we have a P-value that is very small. This indicates that at least one of the variable coefficients in the model is non-zero. So something in this model helps to improve our predictions of mpg.

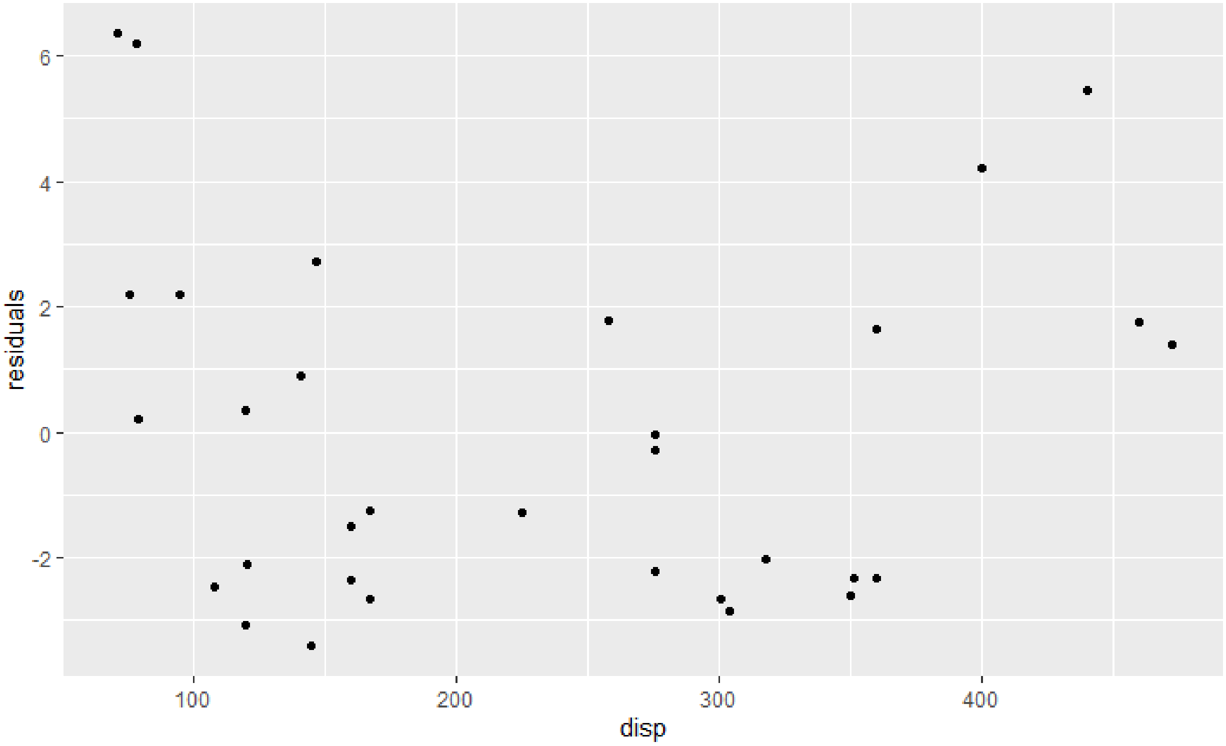
The multiple R^2 is about 78% (we’ve reduced the variability by this much). The adjusted-R^2 is similar but a bit smaller.

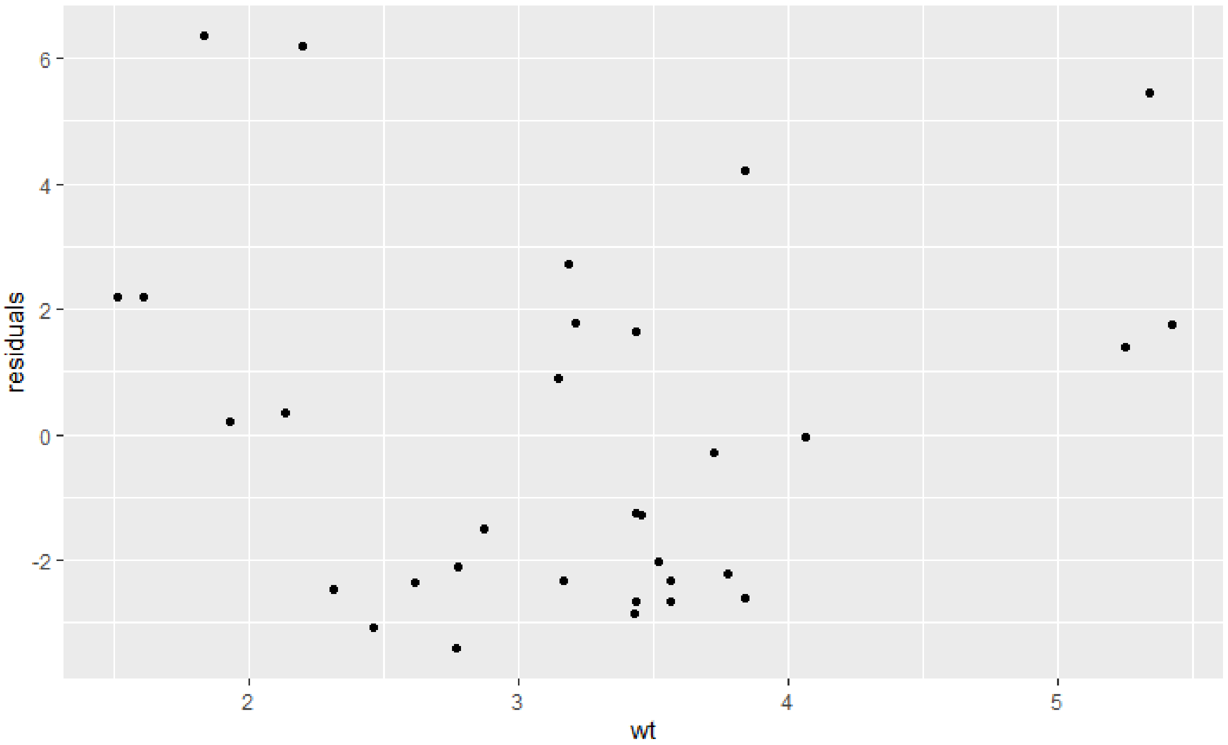
Let’s look at our table of coefficients. The P-value for the intercept is quite small, so it’s not zero. The P-value for the disp variable is 0.06, which is significant at the 10% level, but too high at the 5% level. The coefficient for weight, however, the P-value is 0.007 which is much less than 5%, so this variable should be kept in the model. If we think that the disp variable lacks significance, we can remove it from the model and rerun the test.

We should check for collinearity in our “independent” variables. We can do this by testing the correlation between them. We find that the Pearson correlation is 88%, which is probably why both variables are not statistically significant. Only one of them is contributing new information. We may wish to test the simple linear models with each variable to see which does a better job predicting mpg on their own.

We should look at the residual plots and the normality plot for the residuals. When we have more than one x-variable, we plot the residuals against each variable separately.







The normality plot indicates that the residuals are not very normal. The weight residual plot looks a bit better than the disp residual plot. This one seems to be much more widely dispersed on the ends and more positive, while in the middle the values are smaller and more likely to be negative. An effect that is not as extreme in the second plot. These may be signs of lack of linearity or a lack of constant variance. It suggests that our model assumptions have not been met.

It may be that we are willing to trade off some of these assumptions for a easy to understand model that is better than nothing. However, we can return to this when we have developed more tools for dealing with nonlinear models.

In the next lecture we will look more closely at outliers and influential points: detecting them, and what to do with them.

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
9. http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r