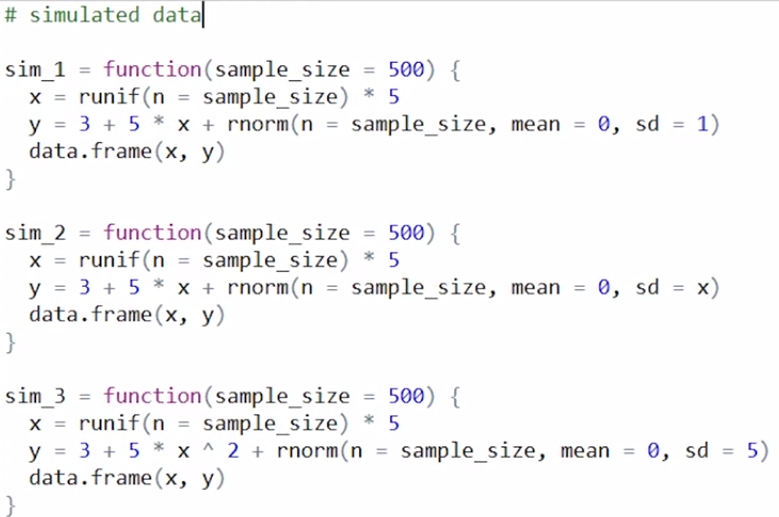
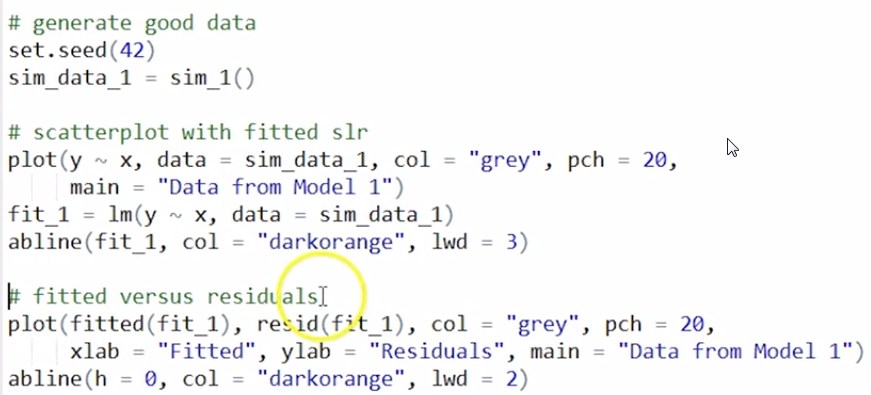
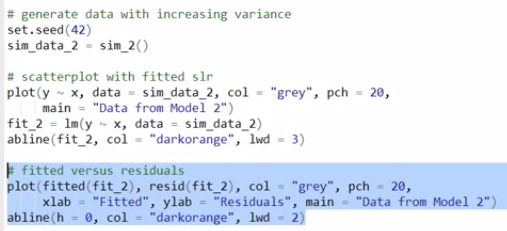
Week 8 - diagnostic tools for assessing the model assumptions. We'll generate some data ourselves so we know exactly which assumptions have been violated. This first function will simulate data that is exactly the Simple Linear Regression Model (no assumption violations) - mean is a linear function of x, and the errors are normally distributed with a constant variance. The second function where the mean is a linear function of x, the errors are normal, but the variance is a function of x (non-constant variance). And this third function has normal errors with a constant variance, but the mean is not a linear function of x (quadratic term) – this can be dealt with.



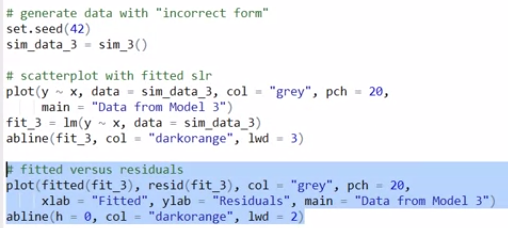
fit a simple linear regression model to that data and then add to the plot and this is a perfectly valid plot. Nothing jumps out at us here as a assumption violation but to double check that, we'll create a fitted versus residuals plot. So what we see here is noise happens to be centered at 0. And this gives us no cause for concern. We would not be worried about this data violating any assumptions.



From scatterplot with fitted SLR plot (for a simple linear regression only) we see there's an obvious non constant variance issue here. We'll move to a fitted versus residuals plot which also works for multiple linear regression. So for the fitted value, which is an estimate of the mean, we expect the residuals to be centered around 0 for any fitted value. Which they are here, but then there's clearly lower variance for low fitted values and higher variance for high fitted values. This plot would suggest a non constant variance issue.







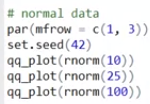
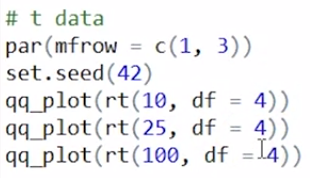
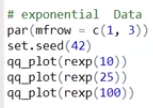


Lastly we'll generate some data that had a quadratic mean. we can immediately see some issues here and the fitted verses residuals plot will make that clear. So we are overestimating here, we are underestimating here and here. This actually has probably, roughly constant variance. The problem is that the residuals are not centered at mean for any fitted value. So we see this clear pattern in the fitted versus residuals plot. This suggests that the form of the model is incorrect. So we would consider that a violation of assumptions, the mean is not a linear function of the predictor.

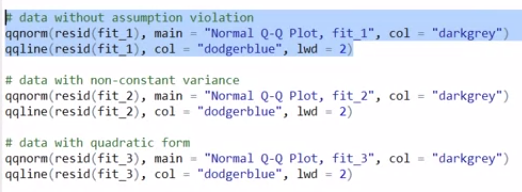
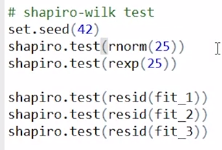
|  |  |
| --- | --- |
|  | So if I perform the bptest, it take as input a fitted model via the lm function. For first model we get a high p value. (null hypothesis was exactly constant variance). So this high p value does not make us doubt that assumption in any way. For the second model, we obtain a very low p-value. So this suggests we have reason to doubt the constant variance assumption which again this plot sort of very clearly tells us. Lastly, for the third data set, the BP test actually does not reject. Suggesting that there's not a non constant variance issue. Which again is correct, the constant variance isn't the issue here. It's the mean is incorrect here. While this plot does suggest an issue, the issue isn't exactly constant variance. It's the non-linear mean function. So the fitted versus residuals plot tells us about the linear mean function and the constant variance, but we'd also like to check the normality assumption, which is rather hard to see in the fitted versus residuals plot. |

|  |  |
| --- | --- |
|  | So one thing you might think to do is to create histograms of the residuals. Here it might not be a stretch to imagine that this first one has normal residuals. The second one has maybe normal residuals but we sort of see fatter tails here. And this third plot is sort of rather weird, so we probably wouldn't think about normality here. Histograms are sort of rather sensitive to the binning selection we create here, and generally it will be sort of hard to tell about normality. We'll want a better tool. That tool will be a qq plot. |

There are two functions, qqnorm and qqline, qqnorm will create points so again putting the observed quantities on the y-axis and what we would expect to see if it was from normal on the x-axis and then qqline we'll add the line we expect to that plot.

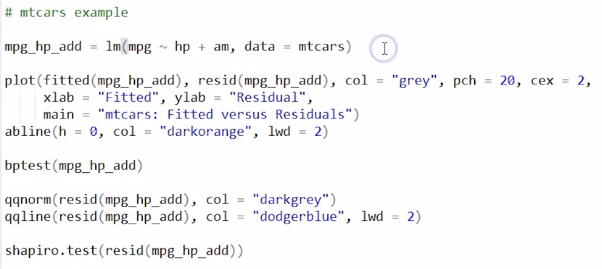
  

with a greater sample size, we sort of see that there's generally not too much deviation from the line, normally it happens sort of at the very end if at all whereas for a low sample size, we can see some deviation that sort of appears fairly big from time to time. I've simulated data from a t distribution with different sample sizes with a fairly low degrees of freedom, sort of giving it a heavy tail. And this is sort of easily seen in the high sample size case with the sort of tail behavior here where it becomes further and further from the line. And then we'll also simulate from some exponential data, which is something that is skewed. This is fairly clear for the medium and high sample sizes and even with the low sample size.

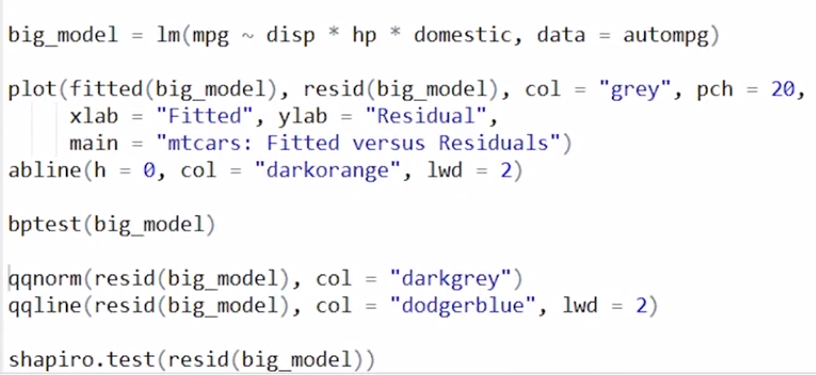
 

normal q q plot for the data that did not have violation of assumptions looks pretty good. we only see some very minor deviations in the tail, but that's somewhat to be expected. When we run this on the data that had non-constant variance, this actually ends up with what sort of looks like fat tails. So, again, technically the errors here were normally distributed. But because their variance was a function of x when we try to fit a single normal distribution to that it sort of breaks down and appears to have fat tails. And a similar issue with the data that had a quadratic mean we see sort of this odd tail behavior here. So this is sort of suggesting that normality assumptions is violated. But, again, it's actually a mis-specified model, if you'd like a test instead of a plot, we talked about the Shapiro-Wilk test.

If we run test on data that was generated from a normal distribution, we would hope that it would not reject. null hypothesis here is that the data sampled from a normal distribution so. High p value and here I'm generating some data from a very non normal distribution, we get a very low p value, and then again returning to the data from the model with no assumptions. High P value, looking good. Model with a non-constant variance, very low P value, we would reject that, gives us cause for concern. And similarly, that third model, also a very low P value. So we reject the assumption that the data could have been sampled from a normal distribution. A minor note here is that Shapiro-Wilk test isn't a test specific to linear models It's about any data in general. And the data that we're giving it is the residuals of a model fit via LM.

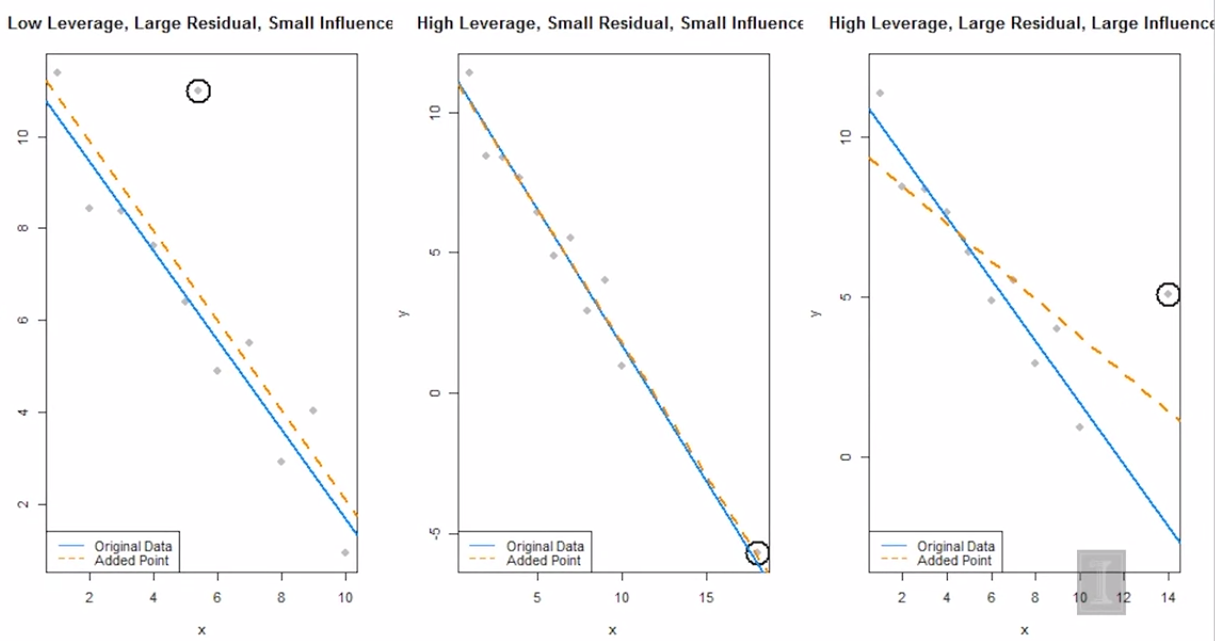


So this is an additive model for the mt(cars) data, where we're trying to model the fuel efficiency using the horsepower and transmission variable, whether or not it was automatic. So we'll fit that, so we would like to assess whether or not there's any violation of assumptions of this model. So we'll first plot a fitted versus residuals plot. Since there's not a ton of data, it would be sort of hard to actually see a pattern here. We could run the BP test and it actually obtains a fairly low p-value. you would reject and claim that there's an assumption violation of non-constant variance. Generally, we sort of like to defer to the plot, and it seems reasonable, so we probably wouldn't cast out here. Similarly, we can look at the normal qq plot, which looks fairly good. We see a couple deviations here but nothing too wild and I believe the Shapiro-Wilk test will confirm that, so high P value so we're not too concerned with normality.



So moving on to the larger auto MPG data set. It was a model with three way interaction between displacement, horsepower and this domestic variable as well as all the lower order terms. So if I plot a fitted versus residuals plot this is a tough one. We sort of don't see an obvious pattern. Maybe if it weren't for these three or so points here you could argue that there is some increasing variance. But it's sort of hard to say. If I performed the BP test here, I get a very low p-value. That actually does maybe together with this plot maybe cast a little doubt on the constant variance assumption. It doesn't seem to be a horrible violation but we do see some evidence that there may be increasing variance for increasing fitted values. And then similarly, the QQ plot here we see this sort of obvious fat tail behavior going on here. And Shapiro-Wilk test will confirm that. And so in the next couple lessons we'll maybe look for ways to improve these two diagnostics, the fitted versus residuals and qq plot to see if we can maybe by changing the model a little bit, and maybe doing some other things, improve upon this possible violation of assumption here.

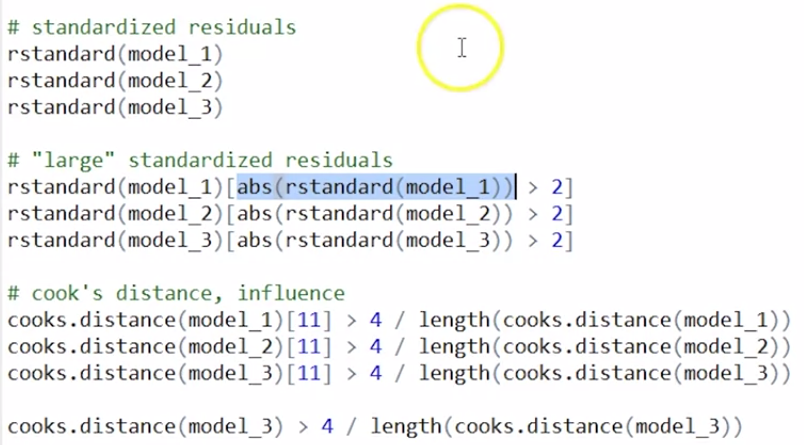
DETECTING UNUSUAL OBSERVATIONS



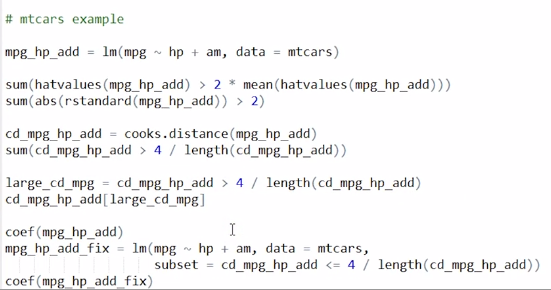
There's 11 data points in each, the 11th data point being the circled one in each, so they all share this data here, but then I have these three points that correspond to the data set for each panel here. the orange dash line is the model fit to all 11 data points, just different in each plot. The blue solid line is a model fit to the ten data points excluding the circled points, which is actually the same in each plot. The first thing we'll want to do is calculate leverages. We said that leverages are the diagonal elements of the hat matrix. hat values function is saying I want the leverages for the observations used to fit the model. So the leverages of these data points are essentially in order, except for this 11th one.

|  |  |
| --- | --- |
|  | But are any of these considered large leverage? And then here we see that, none of these points have a large leverage. Whereas in panels 2 and 3, we can see that both times the circled point indeed is considered a point of large leverage. Another way we sort of assess the unusualness of an observation, are it's residuals. we'd like to detect if any of these are large standardized residuals. We said that standardized residuals were approximately standard normal, we would expect about 5% of them to be greater than 2 and not to happen too often. So the 11th point, we say has a large standardized residual greater than 2. In the second panel here, we actually detect no points with a large standardized residual. And in the third panel, again, the 11th point has a large standardized residual. |

We had this heuristic that said, Cook's distance is considered large, and thus the point is considered influential. If it's larger than 4 divided by n, the number of observations.

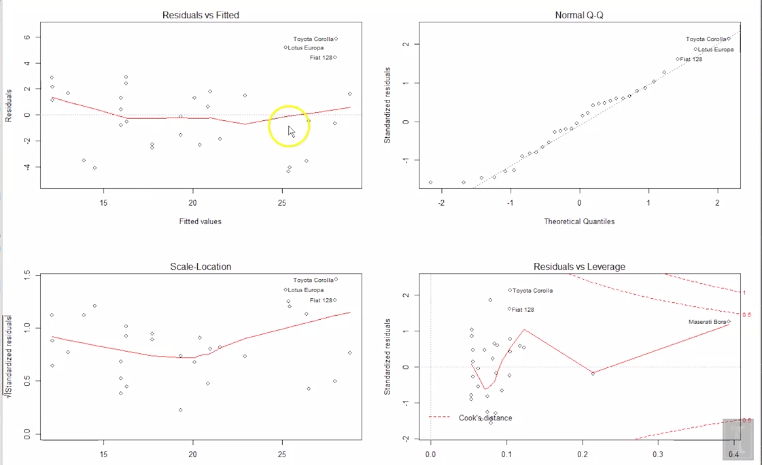


I'm checking if the 11th point is influential. We had said previously that the first case is not, because it's very low leverage. Again, we see the fit in the plot not changing much whether or not it's in the fit. Second case again, it should not be because, while it is a point of high leverage, it has a low residual, and in third case, we expect it to be influential. Running the function, the first one is not influential, that's what we expected. Second one, not influential, that's what we expected. And the third one, is indeed influential. Interestingly, this point has such an influence and pulls the regression so close to it. I believe if we look here, just checking each point for whether or not it is influential. Turns out this first point, so that point here, is also considered influential, but really it's this one that causing the issues. It's pulling the regression so close to it that it's pulling the regression away from this point, and it appears influential.



Going back to our additive model for the mtcars data set, I'm obtaining the leverages, and seeing how many of them are considered large, and we see that there are two points of large leverage. Here, I'm finding how many of the points have a somewhat arbitrary large standardized residual, which that is a magnitude greater than 2. And there is exactly one of those. Maybe that resulted in an influential point, so I'm going to store the Cook's distances. So we believe that there are two influential points here. It's a Toyota and a Maserati. So these are two rather different vehicles here. Toyota Corolla being a very economical car. Whereas a Maserati Bora being a supercar from the 70s both having a large influence on the model. So I'm going to fit the model exactly as I had before, but then subset according to this factor here, which is true for all of the points except for these two influential points. So it'll fit to this data, but only where this condition is true. Now if you look at the coefficients, they haven't changed all that much. Those two influential points, while they did have some influence, it wasn't that big of a difference between the two.

if you call plot on a model object, what I get here are quickly a number of diagnostic plots. So here we see exactly the fitted versus residuals plot. It adds the horizontal line and actually adds this nice smoother to the data. And we see that this smooth red line is not that far from 0, so that's good, we'd like that. We see a normal QQ plot and it labels a few of the more extreme points. But this plot here is an interesting one as well. It plots leverage against standardized residuals. these curves here are contours of Cook's distance. Now we can see this Maserati that we identified and this Corolla here are two of the larger Cook's distances, and those are points that we labeled as potentially influential.



Moving on to the larger data set we fit our three-way interaction model again. We weren't a big fan of this, because it looks like there's some non-normality here, where you're sort of assuming the normality assumption is suspect. What I'm going to do is, I'm going to obtain the Cook's distances here, and check how many of them would be labelled influential, and it's actually a fairly large number. And recall that this would also the Shapiro-Wilks test would reject, so again, that confirms what I'm seeing here. But so now what I'm going to do is, I'm going to remove those data points from the model fitting. So again, fitting the model the same way, but now subsetting two points that aren't influential. Now if I repeat the QQ plot, and perform the Shapiro-Wilks test, the results are much better. I haven't actually justified removing those data points, I removed those data points and everything works out better. Well, we should actually maybe do is think about, are there other models we could be fitting here? Maybe we could consider some additional terms that we're not considering right now, or what we'll see in the next lesson, is maybe consider some transformations of these current predictors. Because again, we would rather not remove data, unless we absolutely have to. So while it does fix a problem here, we haven't really justified the removal of those data points. [MUSIC]

