Slide 1 – Title Slide

Hello and welcome to Week 6, Part 5 of EGM101: Outliers. In this lesson, we'll learn about outliers: what they are, how we can identify them, and how we can handle them when we encounter them in our data.

Slide 2 – What is an outlier?

For starters, what is an outlier? In general, an outlier is any extreme value in our data that lies far away from the rest of the data.

As an example, if our data consisted of the net worth of everyone in this class, and any random person that you picked from a list of people who have far too much money lying around, the random billionaire would be an outlier. Unless most of the rest of the class is also a billionaire, their net worth would be so much larger than everyone else in the class that we would need to ignore it to learn anything useful about the rest of the class.

In the context of this week's lectures, an outlier is very specifically a point that is far away from the regression line.

Slide 3 – Effect of outliers on regression

Remember that when we covered least-squares regression and saw how to calculate the best-fit slope and intercept, the slope is calculated using the co-variance between the explanatory and response variable, and the variance of the explanatory variable. Remember also that because both of these calculations use the square of the difference between the value and the mean, they tend to be more affected by extreme values.

What this means is that when we do linear regression with outliers, the outliers tend to "pull" the regression line toward themselves, in much the same way that billionaires tend to continue to pull wealth toward themselves. In the plot here, we can see the difference between the red line, calculated using only the black points, and the blue line, calculated including the point indicated by the blue square. The slope of the blue line is larger than the red line, because the blue slope is "pulled" upward toward the value plotted using the blue square.

This is especially true if there's a large horizontal distance between the outlier and the rest of the data – in this case, we call the point an "influential point" because of how it influences the regression calculation.

As we discussed in the previous lesson, and as we can see on the plot here, as the distance between the data and the regression line increases, the R-squared value decreases because there's more variability that isn't explained by the model. The black points are all fairly close to the red line, but there's a lot more distance between all of the points and the blue line – this is reflected in the difference between the R-squared values shown in the lower-right corner of the plot.

The increased distance between the data and the regression line also means that the absolute value of the correlation coefficient is lower, because the linear relationship between the two variables is not as strong.

Slide 4 – Identifying Outliers: Graphical

Okay, so how do we identify outliers? Is there a better way than just looking at the graph and saying "this point is far away from the others, so it must be an outlier"?

The answer is, of course! As a basic rule of thumb, we can consider any point that is more than 2 standard deviations away from the regression line to be an outlier. Probably. I say "probably" here because this is really only a way to identify potential outliers – deciding whether or not the data point is actually an outlier is something that we will discuss in a few slides.

Note that this is the standard deviation of the residuals, not the observations – we're comparing how the observations compare to the model, not the dispersion of the observations.

And a second note: when we do this calculation, we divide by n minus 2 instead of n minus 1 – this has to do with something called the "degrees of freedom", which is something that we'll cover more next week.

So, to identify outliers on a graph, we need to first plot the points and the regression line. Second, we need to plot the regression line plus or minus 2 times the standard deviation of the residuals – in other words, we shift the regression line up by 2 times the standard deviation of the residuals, and then we shift it down by 2 times the standard deviation of the residuals.

Finally, any points that are outside of the gray envelope shown here are outliers. Probably.

Slide 5 – Identifying Outliers: Statistically

The procedure for identifying outliers statistically follows in much the same way.

First, we have to find the residuals, y minus y hat.

Then, we square the residual values, take the sum of the squares and divide by n minus 2, then take the square root of this value. Now, we have the standard deviation of the residuals.

Next, we take the absolute value of the residual values that we calculated in the first step, and compare this to two times the standard deviation. Any residual whose absolute value is larger than two times the standard deviation is probably an outlier, and we will need to look further at it.

This is one way to identify outliers, but it's not the only way – most of the other methods work in a similar way, though – we calculate the residuals, compare each of the residuals to some other value, and if they exceed that value we identify them as a potential outlier.

Slide 6 – Handling Outliers

Okay, so now that we've identified a potential outlier in the data, what do we do about it?

There are any number of reasons why a value might appear to be an outlier – we need to look closer at it and see what it is that makes this value an outlier.

For example, it could be because we've made a mistake in measuring or recording a value – maybe we forgot a decimal place, or wrote down the wrong number. If this is the case, then we should try to correct the error, or remove the erroneous value.

A value might also appear to be an outlier because of an actual difference, though – maybe while we were studying the weights of apples at the supermarket, we accidentally included an orange or a watermelon in our data.

Or, maybe the value appears to be an outlier, but it's just an extreme value – there are any number of valid reasons why extreme values might appear in our data, and just because something is an extreme value doesn't mean that we shouldn't include it in our study.

In other words, what we do to handle the outlier is going to depend on what it is that makes the value an outlier to begin with.

Slide 7 – Handling Outliers: A Flow Chart

Now, let's put this all in a handy flow chart, and let's say we've identified a potential outlier in our data. As discussed on the previous slide, we need to determine why it is that this value is a potential outlier.

For the first case, let's say that the value is an outlier because it's a measurement or data entry error. If we can correct the error, either by repeating the measurement or fixing the typo, then we should do so! Then, we should re-run our analysis with the corrected data.

We might not be able to correct the error, though. Maybe we no longer have access to the equipment that we need, or we don't have access to whatever it is that we're measuring to check. If we're not able to correct the error, then we should delete the observation and record why it is that we deleted it. Then, we re-run the analysis with the corrected data.

On the previous slide, we also saw that sometimes, we might have an observation of something that's not part of the study population – for example, we accidentally included an orange in our observations of the weights of apples. If the observation is of something that's legitimately not part of the study population, we should delete it and record why, then re-run the analysis with the updated data.

The other reason we've discussed for why we might have a potential outlier is because of natural variability – the potential outlier might be a legitimately extreme value. If that's the case, then we just have to accept that the world is strange and full of wonders, and that sometimes there are going to be extreme values that confound our analysis.

Now, you might say "that's fine and all, but this one outlier value is ruining my model results! If I delete this value, my results look way better!" And that's a commonly-heard complaint. But it's also not a valid reason to delete the observation. The world is full of natural variation, and it doesn't always fit neatly into a mathematical model. We just have to deal with this fact.

Slide 8 – Summary

In this lesson, we've discussed how outliers are observations or data values that lie far away from the rest of our values – in the context of linear regression, they're points that lie far away from the regression line.

We looked at two ways to identify outliers in the data, both graphically and statistically. The basic approach is the same – identifying regression outliers involves calculating the dispersion of the residual values and finding values that are much larger than the "average" value.

Once we have identified potential outliers, though, we're not done – we have to look closer to determine how to handle them. This might mean that we correct measurement or data entry errors, remove observations of things that aren't part of the study population, or it might mean that we just have to accept that our data don't always fit into the nice mathematical box that we want them to.

Above all, though, remember that "it makes my model work better" is never a valid reason for removing an outlier from the data.

Slide 9 – Additional resources

You can read more about the topics we've discussed here in the textbooks – Illowsky and Dean, Chapter 12.6. I've also included links to two articles about finding and handling outliers in data, which provide a deeper discussion of the ideas that we've covered here.

That's all for this lesson – I hope you found it interesting, and if you have any questions, please don't hesitate to e-mail me or post in the discussion forum on blackboard. Bye!