OUTLIERS

An outlier is a data value that is far from the bulk of the data. It is unusual.

OUTLIERS ARE INTERESTING

Outliers provide valuable insights. In many ways they are more important than the “common” data.

• In production, outliers identify conditions where the production process is “out of control,” indicating need for immediate remedial action.
• A famous engineer/statistician, Fred Wood, reportedly claimed that many of his patents were the results of locating outliers that were “good”, uncovering their mechanisms, replicating the effects, and then patenting the results.
• In health care, people can be occasionally allergic to certain drugs, foods, or therapies, resulting in outlier data values. These type of people need to be identified so that you don’t make them sick or kill them inadvertently.
• Nassim Nicholas Taleb, in his book *The Black Swan*, notes that most of the money that changes hands in financial markets is caused by outliers. The “80 – 20” rule states that 80% of the activity is caused by 20% of the potential causes; he argues for a “99 – 1” rule rather than an “80 – 20” rule, where 99% of the change of money is caused by 1% of the financial instances (these 1% are the outliers).

Thus outliers should ALWAYS be IDENTIFIED and DISCUSSED. They should NEVER be ignored, or "swept under the rug."

OUTLIERS in Y|X space and Their Effects

An outlier in Y|X space is a data value yᵢ that is either unusually small or unusually large considering the given cohort of observable data where X=xᵢ. As always, the subscript “ᵢ” indicates observation label, typically a row in your data frame.

The starting point of regression is to state a model p(y|x), which describes the distribution of possible outcomes y given a particular outcome X=x. The problem with outliers in Y|X space is that if you assume a distribution form for p(y|x) that does not produce outliers (eg, normal), then the maximum likelihood (ML) estimation procedure will have a hard time accommodating the outliers, simply because the model does not expect them to occur.

For example, if you assume normality, and if there is a yᵢ observation that is 6 standard deviations from the regression line, then the estimation procedure (typically OLS, which is the same as ML when you assume normality) will be confused. It does not expect such an extreme observation. So instead, the estimation procedure forces the estimate of the E(Y|X=x) function (often assumed to be a straight line) closer to the outlier yᵢ, so that the outlier does not look so unexpected. This results in an estimated regression function that is pulled toward the outlier yᵢ, far from the main body of the data where you expect the regression function to be. Again, the estimation procedure does this because it is not
expecting the outlier, and instead gives an estimated function where the outlier does not look quite so unexpected.

On the other hand, if you choose a model \( p(y|x) \) that does expect outliers, and use ML, then outlier observations are expected by the ML estimation procedure, and the estimates are not badly influenced. Rather, since the model expects occasional outliers, the estimation procedure sees no need to force the estimated regression function closer to the outlier.

A great example of this was discussed in class. Please review the document “Why do assumptions matter”, where the ML estimation using normal distributions was shown to provide bad results, while the ML estimation using the Cauchy distribution provided sensible results in the presence of an outlier.

OUTLIERS IN X SPACE and Their Effects

An outlier in X space is a data value \( x_i \) (which is a list of values \( x_i = (x_{i1}, x_{i2}, ..., x_{ik}) \) in multiple regression with k predictors) that is far from the rest of the X data values. In multiple regression, these points \( x_i \) are in multidimensional space, so you need to visualize a point \( x_i \) in space that is far from the other X points in space, like a wayward star that has wandered far from its galaxy.

First, note that the model \( p(y|x) \), being conditional on the particular values of the X variable(s), is agnostic concerning the distributions of the X variable(s). There is no assumption that the X variable(s) have any particular distributions, normal or otherwise. Sometimes they are not even random, as in the case of designed experiments, where you yourself set the X variable values, then observe the resulting Y. For example, if you are making a food product, you might set the amount of water, the amount of salt, the length of time cooked, and other variables to particular values, make the product, then observe Y, the taste of the resulting product. Then you would repeat the experiment for other settings of the X variables, observe the taste Y, and try to identify the best combination of X variables for optimal taste (Y). In this example, the X data are not random at all because they are fixed and known in advance of the data collection.

Despite the fact that the distribution of the X data is of no concern for regression analysis, outliers in X space can still cause problems in the estimation procedure, no matter which procedure you use (OLS, ML, etc.) The problem is this: If the \( X_i = x_i \) observation is an outlier, then by definition there is little or no data near \( x_i \). Thus there is little information about the nature of \( p(y|x) \) near \( x_i \). Nevertheless, this outlier data point will be highly influential in determining the estimates of \( p(y|x) \), because of the constraints you assume on \( p(y|x) \). For example, suppose you assume linearity. This constraint forces the means to pass through a single straight line through the range of the x data. If there is an outlier x value, say \( x_i \), the estimation procedure will give you an estimate of the mean function that is highly influenced by whatever the value \( y_i \) happens to be. It will pull the line upwards if the \( y_i \) value is very high, and it will pull the line downwards if the \( y_i \) value is very low. This is all fine and good if the linearity assumption is precisely true – in this case the outlier provides valuable information. However, you know that \( E(Y|X=x) \) is never precisely linear. So if the true regression function \( E(Y|X=x) \) curves dramatically as it approaches the outlier, the estimated regression line will be pulled towards the outlier. So again, as in the case where there is an outlier in \( Y|X \) space, an outlier in X space results in an estimated function that does not represent the bulk of the data well; instead, it attempts to accommodate the outlier.
And again, this is all fine and good if the assumed form of the mean function \( E(Y|X=x) \) applies throughout the range of the observed data. However, when the data value \( x_i \) is an outlier, you have little data in the neighborhood of \( x_i \) to assess whether the assumption is reasonable. So the excessive influence of this outlier is untrustworthy in this case.

To summarize, an outlier \( x_i \) in X space forces the estimated regression function toward the \( y_i \) value corresponding to the outlier \( x_i \). This is fine when your assumed form of \( E(Y|X=x) \) is correct throughout the range of the \( X \), including near the outlier, but is ultimately highly troublesome because (i) you know your assumed function form for \( E(Y|X=x) \) is nearly always incorrect, and (ii) since the point is an outlier, you have no data to assess whether your assumed function form of \( E(Y|X=x) \) is even reasonable in the vicinity of the outlier.

WHAT TO DO

1. Identify outliers using graphical and or numeric methods.
   a. Graphical methods: \((X,Y)\) scatterplots, residual plots, normal q-q plots, multivariate plots (such as principal component plots discussed in the multivariate analysis class), boxplots, etc.
   b. Numeric methods: Z-values (for all of the Y and X variables), studentized residuals, leverage values, Cook’s D statistics, etc.

2. Discuss the outliers.
   a. What information content about the subject of your study do they provide? Outliers are interesting. They may be the best information you can get from your study!
   b. Identify whether they are mistakes; if so, fix the mistakes or delete.
   c. Identify whether the outliers represent unusual circumstances that differ dramatically from your study objectives, which are not part of the data-generating processes you are modelling. For example, the outliers are from Oklahoma, and you intended to study Texas, then just delete them, but explain this clearly. But note: Deleting a value because it is “unusual” does not fall into this category. You are studying \( p(y|x) \) for a given study objective, and if outliers are part of the process for this objective, then they are indeed part of the data-generating processes that you are studying.

3. Just use OLS but correct the inferences. If you really want to estimate the means of the distributions \( p(y|x) \), then OLS will do that. The estimates will be pulled towards the outliers as mentioned above, but that is what outliers do to averages always, regression or just ordinary data. But because the distributions are so clearly non-normal the ordinary inferences (confidence intervals, hypothesis tests) are not valid. The bootstrap provides better intervals and tests in this case. This is a viable option if (i) you really do want to estimate the mean of the distribution, not the median or some other measure of the center, and (ii) if the outliers are not too extreme.

4. Use log or other transformation and proceed with the standard models. If your assumed model (e.g., classic regression, which assumes linearity, normality, and constant variance) produces data that look like your transformed data, then this is a viable solution.
5. Choose a reasonable model $p(y|x)$ that expects outliers, and use maximum likelihood. This is the gold standard. It is optimal with the correct choice of $p(y|x)$, and it is a viable solution if your choice of the model $p(y|x)$ is reasonable. (Recall: A reasonable model is one that produces data that look like your observable data.) An additional advantage is that ML methods are the first step to Bayesian analysis, which can provide further benefits.

6. Use quantile regression. This is not as good as ML, but unlike ML it has the advantage that it does not require you to assume a distribution form for $p(y|x)$. It is another viable solution to estimating $p(y|x)$, but it still assumes correct functional specification of the quantile functions. For example, if you model the median of $Y$ as a function of $X$, then you must assume that the medians fall exactly on a straight line (or whatever other function you use). It is not the gold standard that is MLE, because it does not make use of the information about the specific distribution form, thus quantile regression is inefficient relative to MLE. But it is nevertheless a viable solution.

7. Some people like to use Winsorization-type methods, but this is just a bad idea. These methods simply throw out extreme data values or replace them with something less extreme. On its face the idea is absurd: How can you estimate a distribution $p(y|x)$ when you throw out or change the extreme values? You are missing all the information in the tails. If your interest really is in the means, then the outliers are an important part of the estimates. Trimmed (Winsorized) means do not have a clear-cut use or interpretation, unlike the ordinary mean and the ordinary median.

To be fair, these methods can produce OLS estimates with more accuracy than OLS estimates that leave the outliers alone. But Winsorization gives biased estimators and inaccurate inferences. Further, there is an implicit assumption made by Winsorizers that by deleting outliers, they are removing “incorrect” data. But outliers are not necessarily incorrect. And even if they are incorrect, it is very likely that some non-outliers are also incorrect, and they will not have solved the problem by removing just the outliers. More to the point, simple removal of the extreme data completely misses the point of regression, which is to model $p(y|x)$.

The argument that Winsorized OLS estimates beat non-Winsorized OLS estimates, while correct, is not a compelling argument. Maximum likelihood and quantile regression estimation methods are better, readily available, and provide consistent estimates of the models $p(y|x)$ in outlier-prone and other situations. So Winsorization-type methods should not be used.

Here is a great document that suggests you need to do exactly the opposite of Winsorization when you estimate $p(y|x)$: You need to throw away the good data and just analyze the outliers!