Residual and Outlier Diagnostics

The purpose of this handout is to summarize the diagnostic techniques we have learned for detecting violations of the standard regression diagnostics and the presence of outliers or high leverage points.

Regression Assumptions

When we fit a standard simple or multiple linear regression model we make a set of assumptions. These assumptions have to do with the error terms in our model, the ϵ's:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

Remember that the error terms tell you about the variation in Y that is NOT attributable to your predictor variables. If you have the right model then those basically the error terms should just represent random noise or measurement error. The four main assumptions we make are that the ϵ's are

- **Mean 0:** The average value of the error terms should be 0 **for all combinations of the X values.** This means that the points are centered about the regression line (or plane) at all points, suggesting that you have the correct model shape. This also means that our estimates of the regression coefficients (the β's) will be unbiased.
- **Independent:** The errors should not show a systematic pattern that would imply a relationship with each other, the predictor variables, the fitted values, the order in which the data were collected, etc. Basically the presence of such a pattern would mean that there was more that could be added to the model to improve the fit. You can think of independence as being like having a i.i.d. or simple random sample after adjusting for your X values.
- **Constant Variance:** The errors should be **homoscedastic** or have the same spread about the regression line/plane at each combination of X values. That common standard deviation is estimated by the RMSE. Constant variance makes it much easier to obtain standard errors for the regression parameter estimates. Violations of this assumption are referred to as **heteroscedasticity**.
- **Normally Distributed:** At each combination of X values the errors should have a normal distribution, i.e. more errors near 0 and fewer large ones in a symmetric pattern. This allows us to use t and F tests for inference about our model parameters. It also results in the least squares estimates being the maximum likelihood estimates for the regression coefficients. Recall that the maximum likelihood estimates of a parameter are the values most likely to have produced the observed data.

Regression Diagnostics

To test the regression assumptions listed above we use **residuals.** The residuals can be thought of as the "estimated errors." They are calculated by taking the distance between each data point and the fitted regression line:

$$e_i = Y_i - \hat{Y}_i$$

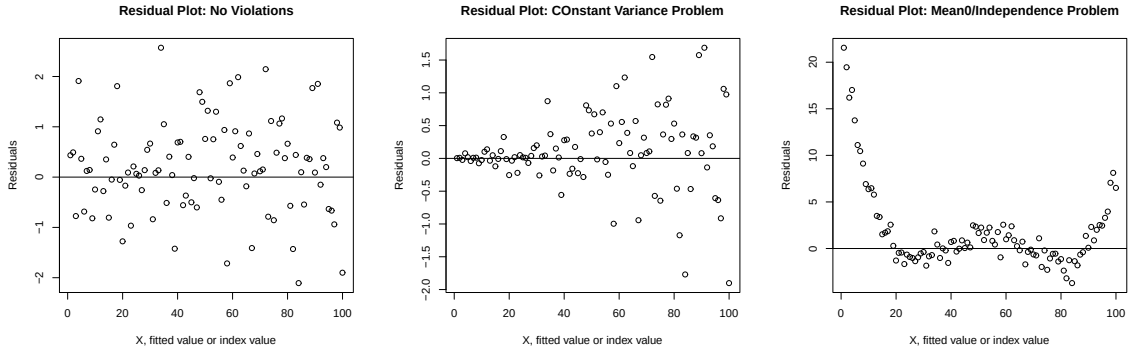
where Y_i is the observed or actual Y value and $\hat{Y}_i$ is the value predicted for that subject by the regression equation. If our error assumptions are correct then the residuals should also be mean 0, constant variance, normally distributed and show no systematic patterns. We check this with several diagnostic plots:

- **Histogram:** This plot is used to check the normality assumption. A histogram of the residuals should look approximately bell-shaped. It's usually easy to get an idea whether the residuals look symmetric about 0 and have a hump in the middle. However it is hard to tell by eye if the rate at which the hump tails off too either side really fits the requirements of the normal distribution, especially if the data set is small. Therefore there is a second plot that is also used to check normality....

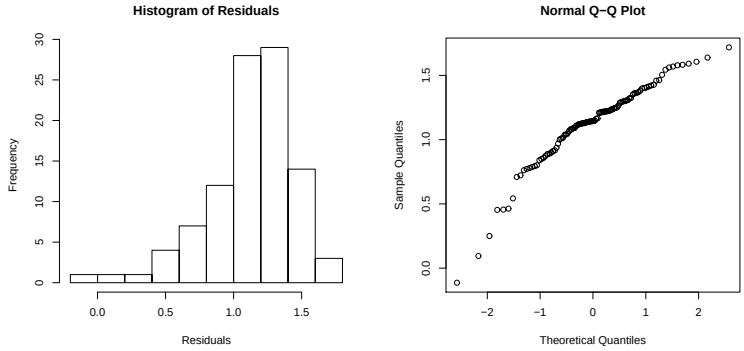
- **Quantile-quantile or normal probability plot:** This plot is also used to assess normality. The basic idea is that you compare the residuals you actually got with the ones you would have **expected** to get if the errors were really normal. You can do this essentially by calculating the appropriate percentiles of a normal distribution for the number of data points in your sample. You plot the sorted residuals from your regression model against the expected error values. If the normality assumption is correct these two sets of values should be the same and the resulting plot will look like a straight line.

- **Residual Plot:** The other three regression assumptions are checked with a **residual plot**. This plot has the residuals on the Y axis and on the X axis has any of the predictor variables, the fitted values from the regression model ($\hat{Y}$'s), or some other index variable. If the assumptions are met then this plot will look like random scatter. The points will line in a band, centered around the $Y = 0$ line. The band will be the same width all the way along and there will not be any discernible patterns.

The following plots show examples of a valid residual plot, a residual plot with a constant variance violation (the residuals are getting larger as the index value increases) and a plot with a mean 0 and independence violation. In the latter case it appears that some sort of curved model would fit the data better. Note that typically the mean 0 and and independence assumptions go together.



The next two plots show evaluation of the normality assumption, first with a histogram and second with a qq plot. We see that the distribution of the residuals is very skewed and as a result the qq plot does not follow a straight line. This suggests that in this case the normality assumption is violated.



Outlier Diagnostics

An **outlier** is a point that has an unusual Y value, an unusual combination of X values, or both. A **high leverage point** is a point that is on the "edge" of the data set in terms of it's X values and therefore has more **potential** to have a large effect on the data set. An **influential point** is a data point that by itself has a large effect on the estimated regression equation (i.e. the fit will be very different depending on whether or not the point is included in the data set). Such points can cause a great deal of trouble when running a regression even if you are using what is basically the right model. If you are doing a simple linear regression, outliers and influential points are easy to spot because they stick out like sore thumbs in a scatter plot. However in a multiple regression analysis they are harder to identify. There are a number of diagnostic statistics that can help you to identify such points. Some common ones include:

- **Standardized Residuals:** The standardized residuals are obtained by dividing the raw residuals by the root mean squared error: $e_i/RMSE$. The idea here is to get the residuals on a unit free scale so it is easy to tell what constitutes a "large" value. It turns out that the standardized residuals have roughly a t distribution so that values above 2 (small data set) or 3 (large data set) in absolute value are considered "large."

- **Leverage:** The leverage represents a point's **potential** to influence the line although not all high leverage points are influential. Leverage has to do in large part with how far the point is from the center of the data set. Recall that our estimated regression equation is

  $$\hat{Y}_i = b_0 + b_1 X_{i1} + \cdots b_p X_{ip}$$

  Note that the b's are functions only of the observed X and Y values. It turns out that we can rewrite the estimated regression equation as

  $$\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j$$

  where the h_{ij}'s are functions purely of the X's. The h_{ij}'s make up what is called the "hat matrix" because the take the original Y values and put the "hat" on them that indicated the fitted value. The diagonal elements of the hat matrix, h_{ii} which say how the observed value Y_i contributes to it's own fitted value is called the **leverage** of the ith point. What counts as "high leverage"? It turns out that after lots of algebra $0 \leq h_i i \leq 1$ where 1 represents the highest possible leverage and 0 represents no leverage. You can show that in a model with n data points and p predictors that the average leverage should be $(p+1)/n$. Because of this the most common rule of thumb is to declare values two ttimes this value as high leverage, i.e. $h_{ii} \geq 2(p+1)/n$. However it is often enough simply to look for points that have a much higher leverage than the rest.

- **Studentized Residuals:** The studentized residual is like the standardized residual except it adjusts for the leverage of the point. Specifically, the studentized residuals are given by

  $$\frac{e_i}{RMSE\sqrt{1-h_i i}}$$

  There are two versions of the studentized residual. The first is called the **internally studentized residual** and uses the RMSE from the model fit with all the data points. The **externally studentized residual** for the ith subject uses the RMSE from the model fit with the ith data point removed. Essentially the idea of removing the point is that if it's an influential point it will effect the fit a lot and so if you have it in the model it's residual won't be as big as it "should" be since it will have pulled the line towards itself. Studentized residuals have approximately a t-distribution with $n-p-1$ degrees of freedom so the usual rule of thumb is that values above 2 or 3 in absolute value are large, just as with standardized residuals.

- **DF Fits:** As it's name implies, this diagnostic has to do with how much the fitted or predicted value, $\hat{Y}_i$, changes when the ith point is removed from the data set. The formula is

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(-i)}}{RMSE_{-i}\sqrt{h_{ii}}}$$

  where the "-i" subscripts indicated the values are taken from the model with the ith point deleted. What counts as a large DFFITS value? For small data sets the cutoff (in absolute value) is usually taken to be 1 while for larger data sets $2\sqrt{p/n}$ is used.

- **Cook's Distance:** Cook's Distance is like DFFITS except that it measures the collective influence of the ith point on **all** the fitted values rather than just the fitted value corresponding to i. The formula is rather messy so we won't bother with it. The important thing to note is that as usual large values indicate that the point is unusual or influential. According to our text book values above 1 are bad and values above 4 are really bad. Personally, my favorite approach is to look and see if there are a couple points with much bigger Cook's distances than the others. If you want to be fancier it turns out that you can compare D_i to an F distribution with p and $n-p-1$ degrees of freedom and find the percentile of the distribution to which D_i corresponds (this is like finding $P(F \leq D_i)$ from the F table). It is thought that values below the 20th percentile are OK while values above the 50th percentile are bad. I will illustrate this approach in the homework solutions for those who are interested but you usually don't need to worry about it in practice.

- **DF Betas:** This diagnostic focuses on the influence on the regression coefficient for a particular predictor variable, X_j. Specifically, $DFBeta_{ij}$ says how much b_j changes when the ith point is left out of the model. Thus for each point you get a DFBeta for each of the regression coefficients including the intercept. Generally values above 2 are considered large for small data sets and values above $2/\sqrt{n}$ are considered large for bigger data sets.

Having obtained all this diagnostic information, what do you do with it? If a point is unusual by all or most of the above measures then you should check and see if there was a data entry error or something unusual happened when the point was collected that would justify removing it. If the point seems legitimate you usually report results with and without it and say why you think one of those solutions is more appropriate in the context you are studying. What you can't do is simply remove it and pretend it never existed!

It is important to note that what counts as an outlier depends on the model you fit. A point that looks like an outlier with one model may look perfectly fine with another model. It is therefore usually a good idea to figure out the shapes of the relationships between Y and your X variables **before** you decide something is an outlier.