Regression diagnostics

As is true of all statistical methodologies, linear regression analysis can be a very
effective way to model data, as along as the assumptions being made are true. For the
regression model, these assumptions include that all of the data follow the hypothesized
linear model; that is, there aren’t any cases far off the regression line (outliers). In addition,
cases that are isolated in X–space (leverage points) are also problematic, as they can have
a strong effect on estimated regression parameters, measures of fit, and so on. Once you’ve
identified such a case, it’s very important to identify what it is that makes the case unusual
(e.g., $y$ is surprisingly large for the given $X$ values, the observed values for two $X$–variables
don’t typically occur together, etc.). Further, you need to try to determine what might
have happened in the random process under study that would result in such a case.

As we discussed in class, residual plots are very useful to detect outliers and leverage
points. In a Residuals vs. Fitted plot, points by themselves on the top or bottom are
outliers; points by themselves on the left or right are leverage points. In a normal plot
of the residuals, outliers show up as distinct at the bottom left (negative outliers) or top
right (positive outliers).

Still, it is sometimes the case (particularly for multiple regression data sets) that these
plots don’t identify these cases very well. For this reason, several diagnostics have been
developed to help identify unusual cases. Three of them seem to adequately cover the
many possibilities that have been suggested:

Standardized residuals

By definition, an outlier is a point whose response variable is far from where the
general regression relationship would imply. Thus, its residual should be large (in absolute
value). The standardized residual is the residual divided by the standard deviation of the
residual; that is, it is a residual standardized to have standard deviation 1. More precisely,
the $i$th standardized residual equals

$$
e_i^* = \frac{y_i - \hat{y}_i}{s.e.(y_i - \hat{y}_i)} = \frac{y_i - \hat{y}_i}{\hat{\sigma} \sqrt{1 - h_{ii}}},$$

where $\hat{\sigma}$ is the standard error of the estimate and $h_{ii}$ is the $i$th leverage value, defined
below.

Recalling that the (unknown) errors are assumed to be normally distributed, we can
see that the standardized residuals can be expected to be (roughly) standard normal. For
example, we would expect about 95% of them to be within $\pm 2$. A good guideline for
standardized residuals is that a case with a standardized residual larger than about $\pm 2.5$ should be investigated as a potential outlier, since that would only be expected to occur randomly about 1% of the time. Note, however, that for very large samples, that means that many observations could have standardized residuals outside $\pm 2.5$ while not being outliers. Further, you should not dismiss a single observation with absolute standardized residual greater than 2.5 by saying that “1% of the values are expected to be greater than $\pm 2.5$” if its standardized residual is (say) 5, as that is expected to happen far less than 1% of the time by random chance; if a point (or points) show up as extreme in a normal plot they need to be examined.

Since standardized residuals have all of the same information in them that ordinary (unstandardized) residuals do, but are scaled appropriately, there is NO reason to ever use ordinary residuals in any diagnostic plot.

**Leverage values**

Looking at residuals doesn’t help in the detection of leverage points, since they don’t necessarily fall off the line (and can, in fact, draw the line towards them, thereby reducing their residuals). What is needed is a measure of how far a case is from a “typical” value. This is provided by the so–called leverage value. The leverage value is simply a measure of how far a particular case is (based on only the X–values) from the average of all cases, with distance being measured in such a way that the correlations between the X–variables is taken into account.

The leverage value is sometimes referred to as the “hat” value, or the “diagonal element of the hat matrix”, or the “diagonal element of the prediction matrix.” There are good reasons for this, based on its derivation. Recall the definition of the hat matrix: it is the matrix $H$ such that $\hat{y} = Hy$. Expanding this matrix equation out to the $n$ underlying equations gives the following form for the $i$th equation:

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{ii}y_i + \cdots + h_{in}y_n.$$  

Thus, the $i$th diagonal element of the matrix, $h_{ii}$, represents the potential effect that the $i$th observed target value $y_i$ can have on the $i$th fitted value $\hat{y}_i$. Since leverage points, by being isolated in the $X$–space, draw the regression line (or plane or hyperplane) towards them, $h_{ii}$ is an algebraic reflection of the tendency that an observation has to draw the line towards it. The connection between large values of $h_{ii}$ and an unusual position in
X-space is particularly clear in the case of simple regression, where

\[ h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{X})^2}{\sum_j (x_j - \bar{X})^2}; \]

the farther \( x_i \) is from the center of the data (as measured by the sample mean of the \( x \)'s), the higher the leverage.

It can be shown that \( 0 < h_{ii} < 1 \) for all \( i \), and the sum of the \( n \) leverage values equals \( p + 1 \), where \( p \) is the number of predicting variables in the regression. That is, the average leverage value is \( \frac{p+1}{n} \). A good guideline for what constitutes a large leverage value is \( 2.5\left(\frac{p+1}{n}\right) \); cases with values greater than that should be investigated as possible leverage points.

Note, by the way, that based on equation (1) above, leverage points have residuals with less variability than residuals from non–leverage points (since \( h_{ii} \) is closer to 1, resulting in a smaller standard error of the residual). This makes sense; since a leverage point is characterized by fitted value close to the observed target value, its residual is likely to be closer to zero. Another way to see this is from the result

\[ h_{ii} + \frac{e_i^2}{\hat{\sigma}^2(n-p-1)} \leq 1. \]

This shows that as \( h_{ii} \) gets closer to 1, \( |e_i| \) gets closer to 0.

You might hear the distinction between “good” leverage points and “bad” leverage points. A “good” leverage point is one that is not an outlier; that is, its target value is consistent with the regression relationship implied by the other observations. A “bad” leverage point is one that is also an outlier. The following picture shows the two types:
The point in the upper right is a “good” leverage point, while that in the lower right is a “bad” one.

What makes a “good” leverage point good? It comes from the fact that the variability of the fitted regression coefficients is inversely related to the variability of the predictor variables. For example, in a simple regression,

\[ V(\hat{\beta}_1) = \frac{\sigma^2}{\sum_j (x_j - \bar{X})^2}. \]

Thus, adding a leverage point to a data set increases the denominator above, and hence results in more accurate estimation of \( \beta_1 \). Since the “good” leverage point has the “right” value of \( y_i \), the regression coefficients haven’t been moved away from where they should be, and are more accurate.

I’m not a big fan of this idea. Examination of the plot above shows that a “good” leverage point can become a “bad” one if its position is changed only slightly. Also, the apparent increased precision of the regression coefficients seems to be artificial to me, since it’s being driven by only one observation in the sample. I believe that since all leverage points are potentially bad, they should all be treated with great caution.

**Cook’s distances**

A different way to look at the unusual case problem is to focus on the effect a case has on the regression. A case that, if it were removed, would result in a large change in the regression is an *influential point*, and obviously dangerous to leave in. A common measure
of influence (although not the only one) is *Cook’s distance*, which measures the change in the fitted regression coefficients if a case were dropped from the regression, relative to the inherent variability of the coefficient estimates themselves. Cook’s $D$ combines the notions of outlyingness and leverage in an appealing way, since

$$D_i = \frac{(e_i^*)^2 h_{ii}}{(p + 1)(1 - h_{ii})};$$

thus, observations that are outliers (large $|e_i^*|$) or leverage points (large $h_{ii}$) are potentially influential, and points that are both (“bad” leverage points) are the most influential. A value of Cook’s $D$ over 1 or so is flagging a point that should probably be studied further.

It is important to remember that Cook’s $D$ only measures one particular form of influence, and shouldn’t be viewed as the final judge of whether a point is worth investigating or not. For example, outliers that have low values of Cook’s $D$ can still have a large effect on test statistics, $R^2$, the standard error of the estimate, and so on. **A good general rule is that if any of the diagnostics flag a point, it is worth looking at it more carefully; that is, you MUST explore the effects of an outlier or leverage point on a regression even if it does not have a large Cook’s distance.**

Minitab (and all other good statistical packages) provides these diagnostics as a standard option from the regression. These values should always be determined and looked at. That can mean simply printing them out to look at; using univariate pictures of them, like histograms, stem–and–leaf plots or boxplots, can also be useful, as that might make unusually high or low values stand out more. **You should ALWAYS examine and report the diagnostics in a statistical report in this class; you can print them out, or (particularly if the sample is large) display them in observation order (a so-called index plot) to help make identification easier. It is NOT adequate to merely report that you have examined diagnostics without providing them, and it is NOT adequate to only report the diagnostics for observations that you have flagged as unusual.**

It is worth noting what these diagnostics are not so good at. Specifically, they are all sensitive to the so–called **masking effect**. This occurs when several unusual cases are all in the same region. When this happens, the diagnostics, which all focus on changes in the regression when a *single* point is deleted, fail, since the presence of the other outliers means that the regression line changes very little. The problem of multiple outliers in regression is one of the hardest problems in statistics, and is a topic of ongoing research. One approach

Note that if you have omitted an observation or observations from your data set, you’ve effectively created a new data set. What that means is that your results from the analysis with the observations included in the data are no longer valid, and you need to start from the beginning in deciding the variables you want in the model, fitting the model, checking assumptions, etc.

This is a good time to debunk an argument that you might hear regarding unusual observations and statistical modeling. You might hear people say that they are not going to omit unusual observations from their data, because all of the observations in the data are “legitimate” (that is, the observations don’t correspond to coding errors, observations that shouldn’t have been in the original sample, or other obvious mistakes), and they want to keep the data “as they really are.” This is a fundamentally incorrect attitude, as it ignores the key goal of any statistical model, which is to describe as accurately as possible what is going on in the data. Consider the following simple situation:

![Regression Plot](http://www.jstor.org/stable/pdfplus/2291266.pdf)

Say all of these data are “legitimate,” and an analyst fits a regression model based on all of the data. The resultant fitted regression line is obviously an extremely poor representation of what is going on in the data — it does not in any way describe the data “as they really
are,” because of the deficiencies of least squares regression modeling (and its sensitivity to unusual observations). Since a different regression method that was insensitive to unusual observations (a robust regression technique) would lead to a completely different fitted regression (one not affected by the unusual observation), the issue is not whether the data are legitimate, but rather whether the description provided by the statistical model is legitimate. It is apparent that a much better description of what is going on in these data would be to report and discuss the one unusual observation, and then fit the model to the remaining data points, providing a good description of what is going on in the bulk of the data.

You should not appeal to the “Unusual Observations” table produced by Minitab to identify unusual observations (or even pay any attention to it), as it uses a cutoff for outliers that is in my opinion too low, and a cutoff for leverage points that is in my opinion too high. Further, the table does not give any information about the relative values of diagnostics (a diagnostic could be below an arbitrary cutoff but still noteworthy if it is much larger than any other value), and says nothing about influential points.

Minitab commands

To save regression diagnostics when performing a regression, click on Storage, and then click on Standardized residuals, Leverages, and Cook’s distance. Note that you would presumably also create residual plots at the same time; when doing so, you should standardized residuals, rather than ordinary residuals.

An index plot of a variable is actually a time series plot. Click on Graph → Time Series Plot (Simple). Enter the saved diagnostic variables under Series:: and click Okay.
A brief detour into the foundations of statistics

One of the thorniest issues in the philosophical underpinnings of statistics (what is often called the foundations of statistics) is the question of whether to perform statistical inference conditionally or unconditionally. What do we mean by (un)conditional inference? One way to describe the difference is as follows: Conditional inference uses only the data we actually have, while unconditional inference also uses the data we could have had.

Without realizing it, you’ve already been exposed to both of these kinds of inference mechanisms. When we construct a confidence interval for the true mean of a population, we appeal to a “repeated sampling” mechanism to define what we mean by confidence; that is, we refer to data (and resultant confidence intervals) that we haven’t actually seen to understand what confidence means for the given data set and interval. Similarly, tail probabilities refer to the sampling distribution of a test statistic, which is based on data we haven’t actually seen. Thus, these are examples of unconditional inference.

On the other hand, regression is in many ways an example of conditional inference. The theory of regression assumes that the predicting variable values are taken as fixed and known, even when we know that they are actually results from some random process. We get around this by conditioning on the observed values; that is, we only use the (predictor) data we have, rather than the data we might have gotten. You can see this most clearly, perhaps, in the context of prediction: we ask “What is the predicted $y$ for this particular value of $x$?”, not “If I took a random observation from the population, without knowing the actual observed value of $x$, what is the predicted $y$?” (our best answer in the second case is simply $\bar{y}$, which doesn’t use the regression relationship at all). On the other hand, regression is not completely conditional; for example, the sample correlation coefficient is an estimate of inherently unconditional number,

$$
\rho = \frac{E\{[X - E(X)][Y - E(Y)]\}}{SD(X)SD(Y)},
$$

(this is unconditional, since it is based on the distribution of $X$ in the population, not just the observed values of $X$).

Why do I bring this up now? Because this fuzziness between conditional and unconditional inference gets worse as we get into more complex situations. For example, the concept of leverage is unconditional — it only makes sense to talk about “unusual” predicting values relative to some distribution of where they’re “supposed” to be. When we talk about designed experiments and analysis of variance, it’s natural to view things
conditionally (there is no distribution to the predictors, since their values are chosen deliberately), but many analysis of variance models are observational rather than designed (putting us back in the same situation). The problem culminates in this class in the modeling of group membership later on, where two competing methods are each optimal (one conditionally, and one unconditionally). Ultimately, you don’t need to involve yourself in this philosophical argument, but you should at least recognize these two different ways of thinking about inference.