Multiple Regression "Diagnostics"

Basics

The basic tools are the graphical ones familiar from simple linear regression, which we have been using already this semester:

- **distribution of each independent variable** (boxplot, histogram, etc.):
  look for outliers, which will have high leverage and therefore may be influential; normality is not relevant
- **distribution of residuals** (boxplot, histogram, normal QQ plot, etc.):
  look for outliers, which could be influential (or inflate MSE), and for non-normality (judged relative to the sample size)
- scatterplots of **residuals *vs*. each independent variable** and ***vs*. fits**:
  look for nonlinearity, uneven variance, outliers in Y direction (large residuals), high leverage points (extreme in X-axis direction)
- scatterplots of residuals *vs.* time, order, location, etc.:
  look for trends, which would indicate non-independence
- scatterplots of residuals *vs.* any other available variables:
  look for trends, which would suggest overlooked relationships

In multiple regression it also is useful to look at scatterplots of the various independent variables plotted against each other. These can show observations that have unusual **combinations** of the independent variables, which will give them high leverage. Partial-regression plots (next below), however, are better for this.

Formal Tests?

Tests for **normality** are not useful: they are powerful when normality isn't important (large *n*, so Central Limit Theorem applies) and weak when it is (small *n*).

Lack-of-fit test, when possible, can be very helpful.

There are tests for certain kinds of non-independence, e.g. serial autocorrelation (Durbin-Watson test).

There are various tests for unequal variance: Breusch-Pagan, modified Levene's. These test the null hypothesis of equal variances and are more powerful for detecting violations when the sample is large. These are reasonable to use when the variances appears to be correlated with one of the independent variables or the fits, though I usually don't use them, relying instead on informal graphical methods.

Modified Levene's test

Levene's test can be used in regression as follows:

(1) Divide the dataset into two parts based on the value of the independent variable with which the variance appears to be correlated, separating the small values from the large values.

(2) For each of the two parts of the dataset, calculate $\tilde{e}_j$ the median of the residuals in subset *j*.

(3) For each observation, calculate $d_{ij} = \left|e_{ij} - \tilde{e}_j\right|$ (with e_{ij} being the *i*th residual in subset *j*).

(4) Conduct a 2-sample *t* test on these d_{ij} (see text for specific formula).

Breusch-Pagan test

This tests specifically is for (or assumes) a linear relationship between the variance of *Y* and one of the *X*s. To perform it,

(1) Regress *Y* on the *X* that the variance is thought to be related to; get the SSE for this regression.

(2) Regress the squared residuals, e_i^2, on the *X*; get the SSR for this regression.

(3) Calculate the test statistic $X^2_{BP} = \dfrac{SSR/2}{(SSE/n)^2}$.

(4) Compare this statistic to critical values for a chi-square distribution with 1 degree of freedom (χ^2_1).

Partial-Regression Plots

These plots (also called leverage plots) show the "**partial**" relationship between the response variable and a particular explanatory variable as it is in the multiple-regression model, i.e. after taking into account the relationships of these two variables with all the other explanatory variables in the model. The plots therefore are useful in assessing many aspects of the model and detecting important kinds of problems.

Construction

The partial-regression plot for independent variable X_j (out of a total of k independent variables) is produced as follows:

(1) Regress X_k on all the other independent variables, and store the residuals $R_{i;Xk}$.

(2) Regress Y on all the independent variables except X_k, and store the residuals $R_{i;Y(-Xk)}$ (where the notation *Y(-Xk)* indicates that *Y* has been regressed on all the *X*s but X_k).

(3) Plot the residuals from (2) against those from (1): $R_{i;Y(-Xk)}$ [on vertical axis] *vs.* $R_{i;Xk}$ [horizontal axis].

Assessing the form of the partial relationship

The form of the relationship seen in the partial-regression plot is the form that is relevant in the multiple-regression model. It therefore can show **nonlinearity** and thus perhaps the need to either transform that particular independent variable or add polynomial terms in it. It also can show **uneven variance** at different values of this independent variable, which might suggest the need to transform the response variable or use some more advanced remedy.

The slope of the linear relationship seen in this plot will be the coefficient for this independent variable in the multiple regression model. (Thinking about this helps clarify the meaning of the parameters in a multiple regression model, how the meaning of a parameter depends on what other terms are in a model, and thus how confounding can occur.)

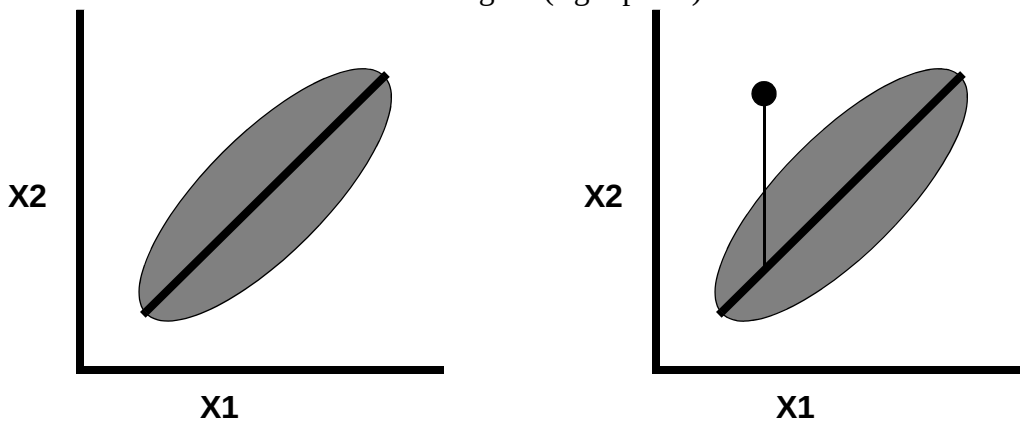
If you were to do a simple linear regression on the relationship in this plot, the regression sum-of-squares (explained variance) and error sum-of-squares (remaining unexplained variance) for this simple linear regression would equal the added-last SS for this independent variable, and SSE, in the multiple regression. Similarly, the correlation coefficient r for this partial-regression plot equals the partial correlation between X_k and Y.

Assessing leverage

An observation which has a value for independent variable X_k which is unusual given its values for the other independent variables will have a large (positive or negative) residual in step (1) of the process above for constructing the partial-regression plot. It therefore will be far to the right or left in the partial-regression plot.

Such observations also have high leverage. So looking for points far to the side in a partial-regression plot is a way to detect high leverage points.

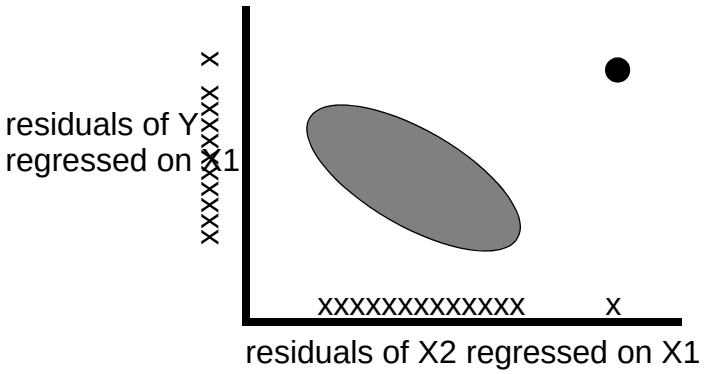
To understand this, think about the situation with only two independent variables. Compare the case when all the observations follow the same general relationship between these two independent variables (left panel in the figure below), compared with the case when one observation is aberrant in this regard (right panel). Note that the aberrant obser-



vation in the right graph is not an outlier in either *X* variable by itself and so would not be noticed in, for instance, histograms of the variables. It will, though, have a very large residual when either of the *X*s is regressed against the other. For instance, if X_2 is regressed on X_1, as shown in the figure, this point will have a very large positive residual. And it therefore will lie far to the right in the distribution of these residuals, and far to the right in the partial-regression plot for X_2 produced from these residuals.

Furthermore, this observation will have high leverage. The response variable *Y* would be on an axis coming straight out of the paper/screen in the figures above. The fitted model would go through the main cloud of points while also "trying" to fit the unusual observation. For simplicity, think of the fitted model as a plane in this 3-D space. A change in its slope perpendicular to the long axis of the main cloud of points would have a much greater effect on the size of the residual for the unusual point — which is much further from that axis — than on the residuals for the other observations. This constitutes high leverage, and would be revealed by the point being far right or left in the partial-regression plot.

If the high-leverage observation also is unusual in the *Y* direction, it would be influential: to minimize the sum of squared deviations, the fitted model would be pulled towards the unusual point (because of its leverage). "Unusual in the *Y* direction" here means have its *Y* value not follow the relationship of *Y* to the independent variables that is seen in the main group of observations. Being unusual in this way would make it likely that if *Y* were regressed on, e.g., X_1, the aberrant point would have a large (positive or negative) residual. So then in the partial-regression plot this point would be far to the right (because it is unusual in X_2 relative to the typical relationship of X_2 to X_1), and also far above or below (because unusual in *Y* relative to the typical relationship of *Y* to X_1). The partial-regression plot would like something like below (the **x**s along the axes represent the univariate distributions of the two sets of residuals):



In the scenario illustrated here, with only two independent variables, the high-leverage point can be seen in the scatterplot of the two *X*s. With more than two variables, however, an observation could have an unusual combination of the *X*s — and thus have high leverage — while not appearing unusual in any bivariate plot of two of the *X*s, just as the unusual point in the figures above was not apparent in univariate plots of either variable. A set of partial-regression plots (one for each of the independent variables), though, would detect this sort of unusual observation, by including measures of how observations follow the typical relationships among <u>all</u> the *X*s.

Leverage: "Hat Matrix"

The leverage of an observation — the unusualness of its combination of independent variables — can be quantified, as well as shown in partial-regression plots. This measure of leverage of observation *i* is denoted as h_i or h_{ii}. It is the *i*th element on the diagonal (i.e. the element in the *i*th row and *i*th column) of what is called the "hat" matrix; the "*h*" notation is short for "hat." The "hat" matrix is

$$H = X^T(X^T \cdot X)^{-1}X$$

Its name arises from its role in determined the fitted values: the "*Y*-hats":are gotten by multiplying the hat matrix by the *Y*s

$$\hat{Y} = HY$$

The diagonal elements of the hat matrix, the h_is, measure how far the values of the *X*s for observation *i* are from the overall mean of the *X*s relative to the variability in the *X*s, that is, how far the vector $(X_{i1}, X_{i2} \ldots X_{i(p-1)})$ is from $(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_{p-1})$, with distance measured in standard deviation units.This directly measures how unusual the set of *X*s are for that observation.

h_i also measures leverage: the equation above for the *Y*-hats shows that h_i measures how much weight is given to Y_i, relative to all the other *Y*s, in determining $\hat{Y}$. The larger h_i is, the closer $\hat{Y}$ will be to Y_i, and the smaller e_i, the residual for that observation, will be.

In turn, if the residual tends to be small, so does its standard error (sampling variability): $SE\{e_i\} = \sigma\sqrt{1-h_i}$. Conversely, since large leverage causes the single Y_i to

largely determine $\hat{Y}$, it causes the standard error of the latter to be large:
$SE\{\hat{Y}_i\} = \sigma\sqrt{h_i}$.

Using h_i *to assess leverage*

The h_is sum to p (the number of Xs, including the intercept), and since there are n of them, the average value is p/n. The recommended guideline is that any value greater than twice this average value, i.e. $h_i > 2p/n$, indicates high leverage.

Outliers in *Y*

The simple residuals ($e_i = Y_i - \hat{Y}$) can be modified iin various ways which facilitate determination of points which lie excessively far in the Y direction from the overall pattern of the data.

Studentized residuals

"Studentizing" is the general term for "standardizing" some statistic by dividing it by its standard error. Applied to residuals, this gives what are generally called "studentized" or "standardized" residuals:

$$\frac{e_i}{se(e_i)} = \frac{e_i}{S\sqrt{1-h_i}}$$

where S is $\sqrt{MSE}$ (the usual estimate of σ),

As noted in the previous section on leverage, points with high leverage pull the model fit towards themselves, reducing the expected magnitude and sampling variability of their residual. Studentizing compensates for this, as well as puts all residuals on a standard scaling, independent of the measurement scale of the Y variable.

Studentized deleted residuals

Because unusual observations can have a large effect on the fitted model, pulling it towards them, they in effect hide themselves: their effect on the model causes their residuals to be small. This can be prevented by the same process underlying the *PRESS* cross-validation measure: the fitted value used to calculate a given residual can be gotten by predicting that observation using the model fit obtained by excluding that observation. Hence the "deleted" part of the name refers to the observation (in effect) having been deleted from the dataset when the model was fit; the "studentized" part of the name again refers to dividing the deleted residual by its standard error.

As was noted for *PRESS*, the model doesn't actually have to be re-fit with each observation deleted in turn: the deleted residuals can be calculated from the regular residuals and the hat matrix. The studentized deleted residuals are

$$r_{(-i)} = \frac{e_i}{S_{(-i)}\sqrt{1-h_i}}$$

where $S_{(-i)}$ is S calculated with observation i omitted from the calculation.

These "studentized deleted" residuals go by different labels: in Minitab they are "TResiduals," in SAS commands they are "rstudent" while in SAS's Insight procedure they are "studentized."

The set of n studentized-deleted residuals, if the model assumptions are met, will have a t distribution with (n-p-1) degrees of freedom (i.e. one less than the error df). This can be used to test whether a given residual is significantly larger than it would be if the model assumptions were met. If this is done by picking out the largest studentized-deleted residual, a Bonferroni adjustment is needed to account for the fact that a total of n residuals were examined to pick the largest. So the largest $r_{(-i)}$ would be compared to the $\alpha/2n$ critical value of the t_{n-p-1} distribution.

Influence

"Influence" means the actual effect of a given observation on some aspect of the model fit or inference; influence usually arises from a combination of high leverage (unusual Xs) and a value of Y which doesn't follow the general trend (which would have a large studentized-deleted residual). There are several statistics which measure different aspects of influence.

Cook's D

Cook's Distance measures the impact of a given observation on all of the model coefficients: how different would they be if observation i were deleted?

$$D_i = \frac{e_i^2 \cdot h_i}{p \cdot S^2 \cdot (1 - h_i)^2}$$

D_i is considered large if it is greater than 1. More precise criteria come from comparing D_i to the F distribution with p and n-p degrees of freedom: if D_i is smaller than the 20th %ile of that distribution, observation i is not influential; if it is larger than the 50th %ile it is considered a possible problem.

(The 20th and 50th %iles are the values of F with 0.2 and 0.5 probability to the left. Statistical software can directly give the %ile of a given value, i.e. of Cook's D_i. If you are using tables of critical values, however, these do not give lower %iles. You therefore need to use the fact that the 20th %ile of the F distribution with p and n-p df is equal to the reciprocal of the 80th %ile for the F distribution with n-p and p df.)

dffits

The dffits (sometimes written as dfits) measure the influence of an observation on the fitted value for that observation. They are calculated as

$$dffits_i = \frac{\hat{Y}_i - \hat{Y}_{i(-i)}}{S_{(-i)}\sqrt{h_i}}$$

where $\hat{Y}_{i(-i)}$ is the estimated value of *Y* for observation *i* predicted by the model fit to all the data except observation *i*, so that the numerator of *dffits* is the deleted residual.

Values of *dffits* are considered large if they exceed 1, or for large sample sizes if they exceed $2\sqrt{p/n}$.

dfbetas

The *dfbeta*s measure the influence of a particular observation on a particular parameter estimate. The *dfbeta* for the effect of observation *i* on b_k is calculated as

$$dfbeta_{i,k} = \frac{b_k - b_{k(-i)}}{S_{(-i)}\sqrt{c_{kk}}}$$

where c_{kk} is the *k*th diagonal element of the matrix $(\mathbf{X}^{tr}\mathbf{X})^{-1}$. The numerator of $dfbeta_{i,k}$ is the difference in b_k between that estimated for the full data set and that estimated with observation *i* excluded. The denominator is the estimate of the standard error of b_k that would be gotten with observation *i* excluded from the analysis.

A dfbeta is considered large if it exceeds 1, or for large sample size if it exceeds $2/\sqrt{n}$.

Using these Leverage, Outlier, and Influence Measures

I find it helpful to examine these various measures of leverage, outliers, and influence graphically, in two ways. First is to look at the distribution of each, by for instance individual-value plots or histograms. In these you simply look for outliers: observations that have values of the statistic that are extreme compared to the values for most of the rest of the observations.

Second, bivariate scatterplots of any two of the statistics can be examined. I particularly like to plot the influence measures (Cook's D, *dffits*, and *dfbeta*s) against either the studentized-deleted residuals or the leverage values (h_i). In these plots again we are interested in finding observations that are extreme in one or more of the measures, for instance to see whether there are observations with unusually high leverage and influence both.

Multicollinearity: VIF

Multicollinearity is the condition of two or more of the independent variables being highly correlated. This includes the situation in which one variable is correlated with some linear combination of two or more other variables, while not particularly correlated with any of the other variables alone. Because of this latter possibility, simple bivariate correlations or scatterplots of the independent variables may not be adequate for detecting collinearity.

The most straightforward measure of collinearity which is adequate to address this situation is what is called the Variance Inflation Factor (VIF). There is a VIF for each term in the model. The VIF for the *j*th term is

$$(\text{VIF})_j = 1 / (1 - R^2_j)$$

where R^2_j is the R^2 for the regression of X_j on all the other Xs.

An individual VIF is considered large — indicative of a problem — if it is larger than 10. In addition, if the average of the VIFs is considerably larger than 1, this too is considered to indicate a problem.

VIFs do not tell how many collinearities there are, or which variables are included in them. There are other more sophisticated measures, based on eigenvalues and eigenvectors of the matrix of Xs, which provide more detailed information about collinearity, if this ever seems like it would be useful to you.