Outlying & Influential Observations

1. Added-Variable Plots for Functional Form & Outlying Observations

1. Uses of Added-Variable Plots

Added-variables plots are also called *partial regression plots* and *adjusted variable plots*.

A partial regression plots is a diagnostic tool that permits evaluation of the marginal role of an individual variable within the multiple regression model, given that other independent variables are in the model. The plot is used to visually assess

- whether a variable has a significant marginal association with Y (given other independent variables already in model) and thus should be included in the model
- the presence of outliers and influential cases that affect the coefficients of an individual X variable in the model
- the possibility of a nonlinear relationship between Y and individual X variable in the model

An added-variable plot is a way to look at the marginal role of a variable X_k in the model, given that the other independent variables are already in the model.

2. Construction of an Added-Variable Plot

Assume the multiple regression model (omitting the i subscript)

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

There is an added-variable plot for each one of the X variables.

To draw the added-variable plot of y on X_1 "the hard way", for example, one proceeds as follows:

1. Regress y on X_2 and X_3 and a constant, and calculate the predictors (denoted $\hat{y}_i(X_2, X_3)$ and residuals (denoted $e_i(y|X_2, X_3)$ for y regression

$$\hat{y}_i(X_2, X_3) = b_0 + b_2 X_{i2} + b_3 X_{i3}$$
$$e_i(y|X_2, X_3) = y_i - \hat{y}_i(X_2, X_3)$$

2. Regress X_1 on X_2 and X_3 and a constant, and calculate the predictors and residuals for X_1

$$\hat{X}_{i1}(X_2, X_3) = b_0^+ + b_2^+ X_{i2} + b_3^+ X_{i3}$$
$$e_i(X_1|X_2, X_3) = X_{i1} - \hat{X}_{i1}(X_2, X_3)$$

3. The added-variable plot for X_1 is the plot of

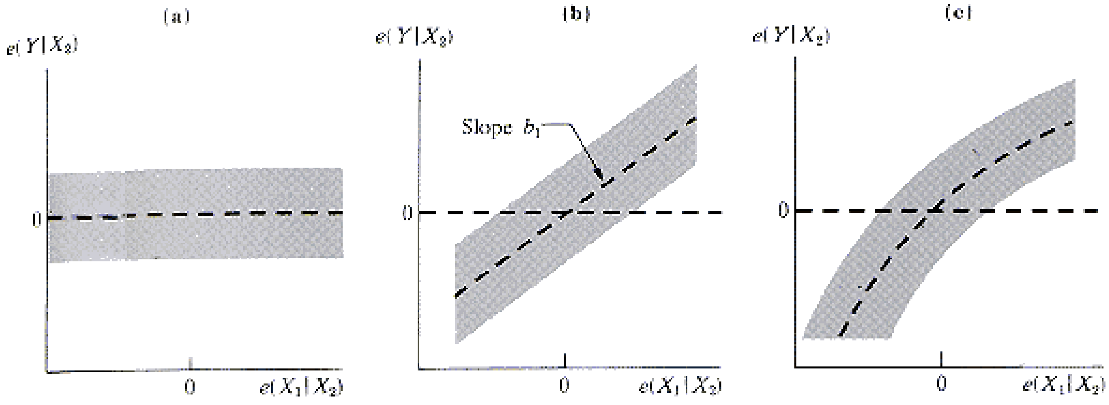
$$e_i(Y|X_2, X_3) \text{ against } e_i(X_1|X_2, X_3)$$

In words, the added-variable plot for a given independent variable is the plot of the residuals of y (regressed on the other independent variables in the model plus a constant) against the residuals of the given independent variable (regressed on the other independent variables in the model plus a constant). This can be done in SAS.

3. Interpretation of an Added-variable Plot

It can be shown that the slope of the partial regression of $e_i(y|X_2, X_3)$ on $e_i(X_1|X_2, X_3)$ is equal to the estimated regression coefficient b_1 of X_1 in the multiple regression model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$. Thus the added-variable plot allows one to isolate the role of the specific independent variable in the multiple regression model. In practice one scrutinizes the plot for patterns such as the ones shown in the next figure

FIGURE 9.1 Prototype Partial Regression Plots.



The patterns mean

1. pattern (a), which shows no apparent relationship, suggests that X_1 does not add to the explanatory power of the model, when X_2 and X_3 are already included
2. pattern (b) suggests that a linear relationship between y and X_1 exists, when X_2 and X_3 are already present in the model; this suggests that X_1 should be added to (or kept in)
3. pattern (c) suggests that the partial relationship of y with X_1 is curvilinear; one may try to model this curvilinearity with a transformation of X_1 or with a polynomial function of X_1
4. the plot may also reveals observations that are outlying with respect to the partial relationship of y with X_1

2. Three Types of Outlying Observations

- an *outlying* case is one with an observation that is well separated from the remainder of the data
- an *influential* case is one that has a substantial influence on the fitted regression function (i.e., the estimated regression function is substantially different depending on whether the case is included or not in the data set)

A case can be outlying with respect to their Y value or X value(s) and an outlying case may or may not be influential.

Diagnostic tests have been developed to identify *three types* of problematic observations:

- observations outlying in the X-dimensions, called *high leverage* observations
- observations outlying in the Y-dimension (discrepancy)
- influential observations

3. Identifying X-Outlying Observations – Leverage

In multiple regression X-outlying observations are identified using the *hat matrix* **H**.

1. Review of the Hat Matrix

We know that a n x 1 vector ŷ of estimated (predicted, fitted) values of y is obtained as

$$\hat{\mathbf{Y}} = \mathbf{HY} \quad \text{where} \quad \mathbf{H} = \mathbf{X(X'X)^{-1}X'}$$

One can view the n x n matrix **H** as the linear transformation that transforms the observations **y** on the dependent variable into their estimated values $\hat{\mathbf{Y}}$ in terms of **X**. The vector of residuals **e** can also be expressed in terms of **H** as

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

It can be shown that the variance-covariance matrix of the residuals **e**, both "true" and estimated are (pg. 204)

$$\boldsymbol{\sigma}^2\{\mathbf{e}\} = \sigma^2(\mathbf{I} - \mathbf{H})$$
$$\mathbf{s}^2\{\mathbf{e}\} = \text{MSE}(\mathbf{I} - \mathbf{H})$$

Thus the variance of an *individual residual* e_i (i.e., the diagonal element of the variance-covariance matrix $\mathbf{s}^2\{\mathbf{e}\}$ corresponding to observation i) is

$$\sigma^2\{e_i\} = \sigma^2(1 - h_{ii})$$
$$s^2\{e_i\} = \text{MSE}(1 - h_{ii})$$

where h_{ii} denotes the ith element on the main diagonal of **H**.

h_{ii} is called the *leverage* of observation i and can be calculated without calculating the whole **H** as

$$h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$$

where $\mathbf{x}_i'$ is the *row* of **X** corresponding to observations on case i.
Thus $\mathbf{x}_i' = [1 \;\; X_{i1} \;\; X_{i2} \ldots X_{i,p-1}]$ (and $\mathbf{x}_i$ is the same thing transposed as a *column* in **X**')

It can be shown that $0 \leq h_{ii} \leq 1$. A larger value of h_{ii} indicates that a case has greater *leverage* in determining its own fitted value $\hat{Y}_i$, i.e. since $\sigma^2\{e_i\} = \sigma^2(1 - h_{ii})$ it follows that the larger h_{ii}, the smaller $\sigma^2\{e_i\}$, and the closer $\hat{y}_i$ will be to y_i.

2. Using the Leverage h_{ii} for Diagnosis

Look at an index plot of h_{ii} for (an) extreme value(s) of h_{ii}

There are 2 rules of thumb for identifying high-leverage observations:

1. consider h_{ii} large if it is larger than twice the average h_{ii}. Since the sum of h_{ii} (trace of **H**) is p (=number of independent variables including the constant term), average h_{ii} is p/n. Thus flag observations for which $h_{ii} > 2p/n$

2. take $h_{ii} > 0.5$ to indicate "VERY HIGH" leverage; $0.2 < h_{ii} < 0.5$ to indicate "MODERATE HIGH" leverage.

4. Identifying Y-Outlying Observations – Discrepancy

Efforts to find the best diagnostic test for outliers in the Y dimension have produced successive improvements to the ordinary residual e_i calculated as

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

or, in matrix notation,

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$$

1. Deleted Residual d_i

A possible measure of discrepancy for case i would be the value of the residual that would be obtained if that case were not included in the regression model. To do this, we can estimate the residual for case i based on a regression that *excludes* (i.e., "deletes") case i. This is the *deleted residual* defined as

$$d_i = y_i - \hat{y}_{i(i)}$$

where the notation $\hat{y}_{i(i)}$ means the fitted value of y for case i using a regression with (n - 1) cases excluding (or "deleting") case i itself.

An equivalent formula for d_i is

$$d_i = e_i/(1 - h_{ii})$$

Observations exhibiting a large value of d_i are cases that are deviant in terms of their residuals when the regression equation is derived based on the rest of the sample. However, we need to put these values on a standardized scale.

2. Studentized Deleted Residual t_i

Since $d_i = e_i/(1 - h_{ii})$ its variance is estimated as

$$s^2\{d_i\} = MSE_{(i)}/(1 - h_{ii})$$

and its standard deviation as

$$s\{d_i\} = [MSE_{(i)}/(1 - h_{ii})]^{1/2}$$

The *studentized deleted residual* t_i is obtained as

$$t_i = d_i/s\{d_i\}$$

What is the distribution of t_i? A useful observation is that $\hat{y}_{i(i)}$ is the same thing as $\hat{y}_{h(new)}$ discussed earlier. $\hat{y}_{i(i)}$ is calculated from a regression with n – 1 cases (excluding case i) while $\hat{y}_{h(new)}$ is calculated for a new set of X values from an original regression based on n cases. It follows from the similarity with $y_{h(new)}$ that

$$t_i = d_i/s\{d_i\} \sim t(n – p – 1)$$

In other words, t_i is distributed as a Student t distribution with (n – p – 1) df. The df are (n – p – 1) rather than (n – p) because there are n – 1 cases left after case i has been deleted.

2. Using t_i in Testing for y-outliers

Knowing that $t_i = d_i/s\{d_i\} \sim t(n - p - 1)$, we can test whether a residual is larger than would be expected by chance. But we cannot just test at the $\alpha = .05$ level by looking for absolute values of t_i greater than about 2, because there are n residuals and thus n tests.

One approach to multiple tests is to use the Bonferroni criterion which divides the original α by the number n of tests. The Bonferroni-corrected critical value to flag outliers is thus

$$t(1 - \alpha/2n; n - p - 1)$$

Cases with values of the studentized deleted residual t_i greater in absolute value than this critical value are flagged as outliers.

5. Identifying Influential Cases - DFFITS, COOK, and DFBETAS

A case is *influential* if excluding it from the regression causes a substantial change in the estimated regression function. There are several measures of influence: DFFITS, COOK (Cook's distance), and DFBETAS.

1. DFFITS

DFFITS measures the influence that case i has on its own fitted value $\hat{y}_i$ a

$$\text{DFFITS}_i = (\hat{y}_i - \hat{y}_{i(i)})/(\text{MSE}_{(i)}h_{ii})^{1/2}$$

In words, DFFITS_i represents the difference in predicted values of y_i from regressions *with* ($\hat{y}_i$) or *without* ($\hat{y}_{i(i)}$) case i, the difference being expressed in units of the standard deviation of the predicted y_i (in which MSE is estimated from the regression without case i). Thus it is a standardized measure of the extent to which inclusion of case i in the regression increases or decreases it own predicted value.

DFFITS_i can be positive or negative. An equivalent formula is

$$\text{DFFITS}_i = t_i(h_{ii}/(1 - h_{ii}))^{1/2}$$

The formula shows that DFFITS_i is the product of the studentized deleted residual t_i multiplied by a function of the leverage h_{ii}. Thus DFFITS_i tends to be large when a case is both a Y-outlier (large t_i) *and* has large leverage (large h_{ii}).

Using DFFITS in Identifying Influential Cases

The following guidelines can be used for identifying influential cases

- $|\text{DFFITS}| > 1$ (small to medium data sets)
- $|\text{DFFITS}| > 2(p/n)^{1/2}$ (large data sets)

2. COOK's Distance D_i

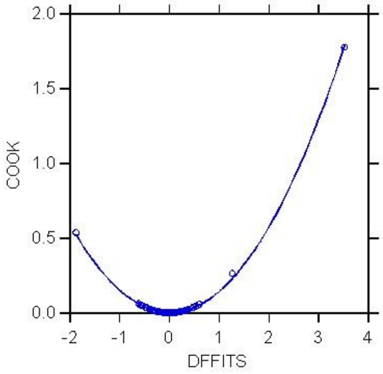
COOK's distance measures the influence of case i *on all n fitted values* $\hat{y}_i$ (not just the fitted value for case i as DFFITS). D_i is defined as

$$D_i = (\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(\mathbf{i})})'(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{\mathbf{i(i)}})/p\text{MSE}$$

where the numerator is the sum of the squared differences between the fitted values of y for regressions with case i included or excluded, respectively. An equivalent formula is

$$D_i = (e_i^2/p\text{MSE})(h_{ii}/(1 - h_{ii})^2)$$

The formula shows that D_i is large when either the residual e_i is large, or the leverage h_{ii} is large, or both. In practice, we find that D_i is almost a quadratic function of DFFITS. As D_i is (approximately) proportional to squared DFFITS it tends to amplify the distinctiveness of influential cases, making them stand out.



Using D_i (COOK) in Identifying Influential Cases

For small samples, one can use the same cutoff values for DFFITS. However, 4/n is a commonly used cutoff for Cook's D for any sample size.

3. DFBETAS

$(\text{DFBETAS})_{k(i)}$ is a measure of the influence of the ith case on the regression coefficient of X_k. Therefore there is a DFBETAS for each X_k (k = 0, ..., p-1). The formula is
$(\text{DFBETAS})_{k(i)} = (b_k - b_{k(i)})/(\text{MSE}_{(i)}c_{kk})^{1/2}$ for k = 0, 1, ..., p-1
where c_{kk} is the kth diagonal element of $(X'X)^{-1}$. (Recall the formula for the variance of b_k, $\sigma^2\{b_k\} = \sigma^2 c_{kk}$.).

NKNW's guideline is to flag a case as influential if

- DFBETAS > 1 (small to medium data sets)
- DFBETAS > $2/(n)^{1/2}$ (large data sets)

The problem with DFBETAS is that there are p of them, which is a lot of output, so it is usually easier to look at summary diagnostic measure such as D_i.

6. Summary of Diagnostics for Outliers & Influential Cases

A useful way to study diagnostics is to construct one's own summary table of guidelines and cutoff points for the diagnostics that are most often used in practice: leverages h_{ii}, studentized deleted residuals t_i, COOK (D_i) and/or DFFITS, and perhaps DFBETAS.

What diagnostics to use for routine work?

- keep in mind that influential cases are more common in small data sets (except for grossly inaccurate coding such as missing values treated as real data)
- look at index plot of Cook's D_i (or DFFITS) as a good summary measure of influence; use the test of COOK based on a comparison with an F distribution to test apparent influential observations
- look at a stem-and-leaf or box plot of STUDENT (studentized deleted residual); look especially for the observations flagged by COOK
- look at a stem-and-leaf or box plot of LEVERAGE (h_{ii}); look especially for the observations flagged by COOK
- look at partial regression plots to see how the "suspects" affect individual regression coefficients

7. Remedies for Outliers & Influential Cases

1. Get Rid of Them Outliers?

Outlying and influential cases should not be discarded automatically.

There are several situations:

- if the case is the result of recording error and such, then
  - if possible, correct the observation
  - if not, discard it
- if the case is not clearly erroneous, *examine adequacy of the model*; influential/outlying observation could be due to
  - omission of an important interaction (so that a case with high values of the two variables involved in the interaction has an extreme value of Y)
  - incorrect functional form
  - omission of an important explanatory variable; Example: several outliers are oil-exporting countries; then maybe including an indicator variable for oil-exporting countries will take care of the problem
- discard cases that cannot be accounted for by above only as last resort, reporting whether results differ with the cases included or excluded; or use robust estimation method to reduce their influence (see below)

2. Robust Regression

1. Varieties of Robust Regression Methods

Robust regression methods are less sensitive than OLS to the presence of influential outliers. There are several approaches to robust regression such as

- least absolute residuals (LAR) regression minimizes the sum of absolute (rather than squared) deviations of Y from the regression function

- least median of squares (LMS) regression minimizes the median (rather than the sum) of squared deviations of Y from the regression function
- trimmed regression excludes a certain percentage of extreme cases at both ends of the distribution of residuals
- iteratively reweighted least squares (IRLS) iteratively reweights cases in inverse proportion to its residual (standardized by the robust estimator of dispersion MAD discussed below) to discount the influence of extreme cases; IRLS refers to a family of methods distinguished by different weighting functions (such as Huber, Hampel, bisquare, etc.)