

NOTICE. Unless otherwise indicated, all materials on this page and linked pages at the blue.temple.edu, astro.temple.edu, and rbtaylor.net addresses are the sole property of Ralph B. Taylor and © 1999-2001 by Ralph B. Taylor. All users have the right to freely access and copy these pages provided that they: acknowledge the source, do not make changes on any pages, and do not charge more than copying costs for distribution.

Simple Regression:

Assumptions, Error, Residuals, Outliers, Residual Analysis

OVERVIEW

In the last note we introduced some assumptions about error terms. In this document we explore some of those assumptions more fully, and talk about how to actually take a look at your error terms. Yes, it can be ugly. But you need to know.

ASSUMPTIONS OF REGRESSION

In simple and multiple regression, there are several assumptions about the way the data are organized (These are covered more fully in the last note; also see Hamilton.) For example:

(1) Data points are distributed uniformly about the regression line. That is, the variance of the Y values is comparable at different ranges of the X values (homoscedasticity of variance).

(2) The Y data values are independent of one another. That is, the score on Y_2 [the score on Y for case number 2] is not correlated with the score on Y_1 [the score on Y for case number 1]. You are most likely to get cases correlated with one another when the dependent variable for the different cases from sequential points in time. For example, your dependent variable may be the crime rate in a city over a period of 50 years. You also can get non-independent error terms when you have clustered or grouped data, like survey respondents clustered by block, or school respondents clustered by classroom. For more information on these matters take my other course on multilevel modeling (not if I can help it you are saying.... the main url for that course is:

<http://www.rbtaylor.net/605sp00.htm>

DONT WORRY ABOUT THIS NOW BUT: If sequential cases are correlated with one another, this is called serial autocorrelation. You get information about this with the Durbin-Watson d statistic, and the first order autocorrelation. It "will tend to be small for positive autocorrelation,

large for negative autocorrelation, and somewhere in the middle for a random series" (Ostrom, Time series analysis, 1978, p. 32) "The first order autocorrelation is the ordinary Pearson correlation of a series of numbers with the same series shifted by one observation ($y_{e1}, y_{e2}; y_{e2}, y_{e3}; \dots y_{en}, y_{en-1}$)" (Wilkinson, Statistics, 1990, p. 421). The number you see indicates correlations between residuals. If you are doing geography of crime then you need to worry about spatial autocorrelation, which is a different type of autocorrelation.

(3) The true relation between Y and X is linear, rather than nonlinear or curvilinear.

You verify these assumptions by directly examining the X Y scatterplot, or by examining the plot of residuals by X values.

THE RESIDUALS YOU SEE

Unless you create a linearly transformed equation, you will have scatter about the regression line. The scatter, or the degree to which the actual y data points are off the regression line, result from error; the inability of the X variable (or variables(s)) to perfectly predict the Y variable.

When $R^2 = 1.0$, and the X variable perfectly predicts the Y variable, then of course there will not be scatter about the regression line, because each $Y_{\text{predicted}} = Y_{\text{actual}}$.

Imagine you are looking at a plot of the $Y * X$. Imagine you also have a plot of $Y_{\text{residual}} * X$. You should be able to imagine how that plot of $Y_{\text{residual}} * X$ will look different from the plot of the $Y * X$ as the correlation between X and Y increases, and R^2 approaches 1.0

Calculating residuals

Remember that for each Y data point it's error -- its residual -- can be calculated:

$$Y_{\text{residual}} = Y_{\text{actual}} - Y_{\text{predicted}}$$

Each $Y_i = M_Y + (Y_{\text{predicted } i} - M_Y) + (Y_i - Y_{\text{predicted}})$ (M=mean)

WHERE DOES ERROR COME FROM AND ASSUMPTIONS ABOUT ERROR STRUCTURES

The last note talked about assumptions about Y. Here we focus the discussion more closely and derive from those more general assumptions about Y, more specific assumptions about error terms. Again, **these assumptions are important**, and if they are violated BIG TIME you can have problems.

This error in a regression -- the portion of Y not explained - may have several components. (1) It may arise from errors in measuring the Y variable (low inter-rater reliability; low internal consistency on an index; low test-retest reliability, and so on). (2) It may also arise from causes of Y that were not included in the regression equation; they were omitted.

So here are the assumptions about the error terms. "If most of these omitted causes have rather minor impacts individually, and if they are operating almost independently of one another, then it will be reasonable to assume that the expected value of the disturbance term ($E(e)$) will be 0, and that e will be normally distributed" (Blalock, Social statistics, p. 387)

A1: $E(e) = 0$

A2: e normally distributed

It is also assumed that the variance of the error terms is relatively constant across different levels of the X variable. This is called the assumption of homoscedasticity of error variance.

A3: $\text{Var } e @ X_1 = \text{Var } e @ X_2 = \dots \text{Var } e @ X_n$

"The really crucial assumption that underlies the use of regression analysis is that X is independent of the error term" (op cit)

A4: $r_{Xe} = 0$

The assumption that X and e are statistically independent will be warranted if "the omitted causes of Y are either: (1) numerous, singly unimportant and not highly inter-related; or (2) unrelated to X in situations where one or two omitted factors predominate" (op cit)

"If one is unwilling to make such an assumption in any particular instance, one should attempt to identify the major disturbing factors that have been omitted and to bring these into the equation explicitly as additional variables." (Blalock, p. 388) In other words, omitted relevant predictors cannot be systematically related to the X variable(s).

This latter point is really important. How do we see if there is an omitted variable related to any predictor? We can't get at the variable because it's omitted. BUT WE CAN LOOK AT THE ERROR TERM. If we have an important omitted variable that is related to X then we can see a relationship between the size and/or direction of the individual residuals, and the value of X, across the cases in the analysis.

If such a relationship does obtain, then what will happen is that as scores increase or decrease on a predictor, the variance of Y (if certain conditions obtain), and the variance of the error terms will either increase or decrease as scores on one of the predictors increases. So we now have

violated the assumption of the homogeneity of variance at different levels of Y, and homogeneity of error variance at different levels of Y.

NOTE: Hamilton uses different terminology to talk about the assumptions behind error terms, but the points are the same.

RESIDUAL DIAGNOSTICS

SPSS can save a bunch of residual statistics; it also can print out some facts about the residuals. Here are some of the things it tells you about.

Residuals are important for a number of reasons. First, they tell us where a case is with respect to a regression line. In addition, we can use this information, along with other information, to decide how much "influence" a case has on a simple or multiple regression. The concept of influence is explained below. If a case has a large influence, it has a large say in shaping the unstandardized regression coefficient (or coefficients in the case of multiple regression) that are produced.

First, let's look at the features of the residuals themselves that can be saved. These are in the "SAVE" option in linear regression.

RESIDUALS: UNSTANDARDIZED. These are the raw residual scores computed according to the classic and well known formula (see last note)

$$e_i = Y_{\text{actual } i} - Y_{\text{predicted } i}$$

RESIDUALS: STANDARDIZED. Residuals are divided by the standard error of the estimate (SEE) (see last note). "Standardized residuals are often helpful in judging the magnitude of the regression outliers" (McLendon MJ (1994) Multiple regression and causal analysis. Itasca, IL: Peacock Publishers, p. 52.).

RESIDUALS: STUDENTIZED these are studentized residuals. For each case the studentized residual provides a scale-free measure of distance from the regression line (or plane or hyperplane when we get to multiple regression). To see if significant you do a single sample t-test with N-p-2 degrees of freedom. (See Darlington p. 357.)

How does it do this standardization to make things "scale free." Good question. Somehow the program comes up with a measure of the "standard deviation" for this case - how far out is it from other cases on its X scores. "Studentized residuals are corrected for the heteroskedasticity created by leverage and standardized to a t distribution" (McClendon, MJ (1994) Multiple regression and causal analysis. Itasca, IL: Peacock Publishers, p. 175.).

RESIDUALS: DELETED. Suppose you left out a particular case when you were doing the regression; you just omitted it from the calculations of Rsquared and b and beta. Then, having left that case out, you look to see what its residual is from the regression line or plane or hyperplane that was calculated without its "input." In cases where the deleted residual differs markedly from the unstandardized, it would suggest to you that that case had a significant influence on where the regression line (or plane or hyperplane) was placed. Stated differently, that case had a big "say" in what happened. You will want to do something about it.

RESIDUALS: STUDENTIZED DELETED. Same as deleted residual immediately above, but now you are studentizing the deleted residual, as was done for the "regular" studentized residuals. Again, you could compare the studentized deleted with the studentized. The advantage of this comparison, instead of the raw vs. deleted raw comparison, is that you have now controlled for residual "spread" with the studentizing.

SPSS also gives you a number of influence statistics. Some of these are under INFLUENCE, some are under DISTANCES. These tell us about the impacts of individual cases on the results you have obtained. These impacts refer to the idea of case **influence**.

"Influence is a function of being at least somewhat unusual with respect to both X and the regression equation. Having the characteristics of an outlier with respect to one frame of reference but not with respect to the other will lead to little or no influence on the regression coefficients" (McLendon, p. 52-53).

Generally, how important a case is depends on how "deviant" it is compared to other cases in the sample. A case can be deviant in one of two ways. It can be deviant in terms of its values on X (or X1 or X2 or X3 ... or any combination of the predictors). If it lies far from the mean of X it will be an **X outlier** (McLendon, p. 50). Typically you square the z-scored value of x to determine how much of an X outlier a case.

Alternatively, a case can be deviant (extreme) "if it is unusual relative to the regression line" (op cit). Its value of Y is **unusual relative to its value of X**. In this case it is a **regression outlier**.

COOKS D provides a measure of **influence** - how much that data point actually effects the position of the regression line (or plane or hyperplane). It tells you how much the regression line (plane) moves if that case is deleted. (See Darlington p. 345 on.) "Cook's D ... is a measure of the total influence on the multiple regression equation of deleting a particular case. D takes into account the changes in slopes plus the change of the intercept" (McLendon, p. 107).

D is a product of how far the case is from the regression line or surface, as measured by a standardized residual (and yes, I do mean standardized residual, not studentized residual), and its leverage (Hamilton, p. 359).

So notice that Cook's D builds on leverage, and takes additional information into account. It is a

more sophisticated measure of influence than is leverage.

Cook's D is PROPORTIONAL to "the sum of squared changes in values of $Y_{\text{predicted}}$ if case i is removed from the sample" (359).

DFBETA is related to the difference between a residual and a deleted residual and is another measure of influence on the slope of the regression. If the slope with the ith case deleted is referred to as:

$b(-i)$

then

$$\text{DFBETA} = [b - b(-i)]$$

In other words, it tells you how much the slope shifts when you drop that case.

LEVERAGE tells you about the potential for a case to influence the regression line (or plane). It is really only relevant when you have two or more predictors. (We are not there yet.) It tells you about the extent to which a case has an atypical pattern of scores on the predictor variables. (See Darlington pp. 352 & 353.)

Leverage is determined entirely by the case's pattern of scores on the X variables; the more atypical its pattern, the higher the leverage. In simple regression, the farther from the mean on X, the higher the leverage (p. 351). See Figures 14.2 and 14.3 in Darlington.

SEPREP reflects the standard error of the predicted Y score, or the relative uncertainty with which that Y case is predicted.

PUTTING RESIDUAL DIAGNOSTICS INTO A COOKBOOK FORM.

So here are the steps you want to move through (taken from Darlington 1990, pp. 351 on). Our purposes in checking regression diagnostics are threefold

- 1) to check for clerical errors;
- 2) to see if we have violated the assumptions of regression and
- 3) to see if our results are unduly influenced by particular cases, and if so to see if our results are "robust" enough to hold up when we remove the case(s).

1. To see if we have violated the assumption of homoscedasticity: look at plots of RESIDUAL*X1, RESIDUAL*X2, and so on, to be sure that the spread of residuals is comparable at different levels of each predictor. X1 is our first predictor; X2 is our second predictor, and so on.

(I know, I know; up to now we have had only 1 predictor. But soon we will have two. And more.)

2. To see if we have violated the assumption of linearity: look at the plot of RESIDUAL*PREDICTED. If you have curvilinearity you will have certain regions of PREDICTED where all the residuals are above the line, and other regions where the residuals are all below the line. This plot is also helpful in detecting other irregularities such as outliers.

The problem of curvilinearity comes about when X has a CURVILINEAR relationship to Y, the outcome. We'll talk more about this later.

If all is normal, this plot should be relatively elliptical, suggesting little or no correlation between the X predicted values (PREDICTED), and what is left over after predicting Y with X1 (i.e., the residuals). At all different values of X (PREDICTED), you should have residuals BOTH above and below the line RESIDUAL=0.

3. To see if we have violated assumptions about bivariate normal or multivariate normal distributions: examine studentized residuals using Bonferroni-adjusted alpha levels. So if your alpha is .05, and $n=50$, your adjusted alpha = $.05/50 = .001$. Therefore the emergence of a studentized t value beyond the t_{critical} associated with .001 is a violation of the assumptions of regression. Remove the case and redo the analysis.

4. To check for errors or highly influential cases:

a. look at RESIDUAL*PREDICTED. Find cases that look odd in the residual file. (Again, this is easier if you are sorted on one of your X variables before you start.) You are looking here for outliers in your plot.

b. Check the cases with the highest scores on LEVERAGE. High leverage does not contradict the assumptions of regression (p. 353), "but these points should usually be examined at least briefly, if only to check for clerical errors." (p. 353). I would go even further and say drop the case if it has leverage above the cutoff, and re-run the regression and see if a substantial change results.

What are the cutoffs? If in a regression the maximum leverage $> .2$ it is "risky" and if it is $> .5$ "avoid if possible." (Hamilton p. 130).

c. Check the cases with the highest INFLUENCE as measured by COOK'S D. How can you tell if you are "too high" on Cook? Hamilton suggests $D > 4/N$. For our 50 cases this gives us a cutoff value of $> .08$.

SUMMARY TABLE ON REGRESSION DIAGNOSTICS

Measure	Tells you	Cutoff	Purpose
Studentized residuals beyond Bonferroni-adjusted alpha levels	Distance of case from regression surface, correcting for leverage	α/n	Detect violations of assumptions about multivariate normality; detect outliers; find clerical errors
Leverage	<p>Leverage:</p> <ul style="list-style-type: none"> -- atypicality of case's pattern of scores on the predictor variables -- "the proportion by which case i lowers its own residual by pulling the regression line or plane toward itself" (355) 	<p>$h > .2$ (risky)</p> <p>$h > .5$ (avoid)</p>	<ul style="list-style-type: none"> - find errors - find highly influential cases
Cook's D	<p>Influence:</p> <ul style="list-style-type: none"> - PROPORTIONAL to "the sum of squared changes in values of $Y_{\text{predicted}}$ if case i is removed from the sample" (359). - product of distance and leverage 	$D > 4/n$	<ul style="list-style-type: none"> - find errors - find highly influential cases

Glossary

Cook's D.

deleted studentized residual

deleted residual

dfBeta

influence

leverage

regression outlier

residual

SEPREP

standardized residual

studentized residual

X outlier