Assumptions of Linear Regression Model

Linear regression model is use to analyze the behavior of a response variable. Statistical analyses based on such models can provide useful results if the correct model has been chosen and other **assumptions underlying the model are satisfied**.

However, if violations of these assumptions are more substantial, the results may not reflect the true population relationships, therefore providing incorrect conclusions that may lead to recommendations or actions that do more harm than good. Unfortunately, such results are rather likely to occur in regression analyses because these are frequently used on data that are ***not*** the result of carefully designed experiments. For this reason it is important to scrutinize the results of regression analyses to determine if problems exist that may compromise the validity of the results, and if such problems are deemed to exist, to search for remedial measures or modify the inferences.

1. The model fits all but one or a few outlier observation.

Recall that, in minimizing the residual sum of squares (SSE), ordinary least squares (OLS) gives equal weight to every observation in the sample. But every observation may not have equal impact on the regression results because of the presence of three types of special data points called **outliers, leverage points,** and **influence points.** It is important that we know what they are and how they influence regression analysis.

In the regression context, an **outlier** may be defined as an observation with a ***"large residual."*** Recall that $e_i = (Y_i - \hat{Y}_i)$, that is, the residual represents the difference (positive or negative) between the actual value of the response variable and its value estimated from the regression model. When we say that a residual is large, it is in comparison with the other residuals and very often such a large residual catches our attention immediately because of its rather large vertical distance from the estimated regression line. Note that in a data set there may be more than one outlier.

An observation that is markedly different from, or atypical of, the rest of the observations of a data set is known as an outlier. An observation may be an outlier with respect to the response variable and/or the independent variables. Specifically, an extreme observation in the response variable is called an **outlier**, while extreme value(s) in the x's (independent variables) are said to have high **leverage** and are often called **leverage points**.

A data point is said to exert (high) **leverage** if it is disproportionately distant from the bulk of the values of a regressor(s). Why does a leverage point matter? It matters because it is capable of pulling the regression line toward itself, thus distorting the slope of the regression line. If this actually happens, then we call such a leverage (data) point an **influential point.** The removal of such a data point from the sample can dramatically affect the regression line.

An observation that causes the regression estimates to be substantially different from what they would be if the observation were removed from the data set is called an **influential observation**. Observations that are outliers or have high leverage are not necessarily influential, whereas influential observations usually are outliers and have high leverage.

Take Note!!!

Automatic rejection of outliers is not always a wise procedure. Sometimes the outlier is providing information that other data points cannot due to the fact that it arises from an unusual combination of circumstances which may be of vital interest and requires further investigation rather than rejection. As a general rule, outliers should be rejected out of hand only if they can be traced to causes such as errors of recording the observations or setting up the apparatus (in a physical experiment. Otherwise, careful investigation is in order.

We begin this lecture by defining a **standardized residual** as the value of the residual divided by the model standard deviation $\hat{\sigma}$. Since we assume that the residuals have a mean of $E(\varepsilon_i) = 0$ and a standard deviation estimated by $\sigma(\varepsilon_i) = \sigma$, a standardized residual is simply the z-score for a residual.

Definition 1:

The **standardized residual**, denoted z_i for the ith observation is the residual for the observation divided by $\hat{\sigma}$, that is,

$$z_i = \frac{e}{\hat{\sigma}} = \frac{y_i - \hat{y}}{\hat{\sigma}}$$

Although we expect almost all the regression residuals to fall within three standard deviations of their mean of 0, sometimes one or several residuals fall outside this interval. Observations with residuals that are extremely large or small (say, more than 3 standard deviations from 0) are called **outliers**. Consequently, observations with standardized residuals that exceed 3 in absolute value are considered outliers.

Take Note!!!

As an alternative to standardized residuals, some software packages compute **studentized residuals**, so named because they follow an approximate Student's t-distribution.

The **studentized residual**, denoted z_i^*, for the ith observation is

$$z_i^* = \frac{e}{\hat{\sigma}\sqrt{1-h_i}} = \frac{y_i - \hat{y}}{\hat{\sigma}\sqrt{1-h_i}}$$

where h_i is the ***leverage***.

Rule of Thumb for Detecting Influence with Leverage

The observed value of Y_i is influential if

$$h_i > \frac{2(k+1)}{n}$$

Where

h_i = the leverage for the ith observation

k = the number of $\beta's$ in the model (excluding β_0)

Jackknife

Another technique for identifying influential observations requires that you delete the observations one at a time, each time refitting the regression model based on only the remaining $n-1$ observations. This method is based on a statistical procedure called the ***jackknife***, which in gaining increasing acceptance among practitioners. The basic principle of the jackknife when applied to regression is to compare the regression results using all n observations to the results with the ith observation deleted to ascertain how much influence a particular observation has on the analysis. Using the jackknife, several alternative influence measures can be calculated.

Cook's Distance

A measure of the overall influence an outlying observation has on the estimated β coefficients. Cook's Distance, D_i, is calculated for the ith observation as follows:

$$D_i = \frac{(Y_i - \hat{Y}_i)^2}{(k+1)\hat{\sigma}}\left[\frac{h_i}{(1-h_i)^2}\right]$$

Note that D_i depends on both the residual $(Y_i - \hat{Y}_i)$ and the leverage h_i for the ith observation. A large value of D_i indicates that the observed Y_i value has strong influence on the estimated β coefficients (since the residual, the leverage, or both will be large).

DFBETA

DFBETA measures the difference in each parameter estimate with and without the influential point. There is a DFBETA for each data point i.e if there are n observations and k variables, there will be $n \times k$ DFBETAs. In general, large values of DFBETAS indicate observations that are influential in estimating a given parameter. Belsley, Kuh, and Welsch recommend 2 as a general cutoff value to indicate influential observations and $\frac{2}{\sqrt{n}}$ as a size-adjusted cutoff.

Grubbs's test

The Grubbs test allows to detect whether the highest or lowest value in a dataset is an outlier. The Grubbs test detects one outlier at a time (highest or lowest value), so the null and alternative hypotheses are as follows:

Ho: The *highest* value is **not** an outlier

Ha: The *highest* value is an outlier

if we want to test the highest value, or

Ho: The *lowest* value is **not** an outlier

Ha: The *lowest* value is an outlier

if we want to test the lowest value. Note that the Grubbs test is not appropriate for sample size of 6 or less $(n \leq 6)$.

Dixon's test

Similar to the Grubbs test, Dixon test is used to test whether a single low or high value is an outlier. So if more than one outliers is suspected, the test has to be performed on these suspected outliers individually. Note that Dixon test is most useful for small sample size (usually $n \leq 25$).

Rosner's test

Rosner's test for outliers has the advantages that:

1. it is used to **detect several outliers at once** (unlike Grubbs and Dixon test which must be performed iteratively to screen for multiple outliers), and
2. it is designed to avoid the problem of masking, where an outlier that is close in value to another outlier can go undetected.

Unllike Dixon test, note that Rosner test is most appropriate when the sample size is large ($n \geq 20$).

Example:

The data in the Excel file presents the sales, *Y*, in thousands of peso per week, for fast-food outlets in each of four cities. The objective is to model sales, *Y* as a function of traffic flow, adjusting for city-to-city variations that might be due to size or other market conditions. Furthermore, we believe that the level of mean sales will differ from city to city, but that the change in mean sales per unit increase in traffic flow will remain the same for all cities (i.e., that the factors Traffic Flow and City do not interact). The model is therefore

$$E(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i}$$

Where

Y_i = (average) sales, *Y*, in thousands of peso per week.

X_{1i} = 1 if the fast-food outlets is in the City 1 and 0 otherwise.

X_{2i} = 1 if the fast-food outlets is in the City 2 and 0 otherwise.

X_{3i} = 1 if the fast-food outlets is in the City 3 and 0 otherwise.

X_{4i} = traffic flow thousands of cars.

A. Fit the model using the Excel file's data and examine the value of the coefficient of determination and the anova table.
B. Check for outliers by plotting the residuals from the model using the Excel file's information.
C. If there are any outliers based on the graph, take corrective action.

Exercise:

A large manufacturing firm wants to determine whether a relationship exists between *Y*, the number of work-hours an employee misses per year, and x, the employee's annual wages. A sample of 15 employees produced the data in the table below.

Employee	Work-Hours Missed	Annual Wages
1	49.0	12.8
2	36.0	14.5
3	127.0	8.3
4	91.0	10.2
5	72.0	10.0
6	34.0	11.5
7	155.0	8.8
8	11.0	17.2
9	191.0	7.8
10	6.0	15.8
11	63.0	10.8
12	79.0	9.7
13	543.0	12.1
14	57.0	21.2
15	82.0	10.9

Determine if there are any outliers using cook's distance plot.

If there are any outliers, remove the outliers and fit the model again. Compare the model with outliers to the model without outliers.

Prepared by:

KATRINA D. ELIZON

Department of Mathematics and Statistics

College of Science