Assumptions of Linear Regression Model

2. There is exact linear relationship or exact collinearity between the X variables.

Often, two or more of the independent variables used in the model for $E(Y)$ will contribute redundant information. That is, the independent variables will be correlated with each other. For example, suppose we want to construct a model to predict the gasoline mileage rating, Y, of a truck as a function of its load, X_1, and the horsepower, X_2, of its engine. In general, you would expect heavier loads to require greater horsepower and to result in lower mileage ratings. Thus, although both X_1 and X_2 contribute information for the prediction of mileage rating, some of the information is overlapping, because X_1 and X_2 are correlated. **Multicollinearity** exists when two or more of the independent variables used in regression are moderately or highly correlated.

Some of the problems arise when serious multicollinearity is present in the regression analysis:

1. High correlations among the independent variables (i.e., **extreme** multicollinearity) increase the likelihood of rounding errors in the calculations of the β estimates, standard errors, and so forth.
2. The regression results may be confusing and misleading specifically, the results of t-test and F - test.
3. Have an effect on the signs of the parameter estimates.

Detecting Multicollinearity in the Regression Model

1. Significant correlations between pairs of independent variables in the model.
2. Nonsignificant t-tests for all (or nearly all) the individual β parameters when the F-test for overall model adequacy $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ is significant.
3. Opposite signs (from what is expected) in the estimated parameters.
4. A variance inflation factor (VIF) for a β parameter based on the rule of thumb:
   - If $1 \leq VIF \leq 5$, it generally considered acceptable.
   - If $VIF > 5$, it indicates moderate multicollinearity, which may warrant closer inspection.
   - If $VIF > 10$, it indicates high multicollinearity.

$$(VIF)_i = \frac{1}{1 - R_i^2}$$

and R_i^2 is the multiple coefficient of determination for the model. For $R_i^2 > 0.90$ indicates extreme multicollinearity case.

Take note that VIF can't be used if there are variables with more than one degree of freedom (e.g. **polynomial and other contrasts relating to categorical variables with more than two levels**) and recommend using the GVIF function (generalized variance inflation factor).

The GVIF is calculated for sets of related regressors, such as a for a set of indicator regressors for some kind of categorical variable, or for polynomial variables:

- For the continuous variables GVIF is the same as the VIF values before.
- For the categorical variables, we now get one GVIF value for each separate category type.

So, variables which *require more than 1* coefficient and thus *more than 1 degree of freedom* are typically evaluated using the GVIF.

We can apply the usual VIF rule of thumb if we squared the $GVIF^{(1/(2*Df))}$ value.

Example:

Which of the following pairs of variables are perfectly multicollinear?

1. X_i and $5X_i$
2. X_i and X_i^3
3. $3X_i$ and $3^2X_i^3$

3. The error terms are not normally distributed.

All the inferential procedures associated with a regression analysis are based on the assumptions that, for any setting of the independent variables, the random error ε_i is normally distributed with mean 0 and variance σ^2, and all pairs of errors are independent. Of these assumptions, the normality assumption is the least restrictive when we apply regression analysis in practice. That is, moderate departures from the assumption of normality have very little effect on Type I error rates associated with the statistical tests and on the confidence coefficients associated with the confidence intervals.

Why do we employ the normality assumption? There are several reasons:

1. ε_i represent the combined influence (on the dependent variable) of a large number of independent variables that are not explicitly introduced in the regression model. As noted, we hope that the influence of these omitted or neglected variables is small and at best random. Now by the celebrated **central limit theorem (CLT)** of statistics, it can be shown that if there are a large number of independent and identically distributed random variables, then, with a few exceptions, the distribution of their sum tends to a normal distribution as the

number of such variables increase indefinitely. It is the CLT that provides a theoretical justification for the assumption of normality of ε_i .

2. With the normality assumption, the probability distributions of OLS estimators can be easily derived because, one property of the normal distribution is that **any linear function of normally distributed variables is itself normally distributed.** Based on the previous lessons, OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear functions of ε_i. Therefore, if ε_i are normally distributed, so are $\hat{\beta}_0$ and $\hat{\beta}_1$, which makes our task of hypothesis testing very straightforward.
3. The normal distribution is a comparatively simple distribution involving only two parameters (**mean and variance**); it is very well known and its theoretical properties have been extensively studied in mathematical statistics.

To determine whether the error terms have a normal distribution, we can conduct the following test:

1. Shapiro Wilk Test
2. Kolmogorov Smirnov Test
3. Anderson Darling Test
4. Lilliefors test
5. Cramer-von Mises Test
6. Jarque Bera Test

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.

Recall that on the previous lesson, there are two influential observations that are outliers, so we remove rows 3 and 5 from the dataset. Using the new dataset,

1. Using VIF, determine if multicollinearity appears in the model. If there is a violation of the assumption, take corrective action.
2. Use the above-mentioned six tests to determine whether the model's residual is normally distributed.
3. Take corrective action if the Shapiro-Wilk test reveals that the residual of the model is not normal.

Prepared by:

KATRINA D. ELIZON
Department of Mathematics and Statistics
College of Science