MULTIPLE REGRESSIONS

The general purpose of multiple regression is to learn more about the relationship between several independent or explanatory variables and a dependent variable. More precisely, multiple regression analysis helps us to predict the value of *Y* for given values of $X_1, X_2, \ldots, X_k$.

In general, the multiple regression equation of *Y* on $X_1, X_2, \ldots, X_k$ is given by:

$$\boldsymbol{Y(t) = a + b_1 X_1(t) + b_2 X_2(t) + \ldots + b_k X_k(t) + \varepsilon(t)}$$

or - in matrix notations -

$$Y=X\cdot B+\varepsilon,$$

where $Y=(y(1),y(2),\ldots,y(t))$, $B=(b(1),b(2),\ldots,b(t))$, $\varepsilon=(\varepsilon(1),\varepsilon(2),\ldots,\varepsilon(t))$,

$$X=\begin{bmatrix} 1 & x_1(1) & \ldots & x_m(1) \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_1(t) & \ldots & x_m(t) \end{bmatrix}.$$

Here *a* is the intercept and $b_1, b_2, b_3, \ldots, b_k$ are analogous to the slope in a simple linear equation and are also called regression coefficients. They can be interpreted the same way as slope. Thus if $b_i = 2.5$, it would indicate that *Y* will increase by 2.5 units if X_i increased by 1 unit while controlling for other independent explanatory variables.

The ordinary least squares estimators of the parameters are given by

$$B = (X^T X)^{-1} X^T Y.$$

The appropriateness of the multiple regression models as a whole can be tested by the *F*-test. A significant *F* indicates a linear relationship between *Y* and at least one of the *X's*.

ASSUMPTION OF NO PERFECT MULTICOLLINEARITY

Along with all classical linear regression assumptions there is another important assumption. It is that of ***non-existence of multicollinearity*** - the independent variables are not related among themselves. At a very basic level, this can be tested by computing the correlation coefficient between each pair of independent variables.
This is a common problem in many correlation analyses. Imagine that you have two predictors (X variables) of a person's height: (1) weight in pounds and (2) weight in ounces. Obviously, our two predictors are completely redundant; weight

is one and the same variable, regardless of whether it is measured in pounds or ounces. Trying to decide which one of the two measures is a better predictor of height would be rather silly; however, this is exactly what you would try to do if you were to perform a multiple regression analysis with height as the dependent (Y) variable and the two measures of weight as the independent (X) variables.

When there are very many variables involved, it is often not immediately apparent that this problem exists, and it may only manifest itself after several variables have already been entered into the regression equation. Nevertheless, when this problem occurs it means that at least one of the predictor variables is (practically) completely redundant with other predictors.

Multicollinearity almost always exists in observational data, and often exists in experimental data as well. The question is therefore not "is there multicollinearity?", but rather "how strong is the multicollinearity?"

Generally, the higher the correlations among the *X*'s, the greater the degree of multicollinearity. The extreme case is when the columns are perfectly linearly dependent, in which case there are no unique least squares estimates.

In practice, scholars almost never face perfect multicollinearity. However, they often encounter near-perfect multicollinearity.

EFFECTS OF (NEAR-PERFECT) MULTICOLLINEARITY

1. Increased standard errors of estimates of the b_i's and, therefore,

    - decreased reliability (lack of precision);
    - "insignificant" t-ratios;
    - wider confidence intervals.

2. The OLS estimators and their standard errors can be sensitive to small changes in the data. In other words, the results will not be robust.

3. Often confusing and misleading results: wrong signs and unreasonable values of the parameters estimates.

DEALING WITH MULTICOLLINEARITY

There are several ways for dealing with multicollinearity when it is a problem. The first, and most obvious, solution is to eliminate some variables from the model. If two variables are highly collinear, then it means they contain highly redundant information. Thus, we can pick one variable to keep in the model and discard the other one.