6.2: Multiple Regression

Different Methods of Multiple Regression

Before we go into the statistical details of multiple regression, I want to first introduce three common methods of multiple regression: forced entry regression, hierarchical regression, and stepwise regression. The differences between these methods of multiple regression lies in how the variables are entered into a regression model. Let's look at each of them.

In forced entry regression, we choose independent variables, or predictors, based on theories and/or empirical literature to include in the regression model. Like the name suggests, we will force enter all chosen independent variables into the regression model simultaneously and study them altogether. In this case, all the predictors are treated equally. In other words, we don't have a preference or hold more interest in one predictor over the other predictors.

In hierarchical regression, just like in forced entry regression, we rely on previous theoretical and/or empirical evidence to select the independent variables to be included. But unlike forced entry regression, the predictors are not entered simultaneously. Instead, we, the researchers, determine in which order the predictors are entered. This is where the hierarchy comes in. It is essentially the order how the predictors are entered. Each step is considered a block, and one or multiple predictors can be included/entered in each block. Each block is considered a model. So hierarchical regression typically include multiple models. For example, if you enter two predictors (IV1, IV2) in block 1, and then enter another predictor (IV3) in block 2, then you will have two models, model 1 from block 1, and model 2 from block 2. In model 1, there are two predictors: IV1 and IV2. In model 2, there are three predictors (IV1 and IV2 from model 1, plus IV3 we just added). Keep in mind, each predictor can be entered only once. Once a predictor is entered into a block, it stays there for all the following blocks and you can't take the predictor(s) out once they are in. That's why IV1 and IV2 remain in model 2 above.

So how do we decide which order to enter the predictors? Generally speaking, in the first model, we would include demographic variables, such as gender, ethnicity, education levels, etc. These predictors likely will influence the dependent variables even though they may not be the focus of our research study. In the next model (model 2), we would include any variables that are known predictors for the dependent variable(s). In the next model (model 3), we will add in new predictors we are particularly interested in. Often times, our goal is to determine if newly added variables could better explain the dependent variable(s), or whether newly added variables could explain significantly more variance in the dependent variable above and beyond the other variables included in the models.

Lastly, unlike the first two methods of regression, stepwise regression doesn't rely on theories or empirical literature at all. It is a purely mathematically based model. All you need to do is throwing in a bunch of IVs, and the software program will sift through all the IVs you entered to identify the ones that best predict the dependent variable(s) by selecting the predictor(s) that has the highest correlation with the dependent variable. It can be done using either forward method or backward method. Regardless, the decision is purely based on mathematical criterion, not on theories. If you know one thing about stepwise regression, that is to avoid it at all cost. As researchers, we want to make sure we choose the predictors based on theories and/or empirical literature.

Multiple Regression

Regression analysis is a statistical technique that can test the hypothesis that a variable is dependent upon one or more other variables. Further, regression analysis can provide an estimate of the magnitude of the impact of a change in one variable on another. This last feature, of course, is all important in predicting future values.

Regression analysis is based upon a functional relationship among variables and further, assumes that the relationship is linear. This linearity assumption is required because, for the most part, the theoretical statistical properties of non-linear estimation are not well worked out yet by the mathematicians and statisticians. There are techniques for overcoming some of these difficulties, exponential and logarithmic transformation of the data for example, but at the outset we must recognize that standard ordinary least squares (OLS) regression analysis will always use a linear function to estimate what might be a nonlinear relationship.

The general linear regression model can be stated by the equation:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + e_i$$

where b_0 is the intercept, b_i's are the slope between Y and the appropriate X_i, and e, is the error term that captures errors in measurement of y and the effect on y of any variables missing from the equation that would contribute to explaining variations in y.

This model works only if certain assumptions hold. We'll look at this next.

Assumptions of the Ordinary Least Squares Regression Model

There are several assumptions of OLS regression. If one of these assumptions fails to be true, then it will have an effect on the quality of the estimates. Some of the failures of these assumptions can be fixed while others result in estimates that quite simply provide no insight into the questions the model is trying to answer or worse, give biased estimates.

1. The error term is a normally distributed with a mean of zero and a constant variance. The meaning of this is that the variances of the independent variables are independent of the value of the variable. Consider the relationship between personal income and the quantity of a good purchased, which is an example of a case where the variance is dependent upon the value of the independent variable, income. It is plausible that as income increases, the variation around the amount purchased will also increase simply because of the flexibility provided with higher levels of income. The assumption is for constant variance with respect to the magnitude of the independent variable called homoscedasticity. If the assumption fails, then it is called heteroscedasticity. Figure 13.6 shows the case of homoscedasticity where all three distributions have the same variance around the predicted value of Y regardless of the magnitude of X.
2. The independent variables are all from a probability distribution that is normally distributed. This can be seen in Figure 13.6 by the shape of the distributions placed on the predicted line at the expected value of the relevant value of Y.
3. The independent variables are independent of Y, but are also assumed to be independent of the other X variables, or other independent variables. The model is designed to estimate the effects of independent variables on some dependent variable in accordance with a proposed theory. The case where some or more of the independent variables are correlated is not unusual. There may be no cause and effect relationship among the independent variables, but nevertheless they move together. For example, you have two variables, household income and socio-economic status (SES), and they are theoretically related to each other. If you want to use both of them as predictors in one model, it would violate this assumption of regression analysis. This condition is called multicollinearity, which will be taken up in detail later.

Figure 13.6 does not show all the assumptions of the regression model, but it helps visualize these important ones.

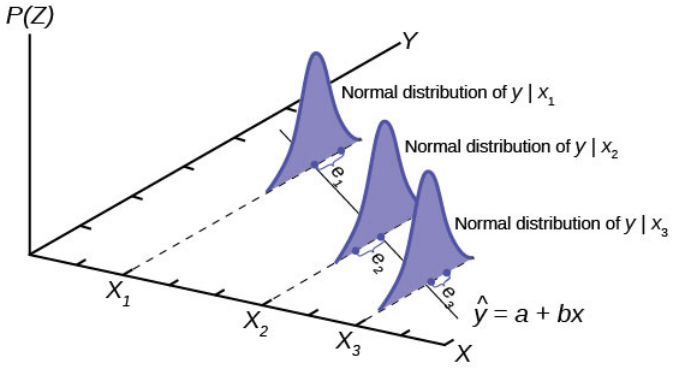
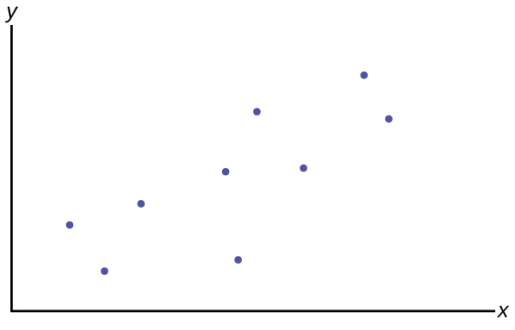


Figure 13.6



$$y = \beta_0 + \beta_1 X + \varepsilon$$

Figure 13.7

Going back to the general linear regression model stated earlier:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + e_i$$

This is the general form that is most often called the multiple regression model. So-called "simple" regression analysis has only one independent variable rather than many independent variables. Simple regression is just a special case of multiple regression. There is some value in beginning with simple regression: it is easy to graph in two dimensions, difficult to graph in three dimensions, and impossible to graph in more than three dimensions. Consequently, our graphs will be for the simple regression case. Figure 13.7 presents the regression problem in the form of a scatter plot graph of the data set where it is hypothesized that Y is dependent upon the single independent variable X.

Let's look at an example. The theoretical relationship states that as a person's income rises, their consumption rises, but by a smaller amount than the rise in income. If Y is consumption and X is income in the equation below Figure 13.7, the regression problem is, first, to establish that this relationship exists, and second, to determine the impact of a change in income on a person's consumption. Each "dot" in Figure 13.7 represents the consumption and income of different individuals at some point in time.

Regression analysis is often called "ordinary least squares" (OLS) analysis because the method of determining which line best "fits" the data is to minimize the sum of the squared residuals or erros of a line put through the data.

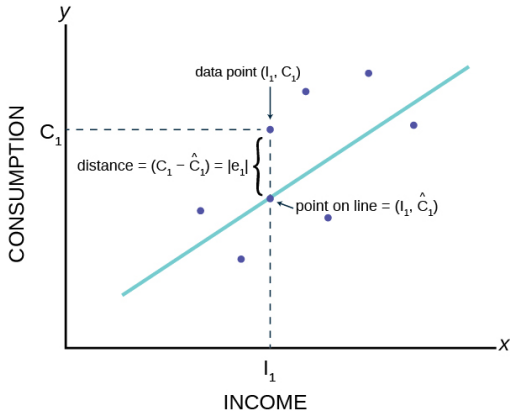


Figure 13.8

Estimated Equation: $C = b_0 + b_1 \text{Income} + e$

Figure 13.8 shows the assumed relationship between consumption and income based on the theory. Here the data are plotted as a scatter plot and an estimated straight line has been drawn. From this graph we can see an error term, e_1. Each data point also has an error term. Again, the error term is put into the equation to capture effects on consumption that are not caused by income changes. Such other effects might be a person's savings or wealth, or periods of unemployment. We will see how by minimizing the sum of these errors we can get an estimate for the slope and intercept of this line.

Consider the graph below. The notation has returned to that for the more general model rather than the specific example of the consumption and income.

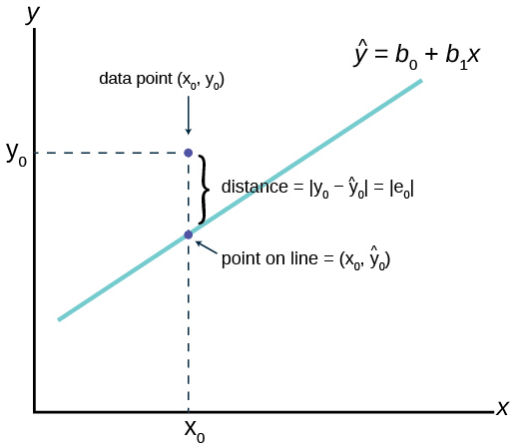


Figure 13.9

The $\hat{y}$ is read **"y hat"** and is the **estimated value of y**. (In Figure 13.8 $\hat{C}$ represents the estimated value of consumption because it is on the estimated line.) It is the value of y obtained using the regression line. $\hat{y}$ is not generally equal to y from the data.

The term $y_0 - \hat{y}_0 = e_0$ is called the **"error" or residual**. The error term was put into the estimating equation to capture missing variables and errors in measurement that may have occurred in the dependent variables. The **absolute value of a residual** measures the vertical distance between the actual value of y and the estimated value of y. In other words, it measures the vertical distance between the actual data point and the predicted point on the line as can be seen on the graph at point X_0.

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for y.

If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for y.

In the graph, $y_0 - \hat{y}_0 = e_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive. For each data point the residuals, or errors, are calculated $y_i - \hat{y}_i = e_i$ for $i = 1, 2, 3, \ldots, n$ where n is the sample size. Each $|e|$ is a vertical distance.

The sum of the errors squared is the term obviously called **Sum of Squared Errors (SS Error)**.

Using calculation, you can determine the straight line that has the parameter values of b_0 and b_1 that minimizes the **SS Error.** When you make the **SS Error** a minimum, you have determined the points that are on the line of best fit. We can further calculate the variance of the squared errors, e^2:

$$s_e^2 = \frac{\Sigma (y_i - \hat{y}_i)^2}{n-k} = \frac{\Sigma e_i^2}{n-k}$$

where $\hat{y}$ is the predicted value of y and y is the observed value, and thus the term $(y_i - \hat{y}_i)^2$ is the squared errors that are to be minimized to find the regression line. One important note is that here we are dividing by $(n-k)$, which is the degrees of freedom. The degrees of freedom of a regression equation will be the number of observations, n, reduced by the number of estimated parameters, which includes the intercept as a parameter.

The variance of the errors is fundamental in testing hypotheses for a regression. It tells us just how “tight” the dispersion is about the line. The greater the dispersion about the line, meaning the larger the variance of the errors, the less probable that the hypothesized independent variable will be found to have a significant effect on the dependent variable. In short, the theory being tested will more likely fail if the variance of the error term is high. Upon reflection this should not be a surprise. As we tested hypotheses about a mean we observed that large variances reduced the calculated test statistics and thus it failed to reach the tail of the distribution. In those cases, the null hypotheses could not be rejected. If we cannot reject the null hypothesis in a regression problem, we must conclude that the hypothesized independent variable has no effect on the dependent variable.

A way to visualize this concept is to draw two scatter plots of x and y data along a predetermined line. The first will have little variance of the errors, meaning that all the data points will move close to the line. Now do the same except the data points will have a large estimate of the error variance, meaning that the data points are scattered widely along the line. Clearly the confidence about a relationship between x and y is affected by this difference between the error variances.