Multiple Regression I (Soc 504)

Generally, we are not interested in examining the relationship between simply two variables. Rather we may be interested in examining the relationship between multiple variables and some outcome of interest. Or, we may believe that a relationship between one variable and another is spurious on a third variable. Or, we may believe that the relationship between one variable and another is being 'masked' by some third variable. Or, still yet, we may believe that a relationship between one variable and another may depend on another variable. In these cases, we conduct multiple regression analysis, which is simply an extension of the simple regression model we have discussed thus far.

1 The Multiple Regression Model

In scalar notation, the multiple regression model is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i$$

We rarely express the model in this fashion, however, because it is more compact to use matrix notation. In that case, we often use:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \ldots & X_{1k} \\ 1 & X_{21} & \ldots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \ldots & X_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

or just $Y = X\beta + \epsilon$. At the sample level, the model is $Y = Xb + e$. In these equations, n is the number of observations in the data, and $k+1$ is the number of regression parameters (the $+1$ is for the intercept, β_0 or b_0).

Note that if you use what you know about matrix algebra, you can multiply out the X matrix by the β matrix, add the error vector, and get the scalar result above for each observation. The column of ones gets multiplied by β_0, so that the intercept term stands alone without an X variable.

2 The OLS Solution in Matrix Form

The OLS solution for β can be derived the same way we derived it in the previous lecture, but here we must use matrix calculus. Again, we need to minimize the sum of the squared error terms ($\sum_{i=1}^{n} e_i^2$). This can be expressed in matrix notation as:

$$F = min\,(Y - Xb)^T\,(Y - Xb)$$

Notice that the summation symbol is not needed here, because $(Y - Xb)$ is an $n \times 1$ column vector. Transposing this vector and multiplying it by itself (untransposed) produces a scalar that is equal to the sum of the squared errors.

In the next step, we need to minimize F by taking the derivative of the expression and setting it equal to 0, as before:

$$\frac{\partial F}{\partial \beta} = -2X^T (Y - X\beta) = -2X^T Y + 2X^T X\beta.$$

It may be difficult to see how this derivative is taken. Realize that the construction above is a quadratic form in β (and X). We could think of the equation as: $(Y - X\beta)^2$. In that case, we would obtain: $-2X(Y - X\beta)$ for our derivative. This is exactly what this expression is. X is transposed so that the multiplication that is implied in the result is possible. Note that, using the distributive property of matrix multiplication, we are able to distribute the $-2X^T$ across the parenthetical.

Setting the derivative equal to 0 and dividing by -2 yields:

$$0 = X^T Y - X^T X\beta$$

Obviously, if we move $-X^T X\beta$ to the other side of the equation, we get:

$$X^T X\beta = X^T Y$$

In order to isolate β, we need to premultiply both sides of the equation by $(X^T X)^{-1}$. This leaves us with $I\beta$ on the left, which equals β, and the OLS solution—$\left(X^T X\right)^{-1} \left(X^T Y\right)$—on the right.

The standard error of the regression looks much as before. Its analog in matrix form is:

$$\sigma_e^2 = \frac{1}{n-k} e^T e.$$

Finally, the variance-covariance matrix of the parameter estimates can be obtained by:

$$Var(\beta) = \sigma_e^2 \left(X^T X\right)^{-1}$$

You would need to square root the diagonal elements to obtain the standard errors of the parameters for hypothesis testing. Notice that this result looks similar to the bivariate regression result, if you think of the inverse function as being similar to division. We will derive these estimators for the standard errors using an ML approach in the next lecture.

3 Matrix Solution for Simple Regression

We will demonstrate the OLS solution for the bivariate regression model. Before we do so, though, let me discuss a little further a type of matrix expression you will see often: $X^T X$. As we discussed above, this type of term is a quadratic form, more or less equivalent to X^2 in scalar notation. The primary difference between the matrix and the scalar form is that in the matrix form the off-diagonal elements of the resulting matrix will be the 'cross-products' of the X variables-essentially their covariances-while the main diagonal will be the sum of X^2 for each X-essentially the variances.

The X matrix for the bivariate regression model looks like:

$$\begin{bmatrix} 1 & X_{11} \\ 1 & X_{21} \\ \vdots & \vdots \\ 1 & X_{n1} \end{bmatrix}$$

For the purposes of exposition, I will not change the subscripts when we transpose this matrix. If we compute $X^T X$, we will get:

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ X_{11} & X_{21} & \dots & X_{n1} \end{bmatrix} \begin{bmatrix} 1 & X_{11} \\ 1 & X_{21} \\ \vdots & \vdots \\ 1 & X_{n1} \end{bmatrix} = \begin{bmatrix} n & \sum X \\ \sum X & \sum X^2 \end{bmatrix}$$

To compute the inverse of this matrix, we can use the rule presented in the last chapter for inverting 2×2 matrices:

$$M^{-1} = \frac{1}{\mid M \mid} \begin{bmatrix} m_{22} & -m_{12} \\ -m_{21} & m_{11} \end{bmatrix}.$$

In this case, the determinant of $\left(X^T X\right)^{-1}$ is $n \sum X^2 - (\sum X)^2$, and so the inverse is:

$$\begin{bmatrix} \frac{\sum X^2}{n\sum X^2 - (\sum X)^2} & \frac{-\sum X}{n\sum X^2 - (\sum X)^2} \\ \\ \frac{-\sum X}{n\sum X^2 - (\sum X)^2} & \frac{n}{n\sum X^2 - (\sum X)^2} \end{bmatrix}$$

We now need to postmultiply this by $X^T Y$, which is:

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ X_{11} & X_{21} & \dots & X_{n1} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum Y \\ \sum XY \end{bmatrix}$$

So, $\left(X^T X\right)^{-1} \left(X^T Y\right)$ is:

$$\begin{bmatrix} \frac{\sum X^2}{n\sum X^2 - (\sum X)^2} & \frac{-\sum X}{n\sum X^2 - (\sum X)^2} \\ \\ \frac{-\sum X}{n\sum X^2 - (\sum X)^2} & \frac{n}{n\sum X^2 - (\sum X)^2} \end{bmatrix} \begin{bmatrix} \sum Y \\ \sum XY \end{bmatrix} = \begin{bmatrix} \frac{\sum X^2 \sum Y - \sum X \sum XY}{n\sum X^2 - (\sum X)^2} \\ \\ \frac{-\sum X \sum Y + n \sum XY}{n\sum X^2 - (\sum X)^2} \end{bmatrix}$$

Let's work first with the denominator of the elements in the vector on the right. The term $\left(\sum X\right)^2$ can be rewritten as $n^2 \bar{X}^2$. Then, an n can be factored from the denominator, and we are left with $n\left(\sum X^2 - n\bar{x}^2\right)$. As we discussed before, this is equal to n times the numerator for the variance: $n \sum (X - \bar{X})^2$.

The numerator of the second element in the vector can be rewritten to be $n\sum XY - n^2\bar{X}\bar{Y}$. An n can be factored from this expression and will cancel with the n in the denominator. So, we are left with $\sum XY - n\bar{X}\bar{Y}$ in the numerator. For reasons that will become apparent in a moment, we can express this term as $-2n\bar{X}\bar{Y} + n\bar{X}\bar{Y}$. The first term here can be rewritten as $-\bar{Y}\sum X - \bar{X}\sum Y$, and the second term can be written as $\sum \bar{X}\bar{Y}$. All four terms can now be collected under a single summation as: $\sum\left(XY - \bar{X}Y - \bar{Y}X + \bar{X}\bar{Y}\right)$, which factors into $\sum(X-\bar{X})(Y-\bar{Y})$, and so the whole expression (numerator and denominator) becomes:

$$\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}.$$

This should look familiar. It is the same expression we obtained previously for the slope. Observe that we now have a new 'computational' formula:

$$\sum XY - n\bar{X}\bar{Y} = \sum(X-\bar{X})(Y-\bar{Y}).$$

What is interesting about the result we just obtained is that we obtained the result without deviating each variable from its mean in the $\left(X^TX\right)$ matrix but the means re-entered anyway.

Now, let's determine the first element in the solution vector. The numerator is $\sum X^2\sum Y - \sum X\sum XY$. This can be reexpressed as $n\bar{Y}\sum X^2 - n\bar{X}\sum XY$. Once again, the denominator is $n\sum\left(X-\bar{X}\right)^2$. So, the ns in the numerator and denominator all cancel. Now, we can strategically add and subtract $\bar{X}n\bar{X}\bar{Y}$ in the numerator to obtain: $\bar{Y}\sum X^2 - \bar{X}n\bar{X}\bar{Y} - \left(\bar{X}\sum XY - \bar{X}n\bar{X}\bar{Y}\right)$. With a minimal amount of algebraic manipulation, we can obtain:

$$\bar{Y}\left(\sum X^2 - n\bar{X}^2\right) - \bar{X}\left(\sum XY - n\bar{X}\bar{Y}\right).$$

If we now separate out the two halves of the numerator and make two fractions, we get:

$$\frac{\bar{Y}\left(\sum X^2 - n\bar{X}^2\right)}{\sum\left(X-\bar{X}\right)^2} - \frac{\bar{X}\left(\sum XY - n\bar{X}\bar{Y}\right)}{\sum\left(X-\bar{X}\right)^2},$$

which is just $\bar{Y} - b_1\bar{X}$.

4 Why Do We Need Multiple Regression?

If we have more than 1 independent variable, the matrices become larger, and, as stated above, the off-diagonal elements of the $\left(X^TX\right)$ matrix contain information equivalent to the covariances among the X variables. This information is important, because the only reason we need to perform multiple regression is to 'control' out the effects of other X variables when trying to determine the true effect of one X on Y. For example, suppose we were interested in examining racial differences in health. We might conduct a simple regression of health on race, and we would find a rather large and significant difference between whites and nonwhites. But, suppose we thought that part of the racial difference in health was attributable to income differences between racial groups. Multiple regression analysis would allow us to control out the income differences between racial groups to determine the residual race differences. If, on the other hand, there weren't racial differences in income (i.e., race and income were not correlated), then including income in the model would not have an effect on estimated race differences in health.

In other words, if the Xs aren't correlated, then there is no need to perform multiple regression. Let me demonstrate this. For the sake of keeping the equations manageable (so they fit on a page) and so that I can demonstrate another point, let's assume that all the X variables have a mean of 0 (or, alternatively, that we have deviated them from their means so that the new set of X variables each have a mean of 0). Above, we derived a matrix expression for $\left(X^T X\right)$ in simple regression, but we now need the more general form with multiple X variables (once again, I have left the subscripts untransposed for clarity):

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ X_{11} & X_{21} & \dots & X_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1k} & X_{2k} & \dots & X_{nk} \end{bmatrix} \begin{bmatrix} 1 & X_{11} & \dots & X_{1k} \\ 1 & X_{21} & \dots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \dots & X_{nk} \end{bmatrix} = \begin{bmatrix} n & \sum X_1 & \sum X_2 & \dots & \sum X_k \\ \sum X_1 & \sum X_1^2 & 0 & \dots & 0 \\ \sum X_2 & 0 & \sum X_2^2 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \sum X_k & 0 & \dots & 0 & \sum X_k^2 \end{bmatrix}.$$

Thus, the first row and column of the matrix contain the sums of the variables, the main diagonal contains the sums of squares of each variable, and all the cross-product positions are 0. This matrix simplifies considerably, if we realize that, if the means of all of the X variables are 0, then $\sum X$ must be 0 for each variable:

$$\left(X^T X\right) = \begin{bmatrix} n & 0 & \dots & 0 \\ 0 & \sum X_1^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sum X_k^2 \end{bmatrix}$$

The matrix is now a diagonal matrix, and so its inverse is obtained simply by inverting each of the diagonal elements. The $\left(X^T Y\right)$ matrix is:

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ X_{11} & X_{21} & \dots & X_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1k} & X_{2k} & \dots & X_{nk} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum Y \\ \sum X_1 Y \\ \vdots \\ \sum X_k Y \end{bmatrix}$$

So, the solution vector is:

$$\left(X^T X\right) = \begin{bmatrix} \frac{1}{n} & 0 & \dots & 0 \\ 0 & \frac{1}{\sum X_1^2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{1}{\sum X_k^2} \end{bmatrix} \begin{bmatrix} \sum Y \\ \sum X_1 Y \\ \vdots \\ \sum X_k Y \end{bmatrix} = \begin{bmatrix} \frac{\sum Y}{n} \\ \frac{\sum X_1 Y}{\sum X_1^2} \\ \vdots \\ \frac{\sum X_k Y}{\sum X_k^2} \end{bmatrix}$$

The last thing we need to consider is what $\sum XY$ and $\sum X^2$ are when the mean of X is 0. Let's take the denominators–$\sum X^2$–first. This is the same as $\sum (X-0)^2$, which, since the means of all the X variables are 0, means the denominator for each of the coefficients is $\sum \left(X-\bar{X}\right)^2$. Now let's think about the numerator. As it turns out, the numerator for each coefficient can be rewritten as $\sum \left(X-\bar{X}\right)\left(Y-\bar{Y}\right)$. Why? Try substituting 0 for $\bar{X}$ and expanding:

$$\sum (X)\left(Y-\bar{Y}\right) = \sum XY - \bar{Y}\sum X = \sum XY - \bar{Y}n\bar{X} = \sum XY - \bar{Y}0 = \sum XY$$

So, each of our coefficients can be viewed as: $\frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sum (X-\bar{X})^2}$. Notice that this is exactly equal to the slope coefficient in the simple regression model. Thus, we have shown that, if the means of the X variables are equal to 0, and the X variables are uncorrelated, then the multiple regression coefficients are identical to what we would obtain if we conducted separate simple regressions for each variable. The results we obtained here also apply if the mean of the X variables are not 0, but the matrix expressions become much more complicated.

What about the intercept term? Notice in the solution vector that the intercept term simply turned out to be the mean of Y. This highlights an important point: the intercept is simply an adjusted mean of the outcome variable. More specifically, it represents the mean of the dependent variable when all the X variables are set to 0 (in this case, their means). This interpretation holds when the X variables' means are not 0, as well, just as we discussed previously in interpreting the coefficients in the simple regression model.