Multiple Linear Regression

ECON 340: Economic Research Methods — Instructor: Div Bhagia

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a dependent variable by fitting a linear equation on the observed data. Let's think about a case with two explanatory variables X_1 and X_2:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{u}_i$$

To estimate the values of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ using ordinary least squares, we minimize the sum of squared residuals, just as we do in simple regression.

1 Adjusted R^2

We can also define three sums of squares — total, explained, and residual — and calculate the R-squared value, which represents the percentage of variation in the dependent variable that can be explained by using X_1 and X_2 to predict it for each observation in the sample.

Total Sum of Squares:

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Explained Sum of Squares:

$$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

Residual Sum of Squares:

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

R-squared as before is defined as:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

It is important to note that the R^2 of a regression model never decreases when an additional explanatory variable is added, no matter how irrelevant that variable may be. This means that the R^2 of an expanded regression model will always be equal to or greater than that of the original regression model. Therefore, we should not rely solely on a high R^2 value when evaluating a model with a long list of explanatory variables, as the value may be arbitrarily increased by adding even unimportant variables.

To address this issue and compare alternative models with different numbers of explanatory variables, we can use an alternative measure called Adjusted R^2, which takes into account the number of variables used in the model. This measure is calculated as follows:

$$AdjustedR^2 = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}$$

where k is the number of variables. The numerator and denominator of R^2 are divided by their respective degrees of freedom to calculate Adjusted R^2. The denominator, which represents the total sum of squares, remains constant for a given dependent variable. However, the numerator, which represents the residual sum of squares, decreases as k is increased. If the residual sum of squares decreases by a larger percentage than the degrees of freedom when adding a variable, then the Adjusted R^2 value will increase, and vice versa. This allows us to compare models with different numbers of explanatory variables more accurately.

2 Interpretation of the Coefficients

However, now the interpretation of the coefficients changes slightly. Let's keep thinking about the case with two explanatory variables X_1 and X_2.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{u}_i$$

In this scenario, $\hat{\beta}_0$ corresponds to the predicted value of Y when both X_1 and X_2 are set to 0. On the other hand, $\hat{\beta}_1$ denotes the impact of a 1-unit increase in X_1, assuming that X_2 remains constant. This is commonly expressed as "holding all other variables constant" or "ceteris paribus."

3 Assumptions for Causal Inference

There are five assumptions for causal inference in the multiple regression model. The first four are the same as the single-regressor model.

1. The population regression model is linear in its parameters and correctly specified as:
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$$
2. The observed data $(Y_i, X_{1i}, X_{2i}, \ldots, X_{ki})$ for $i = 1, 2, \ldots, n$ represent a random sample of size n from the above population model.
3. Large outliers are unlikely: $X_{1i}, X_{2i}, \ldots, X_{ki}$ have finite fourth moments.
4. u_i has conditional mean 0 given $X_{1i}, X_{2i}, \ldots, X_{ki}$
$$E(u_i | X_{1i}, X_{2i}, \ldots, X_{ki}) = 0$$
5. There is no perfect multicollinearity.

If assumptions 1-5 hold, then in large sample $\hat{\beta}_j \sim N(\beta_j, \sigma^2_{\hat{\beta}_j})$. Hence, we can test hypotheses and construct confidence intervals for our estimates as before.

The last assumption requires that the regressors (independent variables) do not exhibit perfect multicollinearity. The regressors are said to exhibit perfect multicollinearity (or to be perfectly multicollinear) if one of the variables is a perfect linear function of the others. At an intuitive level, perfect multicollinearity is a problem because you are asking the regression to answer an illogical question. In multiple regression, the coefficient on one of the regressors is the effect of a change in that regressor, holding the other regressors constant. Now, if the other regressor is a linear function of the first regressor, OLS cannot estimate this nonsensical partial effect.

Interpretation of coefficients: The coefficient β_k measures the effect of a one-unit change in X_k, while holding all other variables constant. To see this, let's consider a two-variable model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

To interpret the coefficient on X_1, we can take the expectation of Y conditional on $X_1 = x_1$ and $X_2 = x_2$, and note that the conditional expectation of the error term u is zero. This gives us:

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{1}$$

Now, let's consider the effect of a one-unit change in X_1, while holding X_2 constant at x_2. We can take the conditional expectation of Y again, this time with $X_1 = x_1 + 1$:

$$E(Y|X_1 = x_1 + 1, X_2 = x_2) = \beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2 \tag{2}$$

By subtracting equation (1) from (2), we obtain:

$$E(Y|X_1 = x_1 + 1, X_2 = x_2) - E(Y|X_1 = x_1, X_2 = x_2) = \beta_1$$

Thus, we see that β_1 represents the effect of a one-unit increase in X_1, holding X_2 constant.

Similarly, we can take the conditional expectation of Y with $X_2 = x_2 + 1$ and $X_1 = x_1$, and subtract it from equation (1) to obtain:

$$E(Y|X_1 = x_1, X_2 = x_2 + 1) - E(Y|X_1 = x_1, X_2 = x_2) = \beta_2$$

This shows that β_2 represents the effect of a one-unit increase in X_2, holding X_1 constant.

4 Control Variables

While there are cases where we might want to evaluate the effect of more than one variable, it is hard to find exogenous variables. A useful application of the multiple regression model is to control for the omitted variables while aiming to estimate the

causal effect of our variable of interest.

Consider the following multiple regression model with two independent variables:

$$Y = \beta_0 + \beta_1 X + \beta_2 W + u$$

In this model, X is our primary variable of interest. However, W is not only correlated with X but also influences Y. Ignoring W would result in a biased OLS estimator for β_1. According to the assumptions discussed earlier, as long as both W and X are jointly exogenous—meaning $E(u|W, X) = 0$—we can interpret β_1 as the causal impact of X while holding W constant and interpret β_2 as the causal impact of W while holding X constant.

However, W may not always satisfy the exogeneity condition. The good news is that it doesn't necessarily have to. Instead of adhering to Assumption (4), we can invoke the "conditional mean independence" assumption:

$$E(u|X, W) = E(u|W)$$

This implies that the error term u becomes independent of X once we control for W.

Replacing assumption (4) with conditional independence, β_1 can be interpreted as the causal effect of X on Y, while controlling for W. Note that β_2 may still be subject to bias.