Introduction to Multiple Regression

R Tutorials for Applied Statistics

Multiple regression analysis extends simple bivariate regression analysis with the inclusion of more than one explanatory variable.

The procedure is used to determine how one or more explanatory variables influence an outcome variable, *while holding all other explanatory variables fixed*.

1 Example: Monthly Earnings and Years of Education

The code below downloads a CSV file that includes data from 1980 for 935 individuals on variables including their total monthly earnings (`MonthlyEarnings`) and a number of variables that could influence income, including each person's IQ (`IQ`), a measure of knowledge of their job (`Knowledge`), years of education (`YearsEdu`), years experience (`YearsExperience`), and years at current job (`Tenure`). The data set originally comes from textbook website for Stock and Watson's *Introduction to Econometrics*.

```
wages <- read.csv("http://murraylax.org/datasets/wage2.csv");
```

Suppose we wish to estimate a linear multiple regression with the above five variables as explanatory variables. The population regression equation has the form,

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_k x_{k,i} + \epsilon_i ,$$

yi=β0+β1x1,i+β2x2,i+...+βkxk,i+ϵi,

where $k = 5$k=5 is the number of explanatory variables. The sample regression equation is therefore given by,

$$y_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + \ldots + b_k x_{k,i} + e_i ,$$

yi=b0+b1x1,i+b2x2,i+...+bkxk,i+ei,

We can use the `lm()` function to estimate the regression. In the code below we call `lm()` with a single parameter that is a formula specifying how the outcome variable, `MonthlyEarnings` depends linearly on the seven explanatory variables. The function `lm()` produces a large list of output, which we assign to an object we name `lmwages`. We call the `summary()` function next to display the coefficient estimates to the screen.

```r
lmwages <- lm(MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience
              + Tenure, data=wages)
summary(lmwages)
```

```
##
## Call:
## lm(formula = MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience +
##     Tenure, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -826.33 -243.85  -44.83  180.83 2253.35
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -531.0392   115.0513  -4.616 4.47e-06 ***
## IQ                 3.6966     0.9651   3.830 0.000137 ***
## Knowledge          8.2703     1.8273   4.526 6.79e-06 ***
## YearsEdu          47.2698     7.2980   6.477 1.51e-10 ***
## YearsExperience   11.8589     3.2494   3.650 0.000277 ***
## Tenure             6.2465     2.4565   2.543 0.011156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 365.4 on 929 degrees of freedom
## Multiple R-squared:  0.1878, Adjusted R-squared:  0.1834
## F-statistic: 42.97 on 5 and 929 DF,  p-value: < 2.2e-16
```

The results indicate the best fitting line is given by,

$$\hat{y}_i = -531.0 + 3.7x_{IQ,i} + 8.3x_{Knowledge,i} + 47.3x_{Edu,i} + 11.9x_{Exp,i} + 6.2x_{Tenure,i}$$

y^i=−531.0+3.7xIQ,i+8.3xKnowledge,i+47.3xEdu,i+11.9xExp,i+6.2xTenure,i.

2 Omitted Variable Bias

A multiple regression is even useful when one is only interested in the impact of *one* of the explanatory variables. The reason is that the regression analysis *holds constant* the effects of the other variables.

Suppose, for example, that you are primarily interested in estimating the impact of education on monthly earnings. Suppose you just estimated the following bivariate regression of education on earnings:

$$y_i = \beta_0 + \beta_1 x_{Edu,i} + \epsilon_i$$

yi=β0+β1xEdu,i+ϵi

The following call to `lm()` and `summary()` produce the estimates for the simple, bivariate model:

```
lmwages_bivariate <- lm(MonthlyEarnings ~ YearsEdu, data=wages)
summary(lmwages_bivariate)
```

```
##
## Call:
## lm(formula = MonthlyEarnings ~ YearsEdu, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -877.38 -268.63  -38.38  207.05 2148.26
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  146.952     77.715   1.891   0.0589 .
## YearsEdu      60.214      5.695  10.573   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 382.3 on 933 degrees of freedom
## Multiple R-squared:  0.107,  Adjusted R-squared:  0.106
## F-statistic: 111.8 on 1 and 933 DF,  p-value: < 2.2e-16
```

The bivariate regression predicts that each year of education is associated with an increase in monthly earnings of $60.21, while the multiple regression above predicted that each year of education is associated with an increase in monthly earnings of only $47.27.

The bivariate regression estimate is a *biased* estimate of the degree to which education affects monthly earnings, because it suffers from **omitted variable bias**. This is when there are possible variables that are not included in the regression which influence both the explanatory variable and the outcome variable. When such variables are omitted from the regression, the regression is not able to hold those effects constant. The regression coefficients estimates for the variables that are included may be picking the omitted effects.

One such omitted variable in the bivariate regression is cognitive ability. People with higher cognitive ability are likely to perform better in their careers and earn more income. This is true both for people who get high levels of education and low levels of education.It happens that people with higher cognitive ability receive higher levels of education on average. When measures of cognitive ability are omitted from the regression, the regression incorrectly attributes the full amount of the higher salaries to higher levels of education, when at least part of this should be attributed to higher cognitive ability.

The multiple regression includes multiple measures of cognitive ability, including IQ and job knowledge. When these measures of cognitive ability are held fixed, and only a marginal increase in education is considered, the estimated impact of education on monthly earnings is $47.27.

Suppose IQ were to be removed from the regression. Call this regression a *restricted* regression, and let b^r_{Edu}bEdur denote the regression coefficient from the restricted regression. We know that the restricted regression is susceptible to omitted variable bias. Let b_{Edu}bEdu and b_{IQ} bIQ denote the regression coefficients from the unrestricted multiple regression that includes IQ. It can be shown that the estimated coefficient on the restricted regression is given by,

$$b^r_{Edu} = b_{Edu} + b_{IQ} \left(d_{IQ,Edu} + d_{IQ,Knowledge} + d_{IQ,YearsExp} + d_{IQ,Tenure}\right)$$

bEdur=bEdu+bIQ(dIQ,Edu+dIQ,Knowledge+dIQ,YearsExp+dIQ,Tenure)

where $d_{Edu,j}$dEdu,j is the regression coefficient on the jthjth variable from a regression with the omitted variable as the outcome variable and all the remaining explanatory variables. The first term in the above equation is the coefficient on the unrestricted regression. All the other terms in the equation add up to the bias created by omitting the variable. The larger are the values for $d_{Edu,j}$dEdu,j, the more correlated the omitted variable is to the existing explanatory variables, the larger will be the bias. Also, the smaller is b_{IQ} bIQ, the less important is the omitted variable in the regression in the first place, the smaller will be the size of the bias.