Chapter 4
Multiple Regression Analysis

The simple linear regression covered in Chapter 2 can be generalized to include more than one variable. Multiple regression analysis is an extension of the simple regression analysis to cover cases in which the dependent variable is hypothesized to depend on more than one explanatory variable. While much of the analysis is an extension of the simple case, we have two main complications. (1) We need to discriminate between the effects of one variable and the effects of the other explanatory variables. (2) We have to decide which variables to include in the regression equation. In this chapter we will focus on the extension of the linear regression model and in (1). In a later chapter we will discuss (2).

4.1 Interpretation of the coefficients

Consider the following population multiple regression model with $(k-1)$ regressors:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + u. \tag{4.1}$$

A simple example of a multiple regression model is:

$$CRIME_i = \beta_1 + \beta_2 POPULATION_i + \beta_3 UNEMPLOY_i + \beta_4 POLICE_i + u_i, \tag{4.2}$$

where i refers to the city, $CRIME$ is crime rates, $POPULATION$ is just the number people in city i, $UNEMPLOY$ is the unemployment rate, and $POLICE$ is the number of police officers. To estimate the βs in Equation 4.2 you may need to observe crime rates and all the other variables for n cities. As before, u is the disturbance term. Because we have more that one regressor, the simple two dimensional characterization illustrated in Figure 2.1 is no longer applicable. Now, we have a $(k-1)$ dimensional problem. In our crime example we would need to have a 4D graph!

The sample counterpart of Equation 4.2 is:

$$CRIME_i = b_1 + b_2 POPULATION_i + b_3 UNEMPLOY_i + b_4 POLICE_i + e_i, \tag{4.3}$$

where the bs are the sample estimates of the βs, and are estimated using computer software via Ordinary Least Squares. We also express this relationship as the 4D fitted plane:

$$\widehat{CRIME}_i = b_1 + b_2 POPULATION_i + b_3 UNEMPLOY_i + b_4 POLICE_i. \tag{4.4}$$

Notice that we no longer write the disturbance term. Moreover, $\widehat{CRIME}_i$ is the fitted or predicted value of $CRIME_i$. The interpretation of the coefficients is the same as before. If the number of police officers increases by one, then the crime rate will change by b_4. Similar interpretation follows for b_2 and b_3.

4.2 Ordinary Least Squares

The OLS estimates are obtained in the same fashion as before. The unknown relationship is given by:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i. \tag{4.5}$$

The fitted OLS regression is:

$$\hat{Y}_i = b_1 + b_2 X_{2i} + b_3 X_{3i} + \cdots + b_k X_{ki}. \tag{4.6}$$

Then, the OLS regression residuals are:

$$e_i = Y_i - \hat{Y} = Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i} - \cdots - b_k X_{ki}. \tag{4.7}$$

Recall that OLS minimizes the sum of squared residuals

$$\min_{b_1, b_2, \ldots, b_k} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2, \tag{4.8}$$

where $RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ is the sum of squared residuals. We need to take the derivative of the RSS with respect to $b_1, b_2, \ldots, b_k$ and obtain k first order conditions. This yields a system of k equations with k unknowns, where the solution is the OLS estimators of the βs.

4.3 Assumptions

1. The model is linear in the parameters and correctly specified

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + u. \tag{4.9}$$

2. There is no exact linear relationship among the regressors in the sample. This is called multicollinearity.
3. The disturbance term has expectation zero

$$E(u_i) = 0 \qquad \text{for all} \quad i. \tag{4.10}$$

4. The disturbance term is homoscedastic.

$$\sigma^2_{u_i} = \sigma^2_u \qquad \text{for all} \quad i. \tag{4.11}$$

5. The values of the disturbance term have independent distributions.

$$u_i \text{ is distributed independently of } u_{i'} \text{ for all } i' \neq i. \tag{4.12}$$

6. The distribution term has a normal distribution.

$$u_i \sim N[0, \sigma^2] \qquad \text{for all} \quad i. \tag{4.13}$$

All the Xs are nonstochastic.

4.4 Properties of the coefficients

4.4.1 Unbiasedness

The OLS estimator b_j of β_j is unbiased:

$$E(b_j) = \beta_j \tag{4.14}$$

4.4.2 Efficiency

Following the results from the Gauss-Markov theorem, we have that OLS yields the most efficient linear estimators, in the sence that they are the one with the smallest variance among all linear estimators.

4.4.3 Precision of the coefficient, t tests, and confidence intervals

Beside our interest on the point estimates, we are also interested in performing hypotheses testing and building confidence intervals. To do this we need a measure of the precision of the coefficients. While we will not show the derivation here (as it required matrix algebra), each of the b_j has an standard error, S_{b_j}.

The null and alternative hypotheses about population coefficient j is written as:

$$H_0 : \beta_j = \beta_j^0 \tag{4.15}$$

$$H_1 : \beta_j \neq \beta_j^0. \tag{4.16}$$

which can be tested using the following t-statistic:

$$t = \frac{b_j - \beta_j^0}{S_{b_j}} \tag{4.17}$$

The null is not rejected if the following condition is met:

$$-t_{n-k,\alpha/2} \leq \frac{b_j - \beta_j^0}{S_{b_j}} \leq t_{n-k,\alpha/2} \tag{4.18}$$

Notice the difference between Equation 4.18 and Equation 3.25. The critical value from the t distribution, $t_{n-k,\alpha/2}$, now has $n-k$ degrees of freedom because we are estimating k parameters, rather than just 2 as in the simple regression model. The intuition behind Figures 3.1 and 3.1 still hold. The computer software will also give you the p-value associated with the t test. If the p-value is below your α, you reject the null hypothesis.

For the construction of the confidence intervals we have:

$$\begin{aligned} 1-\alpha &= P\big(-t_{n-k,\alpha/2} \leq \tfrac{b_j-\beta_j}{S_{b_j}} \leq t_{n-k,\alpha/2}\big) \\ 1-\alpha &= P\big(-t_{n-k,\alpha/2} \cdot S_{b_j} \leq b_j - \beta_j \leq t_{n-k,\alpha/2} \cdot S_{b_j}\big) \\ 1-\alpha &= P\big(b_j - t_{n-k,\alpha/2} \cdot S_{b_j} \leq \beta_j \leq b_j + t_{n-k,\alpha/2} \cdot S_{b_j}\big). \end{aligned} \tag{4.19}$$

4.5 Regression output in Gretl

Gretl is an open-source (free) software package for econometric analysis written in the C programming language. It can be downloaded from:

`http://gretl.sourceforge.net/`

Just follow the instructions to install it in your computer.

Once you loaded the data set in Gretl, to estimate Equation 4.2 you need to go to `Model` → `Ordinary Least Squares`. The regression output is:

```
Model 1: OLS, using observations 1-92
Dependent variable: crimes

             coefficient     std. error    t-ratio   p-value
  ------------------------------------------------------------
  const      2193.34        3918.06         0.5598   0.5770
  pop           0.0652716      0.0106262    6.143    2.30e-08 ***
  unem       -279.291         407.791       -0.6849   0.4952
  officers     15.0406          3.57660      4.205    6.25e-05 ***

Mean dependent var   39663.53   S.D. dependent var   29692.10
Sum squared resid    1.39e+10   S.E. of regression   12548.04
R-squared            0.827293   Adjusted R-squared   0.821405
F(3, 88)             140.5107   P-value(F)           1.90e-33
Log-likelihood      -996.7310   Akaike criterion     2001.462
Schwarz criterion    2011.549   Hannan-Quinn         2005.533

Excluding the constant, p-value was highest for variable 3 (unem)
```

A standard way to present the regression output is:

$$\underset{}{\widehat{\text{crimes}}} = \underset{(3918.1)}{2193.34} - \underset{(407.79)}{279.291}\,\text{unem} + \underset{(3.5766)}{15.0406}\,\text{officers} + \underset{(0.010626)}{0.0652716}\,\text{pop}$$

$$N = 92 \quad \bar{R}^2 = 0.8214 \quad F(3,88) = 140.51 \quad \hat{\sigma} = 12548.$$

(standard errors in parentheses)

To obtain the confidence intervals for the coefficients as presented in Equation 4.19 in the Gretl regression output window you need to go to `Analysis` → `Confidence intervals for the coefficients` to obtain:

```
t(88, 0.025) = 1.987

    VARIABLE         COEFFICIENT      95% CONFIDENCE INTERVAL

         const       2193.34             -5592.97      9979.66
           pop          0.0652716        0.0441542    0.0863890
         unem        -279.291            -1089.69      531.107
     officers          15.0406             7.93282      22.1483
```

4.6 Multicollinearity

Multicollinearity is when two explanatory variables are highly correlated. In addition, if their coefficients have a large population variances, we are at risk of getting erratic estimates of the coefficients. There could also be multicollinearity when there is an approximate linear relationship between more than two variables.

A simple test for multicollinearity is based in the Variance Inflation Factors. To implement this text in Gretl, in the regression output window go to `Test` → `Collinearity`:

```
Variance Inflation Factors
```

```
Minimum possible value = 1.0
Values > 10.0 may indicate a collinearity problem

            pop    4.180
           unem    1.094
       officers    4.371

VIF(j) = 1/(1 - R(j)^2), where R(j) is the multiple correlation
coefficient between variable j and the other independent variables

Properties of matrix X'X:

 1-norm = 2.0266173e+013
 Determinant = 6.6859257e+024
 Reciprocal condition number = 4.681685e-013
```

Based on these results, we do not have a multicollinearity problem in the estimation of Equation 4.2.

4.7 Goodness of fit: R^2 and $\bar{R}^2$

The R^2 in multiple regression analysis has the same interpretation as in a simple regression. It is the proportion of the variation in Y explained by the regression model

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \tag{4.20}$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{4.21}$$

where $\hat{Y}$ represents the fitted values of the regression equation

$$\hat{Y} = b_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k. \tag{4.22}$$

4.8 F tests

Given the population regression model

$$Y = \beta_1 + \beta_2X_2 + \beta_3X_3 + \cdots + \beta_kX_k + u, \tag{4.23}$$

we can use the F test to test if all the slope coefficients $\beta_2, \beta_3, \ldots, \beta_k$ are jointly equal to zero. That is, let the null hypothesis be:

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0. \tag{4.24}$$

The alternative hypothesis (H_0) is that at least one of the slope coefficients is different from zero. The multiple regression version of the F statistic is:

$$F_{k-1,n-k} = \frac{ESS/(k-1)}{RSS/(n-k)}. \tag{4.25}$$

The idea is to compare this F statistic to the critical level found in the F distribution tables with $k-1$ and $n-k$ degrees of freedom. Computer software automatically computes this F statistic and the corresponding p-value for the null in Equation 4.22. This F statistic can also be written in terms of the R^2:

$$F_{k-1,n-k} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}. \tag{4.26}$$

Consider the example presented in Section 4.5. The F statistic is 140.5107 with 3 and 88 degrees of freedom and has a corresponding p-value of 0.000. Then, because the p-value is below $\alpha = 5\%$ then we reject the null hypothesis that the slope coefficients on `pop`, `unem`, and `officers` are jointly equal to zero.

4.9 Adjusted R^2, $\bar{R}^2$

One concern with the R^2 is that it will always go up as we include more variables into the model. Hence, it is a poor way to compare models. On the other hand a similar statistic, the adjusted R^2 ($\bar{R}^2$) is built on the R^2 but with the difference that $\bar{R}^2$ penalizes for the loss of the degrees of freedom as we include more variables into the model. Therefore, the $\bar{R}^2$ can either go up or down as we include more variable into the model. It is defined as:

$$\bar{R}^2 = R^2 - \frac{k-1}{n-k}(1-R^2). \tag{4.27}$$