Multiple Regression Model

Fall 2008

The Multiple Regression Model

- In practice, the key assumption in the simple regression model

$$E\left(u_i|X\right) = 0$$

  is often unrealistic.
- We need to explicitly control for many other (observable) factors that simultaneously affect the dependent variable Y.
- The multiple regression model takes the following form:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + u_i.$$

- The model includes k independent variables and one constant. Thus, there will be $k+1$ parameters to estimate.
- The error term u_i contains factors other than $X_1, ..., X_k$ that affect Y.

Assumption and Interpretation

- Assumption MLR.1 (zero-conditional mean)

$$E\left(u_i | X_1, ..., X_k\right) = 0.$$

- It implies that all independent variables are uncorrelated with the error term.
- The assumption leads to a well-defined ceteris paribus analysis: each coefficient, β_j, measures the impact of the corresponding variable, X_j, on Y, holding all other factors constant.
- Mathematically,

$$\beta_j = \frac{\partial Y_j}{\partial X_{ij}}.$$

Example 1 - Housing Prices and Air Pollution

- Model 1: $\ln(\text{Hprice}_i) = \beta_0 + \beta_1 \ln(\text{Nox}_i) + \varepsilon_i$

Variable	Coefficient	St. Err.
Constant	11.707	0.132
log Nox	-1.043	0.078

Example 1 - Housing Prices and Air Pollution

- Model 1: $\ln(\text{Hprice}_i) = \beta_0 + \beta_1 \ln(\text{Nox}_i) + \varepsilon_i$

Variable	Coefficient	St. Err.
Constant	11.707	0.132
log Nox	-1.043	0.078

- Model 2: $\ln(\text{Hprice}_i) = \beta_0 + \beta_1 \ln(\text{Nox}_i) + \beta_2 \ln(\text{Proptax}_i) + \varepsilon_i$

Variable	Coefficient	St. Err.
Constant	13.176	0.224
log Nox	-0.523	0.098
log Proptax	-0.396	0.050

Multiple Regression with Dummy Variables

- The multiple regression model often contains qualitative factors, which are not measured in any units, as independent variables:
  - gender, race or nationality
  - employment status or home ownership
  - temperatures before 1900 and after (including) 1900
- Such qualitative factors often come in the form of binary information and are captured by defininig a zero-one variable, called *dummy variables*.

$$D_i = \begin{cases} 0 & \text{if} \quad \text{year}_i < 1900 \\ 1 & \text{if} \quad \text{year}_i \geq 1900 \end{cases}$$

Dummy Variables: Intercept Shift

- The dummy variable can be used to build a model with an intercept that varies across groups coded by the dummy variable.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i$$

Dummy Variables: Intercept Shift

- The dummy variable can be used to build a model with an intercept that varies across groups coded by the dummy variable.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i$$

- The model can be interpreted that the observations for which $D_i = 1$ have, on average, a Y_i which is β_2 units higher than otherwise.

Dummy Variables: Intercept Shift

- The dummy variable can be used to build a model with an intercept that varies across groups coded by the dummy variable.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i$$

- The model can be interpreted that the observations for which $D_i = 1$ have, on average, a Y_i which is β_2 units higher than otherwise.
- Example: $\ln(Temp_i) = \beta_0 + \beta_1 \ln(Co2_i) + \beta_2 D_i + u_i$, where

$$D_i = \begin{cases} 0 & \text{if} \quad \text{year}_i < 1900 \\ 1 & \text{if} \quad \text{year}_i \geq 1900 \end{cases}$$

Variable	Coefficient	St. Err.
Constant	0.837	0.708
log CO2	0.243	0.126
Time Dummy	0.010	0.016

Dummy Variables: Slope Shift

- The dummy variable can be also used to vary a slope of one (continuous) independent variable across groups.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i X_i + u_i$$

Dummy Variables: Slope Shift

- The dummy variable can be also used to vary a slope of one (continuous) independent variable across groups.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i X_i + u_i$$

- For observations with $D_i = 0$, a one unit increase in X_i leads to an increase of β_1 units in Y_i. For those with $D_i = 1$, Y_i increases by $(\beta_1 + \beta_2)$ units in Y_i.

Dummy Variables: Slope Shift

- The dummy variable can be also used to vary a slope of one (continuous) independent variable across groups.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i X_i + u_i$$

- For observations with $D_i = 0$, a one unit increase in X_i leads to an increase of β_1 units in Y_i. For those with $D_i = 1$, Y_i increases by $(\beta_1 + \beta_2)$ units in Y_i.
- Example: $\ln(Temp_i) = \beta_0 + \beta_1 \ln(Co2_i) + \beta_2 D_i \ln(Co2_i) + u_i$,

Variable	Coefficient	St. Err.
Constant	0.854	0.719
log CO2	0.240	0.127
Dummy*log CO2	0.002	0.003

Ordinary Least Squares Estimator

- Just as in the simple regression model, the OLS estimator in the multiple regression model is chosen to minimize the sum of squared residuals:

$$\min_{\{\widehat{\beta}_j\}_{j=0}^{k}} \sum_{i=1}^{N} \widehat{u}_i^2 = \sum_{i=1}^{N} \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{i1} - \widehat{\beta}_2 X_{i2} - \ldots - \widehat{\beta}_k X_{ik} \right)^2$$

- By taking a (partial) derivative with respect to each coefficient, we obtain a set of $(k+1)$ equations constituting the first-order conditions for minimizing the sum of squared residuals. These equations are often called the *normal equations*.
- Then, we have the OLS or sample regression line:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{i1} + \widehat{\beta}_2 X_{i2} + \ldots + \widehat{\beta}_k X_{ik}.$$

- Each estimate, $\widehat{\beta}_j$, has a **partial effect** or **ceteris paribus** interpretation: the effect of X_j on Y, while holding other factors constant.

Algebraic Properties of OLS

- Property 1.

$$\sum_{i=1}^{N} \widehat{u}_i = \sum_{i=1}^{N} \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{i1} - \widehat{\beta}_2 X_{i2} - ... - \widehat{\beta}_k X_{ik} \right) = 0.$$

Algebraic Properties of OLS

- Property 1.

$$\sum_{i=1}^{N} \widehat{u}_i = \sum_{i=1}^{N} \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{i1} - \widehat{\beta}_2 X_{i2} - \ldots - \widehat{\beta}_k X_{ik} \right) = 0.$$

- Property 2.

$$\sum_{i=1}^{N} \widehat{u}_i X_{ij} = 0, \forall j = 1, 2, \ldots, k.$$

Algebraic Properties of OLS

- Property 1.

$$\sum_{i=1}^{N} \widehat{u}_i = \sum_{i=1}^{N} \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{i1} - \widehat{\beta}_2 X_{i2} - ... - \widehat{\beta}_k X_{ik} \right) = 0.$$

- Property 2.

$$\sum_{i=1}^{N} \widehat{u}_i X_{ij} = 0, \forall j = 1, 2, ..., k.$$

- Property 3. From Property 1 and $Y_i = \widehat{Y}_i + \widehat{u}_i$,

$$\overline{Y} = \overline{\widehat{Y}}.$$

Algebraic Properties of OLS

- Property 1.
$$\sum_{i=1}^{N} \widehat{u}_i = \sum_{i=1}^{N} \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{i1} - \widehat{\beta}_2 X_{i2} - ... - \widehat{\beta}_k X_{ik} \right) = 0.$$
- Property 2.
$$\sum_{i=1}^{N} \widehat{u}_i X_{ij} = 0, \forall j = 1, 2, ..., k.$$
- Property 3. From Property 1 and $Y_i = \widehat{Y}_i + \widehat{u}_i$,
$$\overline{Y} = \overline{\widehat{Y}}.$$
- Property 4. The point $\left(\overline{Y}, \overline{X}_1, \overline{X}_2, ..., \overline{X}_k\right)$ is always on the OLS regression line:
$$\overline{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{X}_1 + \widehat{\beta}_2 \overline{X}_2 + ... + \widehat{\beta}_k \overline{X}_k.$$

A Case for Two Independent Variables

- Consider the case with $k = 2$ independent variables:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{i1} + \widehat{\beta}_2 X_{i2}.$$

- The solution for $\widehat{\beta}_1$ is

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{N} \widehat{R}_{i1} Y_i}{\sum_{i=1}^{N} \widehat{R}_{i1}^2},$$

  where the $\widehat{R}_{i1}$ are the OLS residuals from a simple regression of X_1 on X_2.
- Note that the residuals $\widehat{R}_{i1}$ have a zero sample average and thus $\widehat{\beta}_1$ is the usual slope estimate from the simple regression of Y_i on $\widehat{R}_{i1}$.
- The residuals $\widehat{R}_{i1}$ is X_{i1} after the effects of X_{i2} have been *partialled out* or *netted out*. Thus, $\widehat{\beta}_1$ measures the sample relationship between Y and X_1 after X_2 has been partialled out.

Goodness of Fit

- As with simple regression, we can define the R-squared:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \widehat{u}_i^2}{\sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2}.$$

- An important fact in R^2 is that it never decreases in the number of independent variables.

- This algebraic fact follows because the sum of squared residuals never increases when additional regressors are added to the model. Thus, just looking at R^2 does not tell us whether an additional independent variable improves the fit.
- One convention is the idea of imposing a penalty for adding additional independent variables to a model, adjusted R^2,

$$\overline{R}^2 = 1 - \frac{\sum_{i=1}^{N} \widehat{u}_i^2 / (N-k-1)}{\sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2 / (N-1)} = 1 - \left(1 - R^2\right) \frac{N-1}{N-k-1}.$$

An Example: Housing Price

- To investigate the deteminants of log housing prices, we include as independent variables: log Nitrogen oxide, log dist, rooms, stratio, and log property tax.

Variable	Coefficient	Std. Err.
Constant	11.798	0.340
log nox	-0.718	0.123
log dist	-0.143	0.042
rooms	0.252	0.018
stratio	-0.041	0.006
log proptax	-0.217	0.042
R^2	0.605	
adjusted R^2	0.601	

Statistical Properties of OLS

- We now turn to the statistical properties of OLS in the multiple regression model for estimating the parameters in an underlying population model.
- As with simple regression, we can obtain the unbiasedness and the efficiency of the OLS estimators with direct extensions of the simple regression model assumptions.
- When an important variable is omitted from the regression, OLS produces the bias, called *Omitted Variable Bias*.
- When an irrelevant variable is included, the regression does not affect the unbiasedness of the OLS estimators but increase their variances.

Assumptions I

- **Assumption MLR1** (zero conditional mean):

$$E\left(u_i | X_1, ..., X_k\right) = 0.$$

- Failure of MLR1
  - omitting a variable
  - measurement error
  - endogeneity bias
- **Assumption MLR 2** (Homoskedasticity):

$$Var\left(u_i | X_1, ..., X_k\right) = \sigma^2.$$

Assumptions II

- **Assumption MLR 3** (no perfect collinearity): There are no *exact linear* relationships among the independent variables.
- Examples of failure of MLR2
  - same independent variable measured in different units
  - one variable is a constant multiple of another: $\ln(X)$ and $\ln(X^2)$
  - regression with a constant term, D_i (dummy variable) and $1 - D_i$.

Unbiasedness and Efficiency of OLS

- (Unbiasedness of OLS) Under Assumptions MLR1 and MLR3,

$$E\left(\widehat{\beta}_k|X\right) = \beta_k, \text{ for } j = 0, 1, ..., k.$$

Unbiasedness and Efficiency of OLS

- (Unbiasedness of OLS) Under Assumptions MLR1 and MLR3,

$$E\left(\widehat{\beta}_k|X\right) = \beta_k, \text{ for } j = 0, 1, ..., k.$$

- (Gauss-Markov Theorem) Under Assumptions MLR 1 through MLR3, $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k$ are the best linear unbiased estimators (BLUE) for the true parameters, $\beta_0, \beta_1, ..., \beta_k$.

Omitted Variable Bias I

- Suppose that the true regression relationship has the following form:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i.$$

- Instead we decide to estimate

$$Y_i = \beta_0 + \beta_1 X_{i1} + \nu_i.$$

- From the OLS of the second regression equation, we will obtain

$$\widetilde{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{N} \left(X_{i1} - \overline{X}_1\right) \nu_i}{\sum_{i=1}^{N} \left(X_{i1} - \overline{X}_1\right)^2}$$

- What is the expected value of the last expression on the right hand side?

Omitted Variable Bias II

- First note that $\nu_i = \beta_2 X_{i2} + u_i$.

Omitted Variable Bias II

- First note that $\nu_i = \beta_2 X_{i2} + u_i$.
- Substituting this into the expression for OLS estimator, we obtain

$$\widetilde{\beta}_1 = \beta_1 + \frac{\beta_2 \sum_{i=1}^{N} \left(X_{i1} - \overline{X}_1\right) X_{i2} + \sum_{i=1}^{N} \left(X_{i1} - \overline{X}_1\right) u_i}{\sum_{i=1}^{N} \left(X_{i1} - \overline{X}_1\right)^2}.$$

Omitted Variable Bias II

- First note that $\nu_i = \beta_2 X_{i2} + u_i$.
- Substituting this into the expression for OLS estimator, we obtain

$$\widetilde{\beta}_1 = \beta_1 + \frac{\beta_2 \sum_{i=1}^{N} \left(X_{i1} - \overline{X}_1\right) X_{i2} + \sum_{i=1}^{N} \left(X_{i1} - \overline{X}_1\right) u_i}{\sum_{i=1}^{N} \left(X_{i1} - \overline{X}_1\right)^2}.$$

- Taking the expectation, we have

$$\begin{aligned} E\left(\widetilde{\beta}_1 | X\right) &= \beta_1 \\ &\quad + \frac{\beta_2 \sum_{i=1}^{N} \left(X_{i1} - \overline{X}_1\right) X_{i2} + \sum_{i=1}^{N} \left(X_{i1} - \overline{X}_1\right) E\left(u_i | X\right)}{\sum_{i=1}^{N} \left(X_{i1} - \overline{X}_1\right)^2} \\ &= \beta_1 + \beta_2 \frac{\sum_{i=1}^{N} \left(X_{i1} - \overline{X}_1\right) X_{i2}}{\sum_{i=1}^{N} \left(X_{i1} - \overline{X}_1\right)^2} \\ &= \beta_1 + \beta_2 \widehat{Cov(X_1, X_2)} / \widehat{Var(X_1)}. \end{aligned}$$

Omitted Variable Bias III

- Thus, the size of the omitted variable bias is

$$Bias\left(\widetilde{\beta}_1\right) = E\left(\widetilde{\beta}_1 | X\right) - \beta_1 = \beta_2 \frac{\widehat{Cov(X_1, X_2)}}{\widehat{Var\left(X_1\right)}}.$$

Omitted Variable Bias III

- Thus, the size of the omitted variable bias is

$$Bias\left(\widetilde{\beta}_1\right) = E\left(\widetilde{\beta}_1|X\right) - \beta_1 = \beta_2 \frac{\widehat{Cov(X_1, X_2)}}{\widehat{Var\left(X_1\right)}}.$$

- There are two cases in which the bias is zero:

Omitted Variable Bias III

- Thus, the size of the omitted variable bias is

$$Bias\left(\widetilde{\beta}_1\right) = E\left(\widetilde{\beta}_1 | X\right) - \beta_1 = \beta_2 \frac{\widehat{Cov(X_1, X_2)}}{\widehat{Var\left(X_1\right)}}.$$

- There are two cases in which the bias is zero:
  - $\beta_2 = 0$.

Omitted Variable Bias III

- Thus, the size of the omitted variable bias is

$$Bias\left(\widetilde{\beta}_1\right) = E\left(\widetilde{\beta}_1 | X\right) - \beta_1 = \beta_2 \frac{\widehat{Cov(X_1, X_2)}}{\widehat{Var\left(X_1\right)}}.$$

- There are two cases in which the bias is zero:
  - $\beta_2 = 0$.
  - $\widehat{Cov(X_1, X_2)} = 0$.

Omitted Variable Bias III

- Thus, the size of the omitted variable bias is

$$Bias\left(\widetilde{\beta}_1\right) = E\left(\widetilde{\beta}_1 | X\right) - \beta_1 = \beta_2 \frac{\widehat{Cov(X_1, X_2)}}{\widehat{Var(X_1)}}.$$

- There are two cases in which the bias is zero:
  - $\beta_2 = 0$.
  - $\widehat{Cov(X_1, X_2)} = 0$.

- Thus, in general, omitting variables, which have an impact on Y, wil bias the OLS estimator of the coefficients of the included variables unless the omitted variables are uncorrelated with the included ones.

Omitted Variable Bias III

- Thus, the size of the omitted variable bias is

$$Bias\left(\widetilde{\beta}_1\right) = E\left(\widetilde{\beta}_1|X\right) - \beta_1 = \beta_2 \frac{\widehat{Cov(X_1, X_2)}}{\widehat{Var}(X_1)}.$$

- There are two cases in which the bias is zero:
  - $\beta_2 = 0$.
  - $\widehat{Cov(X_1, X_2)} = 0$.
- Thus, in general, omitting variables, which have an impact on Y, wil bias the OLS estimator of the coefficients of the included variables unless the omitted variables are uncorrelated with the included ones.
- The direction and size of the bias (negative or positive bias) depend on the signs and sizes of β_2 and $\widehat{Cov(X_1, X_2)}$.

An Example: Housing Prices

- Suppose the true model is

$$\ln\left(Hprice_i\right) = \beta_0 + \beta_1 \ln\left(Nox_i\right) + \beta_2 \ln\left(proptax_i\right) + u_i.$$

- BUT, one omits the proptax variable in the regression:

$$\ln\left(Hprice_i\right) = \beta_0 + \beta_1 \ln\left(Nox_i\right) + \nu_i.$$

Var.	Coeff.	St. Err.	Var.	Coeff.	St. Err.
Constant	11.707	0.132	Constant	13.176	0.224
log Nox	-1.043	0.078	log Nox	-0.523	0.098
			log Proptax	-0.396	0.050

- The sample correlation between log Nox and log Proptax is 0.667.

Including an Irrelevant Variable I

- Suppose the true model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + u_i.$$

- But, we include an irrelevant variable, X_{i2}, in a regression and have an estimate $\widetilde{\beta}_1$. Let $\widehat{\beta}_1$ be the OLS estimator from the correct specification.
- It can be shown that $E\left(\widetilde{\beta}_1 | X\right) = \beta_1$.

Including an Irrelevant Variable II

- For the variances, we have the following relationship:

$$\begin{aligned} Var\left(\widehat{\beta}_1 | X\right) &= \frac{\sigma^2}{\sum_{i=1}^{N}\left(X_{i1} - \overline{X}_1\right)^2} \\ &\leq \frac{\sigma^2}{\left(1 - R_1^2\right)\sum_{i=1}^{N}\left(X_{i1} - \overline{X}_1\right)^2} = Var\left(\widetilde{\beta}_1 | X\right), \end{aligned}$$

  where R_1^2 is the R-squared from the regression of X_1 on X_2.

- Unless X_1 and X_2 are uncorrelated in the sample, including X_2 increases the variance for the estimator of β_1.