Introductory Econometrics

Lecture 5: Multiple Regression, Properties

Alex Alekseev

November 14, 2022

University of Regensburg, Department of Economics

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

$$Y = \mathbf{x}'\beta + U,$$

Recap

Population model

- In the last lecture, we have introduced the multiple regression model
- The population version of this model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + U.$$

- We have also learned that we can write this compactly using the tools of linear algebra as

$$Y = \mathbf{x}'\beta + U,$$

where

$$\mathbf{x} = \begin{pmatrix} X_0 \\ \vdots \\ X_k \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}, \quad X_0 = 1.$$

Sample model

- The sample regression model takes the form

$$y_i = \hat{\beta}_0 x_{i0} + \ldots + \hat{\beta}_k x_{ik} + \hat{u}_i, \quad i = 1, \ldots, n, \quad x_{i0} = 1, \text{ for } i = 1, \ldots, n.$$

- The sample version can be compactly written in a vector form as

$$\mathbf{y} = \mathbf{X}\hat{\beta} + \hat{\mathbf{u}},$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{10} \ldots x_{1k} \\ \vdots \ddots \vdots \\ x_{n0} \ldots x_{nk} \end{pmatrix}, \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}, \quad \hat{\mathbf{u}} = \begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_n \end{pmatrix}.$$

OLS

- We then showed that the OLS estimator of the coefficients is

OLS formula

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Assumptions

1. **Linear CEF**
   The CEF of Y given $X_1, X_2, \ldots, X_k$ is linear:
   $$\mathbb{E}[Y \mid X_1, X_2, \ldots, X_k] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k,$$
   which allows us to write down the population model as we wrote it and implies that U is mean-independent of $X_1, X_2, \ldots, X_k$: $\mathbb{E}[U \mid X_1, X_2, \ldots, X_k] = 0$.
2. **Random Sampling**
   Our sample $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)_{i=1}^n$ is a random sample from the population, i.e., the observations are pairwise independent and identically distributed. This assumption allows us to derive the statistical properties of the OLS.
3. **No Perfect Collinearity**
   None of the predictors is a linear combination of other predictors. This assumption guarantees that the OLS estimator exists.

Algebraic properties of the OLS

Residuals and predictors

- The sample covariance between each predictor and the residuals is zero **by construction**
- The mean of the residuals is zero **by construction**
- Start with the normal equations written in the matrix form

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}.$$

- Now substitute for $\mathbf{y}$

$$\begin{aligned} \mathbf{X}'\mathbf{X}\hat{\beta} &= \mathbf{X}'(\mathbf{X}\hat{\beta} + \hat{\mathbf{u}}) \\ \mathbf{X}'\mathbf{X}\hat{\beta} &= \mathbf{X}'\mathbf{X}\hat{\beta} + \mathbf{X}'\hat{\mathbf{u}} \\ \mathbf{X}'\hat{\mathbf{u}} &= 0. \end{aligned}$$

Linear algebra: Transpose

The transpose operation, $'$, flips the rows and columns of an object: the first column in the new object is the first row in the old object, etc.

$$\begin{pmatrix} x_{10} \dots x_{1k} \\ \vdots \ddots \vdots \\ x_{n0} \dots x_{nk} \end{pmatrix}' = \begin{pmatrix} x_{10} \dots x_{n0} \\ \vdots \ddots \vdots \\ x_{1k} \dots x_{nk} \end{pmatrix}$$

Residuals and predictors

- The product on the left is

$$\mathbf{X}'\hat{\mathbf{u}} = \begin{pmatrix} x_{10} \dots x_{n0} \\ \vdots \ddots \vdots \\ x_{1k} \dots x_{nk} \end{pmatrix} \begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_{i0}\hat{u}_i \\ \vdots \\ \sum_{i=1}^n x_{ik}\hat{u}_i . \end{pmatrix}$$

- Therefore, each term $\sum_{i=1}^n x_{ij}\hat{u}_i = 0, j = 1, \dots, k$ and hence $\frac{1}{n}\sum_{i=1}^n x_{ij}\hat{u}_i = 0, j = 1, \dots, k$
- These terms are the sample covariances between the predictors and the residuals
- Notice also that since $x_{i0} = 1$ for $i = 1, \dots, k$, then

$$\sum_{i=1}^n x_{i0}\hat{u}_i = \sum_{i=1}^n \hat{u}_i = 0,$$

which implies that the mean of the residuals is zero

Residuals and fitted values

- The sample covariance between the fitted values and residuals is zero **by construction**.
- Consider the term

$$\hat{\mathbf{y}}'\hat{\mathbf{u}} = (\mathbf{X}\hat{\beta})'\hat{\mathbf{u}} = \hat{\beta}'\mathbf{X}'\hat{\mathbf{u}} = 0.$$

- Now notice that this term also equals

$$\hat{\mathbf{y}}'\hat{\mathbf{u}} = (\hat{y}_1, \ldots, \hat{y}_n)\begin{pmatrix}\hat{u}_1 \\ \vdots \\ \hat{u}_n\end{pmatrix} = \sum_{i=1}^{n} \hat{y}_i \hat{u}_i.$$

- Therefore, $\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0$ and hence $\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0$, which is the sample covariance between the fitted values and the residuals.

Mean of predictors and mean of outcome

- If we plug in the means of the predictors in the equation for the regression line, we get the mean of the outcome.

Linear algebra: Summation

Define

$$\mathbf{1} \equiv \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Then $\mathbf{1}'$ is the summation operator:

$$\mathbf{1}'\mathbf{y} = \sum_{i=1}^{n} y_i.$$

Mean of predictors and mean of outcome

- If we plug in the means of the predictors in the equation for the regression line, we get the mean of the outcome.
- Consider the sample regression equation

$$\mathbf{y} = \mathbf{X}\hat{\beta} + \hat{\mathbf{u}}.$$

- Now pre-multiple by the summation operator.

$$\mathbf{1}'\mathbf{y} = \mathbf{1}'\mathbf{X}\hat{\beta} + \mathbf{1}'\widehat{\mathbf{u}}.$$

- Notice that the last term $\mathbf{1}'\widehat{\mathbf{u}}$ is zero. Also notice that

$$\mathbf{1}'\mathbf{X} = (1, \ldots, 1)\begin{pmatrix} x_{10} \ldots x_{1k} \\ \vdots \ddots \vdots \\ x_{n0} \ldots x_{nk} \end{pmatrix} = \left(\sum_{i=1}^{n} x_{i0}, \ldots, \sum_{i=1}^{n} x_{ik}\right)$$

Mean of predictors and mean of outcome

- Therefore, we have that

$$\sum_{i=1}^{n} y_i = \left(\sum_{i=1}^{n} x_{i0}, \ldots, \sum_{i=1}^{n} x_{ik} \right) \hat{\beta}$$
$$\sum_{i=1}^{n} y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_{i0} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}$$

- After dividing both sides by n, we get

$$\bar{y} = \beta_0 + \beta_1 \bar{x}_1 + \ldots + \beta_k \bar{x}_k.$$

Goodness-of-fit

Sums of squares

- The **total** sum of squares (SST) is

$$SST \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

- The **explained** sum of squares (SSE) is

$$SSE \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

- The **residual** sum of squares (SSR) is

$$SSR \equiv \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{u}_2^2.$$

- The identity continues to hold (the proof is identical)

$$SST = SSE + SSR$$

Root mean squared error

- The root mean squared error (RMSE) is computed as before

Root mean squared error

$$RMSE \equiv \sqrt{\frac{SSR}{n}}$$

R-squared

- The R-squared is computed as before:

R-squared

$$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- The property that R-squared is equal to the square of the correlation between the observed and fitted values still holds in the multiple regression

R-squared

- The R-squared is computed as before:

R-squared

$$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- The property that R-squared is equal to the square of the correlation between the observed and fitted values still holds in the multiple regression

Note

Comparing the observed and fitted values using a correlation or some other metric (e.g., accuracy) is, in general, a useful measure of a model's fit even when using R-squared does not make much sense (e.g., in classification problems)

R-squared

- The R-squared is computed as before:

R-squared

$$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- The property that R-squared is equal to the square of the correlation between the observed and fitted values still holds in the multiple regression
- However, the property that R-squared is equal to the square of the correlation between the outcome and the predictor **no longer holds**

Adjusted R-squared

- Including more predictors in the model mechanically increases R-squared
- To adjust for this, we can use the following formula:

$$R^2_{adj} = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}$$

- The adjusted R-squared
  - can increase or decrease when including an additional predictor
  - always increases if an additional predictor reduces the estimate of the error variance
- In other words, while including a "useless" predictor increases R-squared, it might actually reduce the adjusted R-squared

Regression anatomy

Individual coefficients

- From the simple linear regression, we know that the estimate of the slope coefficient is ratio of the sample covariance between the outcome and the predictor and the sample variance of the predictor:

$$\hat{\beta}_1 = \frac{\widehat{Cov}(Y, X)}{\widehat{Var}(X)}.$$

- In the multiple regression, we get the expression for the whole vector of coefficients

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

- But can we derive an expression for each individual coefficient β_j that would be analogous to the formula from the simple regression?

Wrong answer

Using a direct analogy will **not** work, i.e., in the multiple regression, each β_j **is not** equal to

$$\frac{\widehat{Cov}(Y, X_j)}{\widehat{Var}(X_j)}$$

Regression anatomy formula

- It turns out that each individual coefficient equals

Regression anatomy formula

$$\hat{\beta}_j = \frac{\widehat{Cov}(Y, U_j)}{\widehat{Var}(U_j)},$$

where U_j is the error term from the regression of X_j on all other predictors:

$$X_j = \gamma_0 + \gamma_1 X_1 + \ldots + \gamma_{j-1} X_{j-1} + \gamma_{j+1} X_{j+1} + \ldots + \gamma_k X_k + U_j.$$

- This result is also known as the Frisch-Waugh-Lovell theorem

Regression anatomy explained

- In other words, we can estimate an individual slope coefficient j from the multiple regression by
  1. Regressing predictor j on all other predictors (the auxiliary regression)
  2. Taking the residuals from that regression
  3. Regressing the outcome on those residuals
- This is another sense in which each individual slope coefficient captures the partial effect of an individual predictor on the outcome

Proof

- First consider the sample covariance in the numerator

$$\widehat{Cov}(Y, U_j) = \frac{1}{n}\sum_{i=1}^{n} y_i \hat{u}_{ij} - \frac{1}{n}\sum_{i=1}^{n} y_i \frac{1}{n}\sum_{i=1}^{n} \hat{u}_{ij}.$$

- The last term equals zero since $\sum_{i=1}^{n} \hat{u}_{ij} = 0$
- Now substitute for y_i and split the sum as follows

$$\sum_{i=1}^{n} y_i \hat{u}_{ij} = \sum_{i=1}^{n} (\hat{y}_i + \hat{u}_i)\hat{u}_{ij} = \sum_{i=1}^{n} \hat{y}_i \hat{u}_{ij} + \sum_{i=1}^{n} \hat{u}_i \hat{u}_{ij}.$$

- Consider the second term

$$\begin{aligned}
\sum_{i=1}^{n} \hat{u}_i \hat{u}_{ij} &= \sum_{i=1}^{n} \hat{u}_i (x_{ij} - \hat{\gamma}_0 - \hat{\gamma}_1 x_{i1} - \ldots - \hat{\gamma}_k x_{ik}) \\
&= \sum_{i=1}^{n} \hat{u}_i x_{ij} - \hat{\gamma}_0 \sum_{i=1}^{n} \hat{u}_i - \hat{\gamma}_1 \sum_{i=1}^{n} \hat{u}_i x_{i1} - \ldots - \hat{\gamma}_k \sum_{i=1}^{n} \hat{u}_i x_{ik} \\
&= 0,
\end{aligned}$$

Proof

- Now consider the first term and substitute for $\hat{y}_i$

$$\begin{aligned}\sum_{i=1}^{n} \hat{y}_i \hat{u}_{ij} &= \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_j x_{ij} + \ldots + \hat{\beta}_k x_{ik}) \hat{u}_{ij} \\ &= \hat{\beta}_0 \sum_{i=1}^{n} \hat{u}_{ij} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} \hat{u}_{ij} + \ldots + \hat{\beta}_j \sum_{i=1}^{n} x_{ij} \hat{u}_{ij} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} \hat{u}_{ij} \\ &= \hat{\beta}_j \sum_{i=1}^{n} x_{ij} \hat{u}_{ij},\end{aligned}$$

Proof

- Now consider the first term and substitute for $\hat{y}_i$

$$\begin{aligned}\sum_{i=1}^{n} \hat{y}_i \hat{u}_{ij} &= \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_j x_{ij} + \ldots + \hat{\beta}_k x_{ik}) \hat{u}_{ij} \\ &= \hat{\beta}_0 \sum_{i=1}^{n} \hat{u}_{ij} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} \hat{u}_{ij} + \ldots + \hat{\beta}_j \sum_{i=1}^{n} x_{ij} \hat{u}_{ij} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} \hat{u}_{ij} \\ &= \hat{\beta}_j \sum_{i=1}^{n} x_{ij} \hat{u}_{ij},\end{aligned}$$

- The other terms are zero because the residuals are uncorrelated with the predictors and the mean of the residuals is zero

Proof

- Now substitute for x_{ij}

$$
\begin{aligned}
\hat{\beta}_j \sum_{i=1}^{n} x_{ij}\hat{u}_{ij} &= \hat{\beta}_j \sum_{i=1}^{n} (\hat{\gamma}_0 + \hat{\gamma}_1 x_{i1} + \ldots + \hat{\gamma}_k x_{ik} + \hat{u}_{ij})\hat{u}_{ij} \\
&= \hat{\beta}_j \left(\hat{\gamma}_0 \sum_{i=1}^{n} \hat{u}_{ij} + \hat{\gamma}_1 \sum_{i=1}^{n} x_{i1}\hat{u}_{ij} + \ldots + \hat{\gamma}_k \sum_{i=1}^{n} x_{ik}\hat{u}_{ij} + \sum_{i=1}^{n} \hat{u}_{ij}^2 \right) \\
&= \hat{\beta}_j \sum_{i=1}^{n} \hat{u}_{ij}^2 \\
&= \hat{\beta}_j n \widehat{Var}(U_j).
\end{aligned}
$$

- We therefore have shown that

$$\widehat{Cov}(Y, U_j) = \hat{\beta}_j \widehat{Var}(U_j),$$

which is equivalent to

$$\hat{\beta}_j = \frac{\widehat{Cov}(Y, U_j)}{\widehat{Var}(U_j)}$$

Statistical properties of the OLS

Unbiasedness

- To prove that the OLS estimator is unbiased, start with the OLS formula

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}.\end{aligned}$$

- Taking the conditional expectation yields

$$\mathbb{E}[\hat{\beta} \mid \mathbf{X}] = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\mathbf{u} \mid \mathbf{X}] = \beta,$$

since $\mathbb{E}[\mathbf{u} \mid \mathbf{X}] = 0$ by our first assumption

Omitted variables bias

- Suppose the true population model is

$$Y = \alpha + \gamma Z + \mathbf{x}'\beta + U,$$

  where we think of Z as a treatment variable of interest, while $\mathbf{x}$ are the control variables
- We will call it the **long model**
- What would happen if we omitted those control variables and estimated a simple regression of Y on Z (the **short model**)?

Omitted variables bias formula

- We can show that the estimate of the slope coefficient from that regression takes the following form:

Omitted variables bias formula

The expectation of the estimate of the slope coefficient from a simple regression of Y on Z is

$$\mathbb{E}\left[\frac{\widehat{Cov}(Y,Z)}{\widehat{Var}(Z)} \mid Z, \mathbf{x}\right] = \gamma + \beta_1\rho_1 + \ldots + \beta_k\rho_k,$$

where ρ_j are the slope coefficients from simple regressions of X_j on Z

Proof

- Recall that from in the simple regression of Y on Z, we can write the estimate of the slope coefficient as

$$\frac{\sum_{i=1}^{n}(z_i - \bar{z})y_i}{\sum_{i=1}^{n}(z_i - \bar{z})^2} = \sum_{i=1}^{n} w_i y_i,$$

where

$$w_i \equiv \frac{z_i - \bar{z}}{\sum_{i=1}^{n}(z_i - \bar{z})^2}, \quad \sum_{i=1}^{n} w_i = 0, \quad \sum_{i=1}^{n} w_i z_i = 1$$

Proof

- Then the conditional expectation of this estimate is

$$\mathbb{E}\left[\sum_{i=1}^{n} w_i y_i \mid Z, \mathbf{x}\right] = \mathbb{E}\left[\sum_{i=1}^{n} w_i(\alpha + \gamma z_i + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i) \mid Z, \mathbf{x}\right]$$

$$= \mathbb{E}[\alpha \sum_{i=1}^{n} w_i + \gamma \sum_{i=1}^{n} w_i z_i + \beta_1 \sum_{i=1}^{n} w_i x_{i1} + \ldots$$

$$+ \beta_k \sum_{i=1}^{n} w_i x_{ik} + \sum_{i=1}^{n} w_i u_i \mid Z, \mathbf{x}]$$

$$= \gamma + \beta_1 \mathbb{E}\left[\sum_{i=1}^{n} w_i x_{i1} \mid Z, \mathbf{x}\right] + \ldots$$

$$+ \beta_k \mathbb{E}\left[\sum_{i=1}^{n} w_i x_{ik} \mid Z, \mathbf{x}\right] + \mathbb{E}\left[\sum_{i=1}^{n} w_i u_i \mid Z, \mathbf{x}\right].$$

Proof

- Consider the last term

$$\mathbb{E}\left[\sum_{i=1}^{n} w_i u_i \mid Z, \mathbf{x}\right] = \sum_{i=1}^{n} w_i \mathbb{E}[u_i \mid Z, \mathbf{x}] = 0.$$

- Now notice that the terms $\sum_{i=1}^{n} w_i x_{ij}$ are the estimates of the slope coefficients from regressions of X_j on Z
- Call the true coefficients from those regressions ρ_j
- The conditional expectations of the estimates will be equal to the true coefficients

OVB explained

- Therefore,

$$\mathbb{E}\left[\sum_{i=1}^{n} w_i y_i \mid Z, \mathbf{x}\right] = \gamma + \beta_1\rho_1 + \ldots + \beta_k\rho_k$$

- The estimated coefficient from the short model will be an unbiased estimated of the true γ if
  - either none of the control variables have any effect on the outcome (all the betas are zero)...
  - ...or neither control variable is correlated with the treatment (all the rhos are zero)
- If these conditions do not hold, the control variables have to be included in the regression.
- In other words, if we suspect that there are variables that affect both the outcome and the treatment, we should control for these variables

Notes on OVB

More general formula

The omitted variables bias formula can be written in a more general way by partitioning the predictors into two arbitrary groups, instead just isolating a single predictor as a treatment variable. We will not consider a more general formula here.

- Notice that by estimating a short model instead of a true long model we are effectively violating our first assumption, hence the estimator is no longer unbiased
- In this case, we would call the short model **misspecified**

Example

- Recall our earlier example about the effect of education on income
- The true population model is

$$income_i = \beta_0 + \beta_1 education_i + \beta_2 ability_i + u_i,$$

so that ability affects (in fact, increases) both income and education
- If instead we estimate the short model

$$income_i = \beta_0 + \beta_1 education_i + v_i,$$

our estimate of the effect of education will be biased
- In fact, we can even say in which direction the bias will go

Example

- Recall our earlier example about the effect of education on income
- The true population model is

$$income_i = \beta_0 + \beta_1 education_i + \beta_2 ability_i + u_i,$$

  so that ability affects (in fact, increases) both income and education
- If instead we estimate the short model

$$income_i = \beta_0 + \beta_1 education_i + v_i,$$

  our estimate of the effect of education will be biased
- In fact, we can even say in which direction the bias will go
- Since ability increases both income and education, the estimated effect of education from the short model will be biased **upwards**

Regression output

	naive	**controls**
(Intercept)	54.947	49.976
	(0.321)	(0.320)
education	16.675	9.918
	(0.370)	(0.382)
ability		20.007
		(0.565)
Num.Obs.	5000	5000
RMSE	11.27	10.07
R2	0.289	0.431

Next Time on *Introductory Econometrics*...

Variance of OLS and model selection