Multiple Regression Model

Juergen Meinecke

Roadmap

Omitted Variables and Causal Effects

Bias from Omitted Variables

An oracle tells you that the relationship between X_{1i}, X_{2i} and Y_i can be represented by

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

According to this model, we should regress Y on X_1 and X_2 to obtain OLS estimates of β_1 and β_2

But what would we be estimating if, instead, we only regressed Y on X_1 (ignoring the presence of X_2)?

It should be clear that you get different estimates for β_1 in the following two estimations

- Y on X_1 and X_2
- Y on X_1

Which one is the “correct” one?

Under the multiple regression model from the previous slide, the first estimation is the correct one

But what does the second equation estimate?

To find out, rewrite the model as follows

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \\ &= \beta_0 + \beta_1 X_{1i} + (\beta_2 X_{2i} + u_i) \\ &= \beta_0 + \beta_1 X_{1i} + w_i, \end{aligned}$$

where $w_i := \beta_2 X_{2i} + u_i$ denotes a new error term

In general, $w_i \neq u_i$ (because $\beta_2 \neq 0$)

The last equation now looks like a simple regression model in which the error term is called w_i

Can use knowledge from simple regression model to study what happens when you regress Y_i on only X_{1i} (omitting X_{2i})

Given $Y_i = \beta_0 + \beta_1 X_{1i} + w_i$, the OLS estimator of β_1 is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_{1i} - \bar{X}_1)}{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2}$$

From the simple regression model, we know

$$Y_i - \bar{Y} = \beta_1(X_{1i} - \bar{X}_1) + (w_i - \bar{w})$$

and plug in to get (after some simplifications)

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(w_i - \bar{w})}{\frac{1}{n}\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2}$$

Typically, the argument now would be that X_{1i} and error term w_i are uncorrelated so that $\frac{1}{n}\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(w_i - \bar{w})$ is almost zero

Problem here:
this particular error term is not uncorrelated with X_{1i}

Recall that w_i is not really a random error term

It also contains X_{2i} because $w_i := \beta_2 X_{2i} + u_i$

It has two components

- u_i which is purely random and uncorrelated with X_{1i} (based on OLS assumption 1)
- X_{2i} which is an omitted regressor which could well be correlated with X_{1i}

If X_{1i} and X_{2i} are correlated with each other than the error term w_i will be correlated with X_{1i}

This will lead to bias in the OLS estimate $\hat{\beta}_1$

Going back to our previous result

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(w_i - \bar{w})}{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2}$$

If we are interested in the expected value of $\hat{\beta}_1$, $\mathrm{E}[\hat{\beta}_1 | X_{1i}, X_{2i}]$, the second term on the rhs will not be equal to zero

Instead, we get …

$$
\begin{aligned}
&\mathsf{E}\left[\hat{\beta}_1 | X_{1i}, X_{2i}\right] \\
&= \beta_1 + \frac{\sum_{i=1}^{n} \mathsf{E}\left[(X_{1i} - \bar{X}_1)(w_i - \bar{w}) | X_{1i}, X_{2i}\right]}{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)^2} \\
&= \beta_1 + \frac{\sum_{i=1}^{n} \mathsf{E}\left[(X_{1i} - \bar{X}_1)(\beta_2 (X_{2i} - \bar{X}_2) + (u_i - \bar{u})) | X_{1i}, X_{2i}\right]}{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)^2} \\
&= \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)^2} + \frac{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1) \mathsf{E}\left[(u_i - \bar{u}) | X_{1i}, X_{2i}\right]}{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)^2} \\
&= \beta_1 + \beta_2 \frac{\frac{1}{n} \sum_{i=1}^{n} (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\frac{1}{n} \sum_{i=1}^{n} (X_{1i} - \bar{X}_1)^2} \\
&\simeq \beta_1 + \beta_2 \frac{\mathsf{E}[(X_{1i} - \mu_{X_1})(X_{2i} - \mu_{X_2})]}{\mathsf{Var}\ (X_{1i})} \\
&= \beta_1 + \beta_2 \frac{\mathsf{Cov}(X_{1i}, X_{2i})}{\mathsf{Var}\ (X_{1i})}
\end{aligned}
$$

The second equality holds because
$w_i := \beta_2 X_{2i} + u_i$ and $\bar{w} = \beta_2 \bar{X}_2 + \bar{u}$

The third equality holds because X_{1i} and X_{2i} can be treated as constants

The fourth equality holds because of by OLS Assumption 1

To get the asymptotic result just replace sample averages by population averages

This shows that the expected value of $\hat{\beta}_1$ is not equal to β_1

The OLS estimator $\hat{\beta}_1$ is therefore not unbiased

What is the bias equal to?

This bias term is $\beta_2 \cdot \text{Cov}(X_{1i}, X_{2i})/\text{Var}\ (X_{1i})$

This bias is proportional to the covariance between X_{1i} and X_{2i} and inversely proportional to the variance of X_{1i}

The omitted variables bias could be positive or negative: its sign is determined by the interplay of the signs of β_2 and $\text{Cov}(X_{1i}, X_{2i})$

If you do not like the mathematics of it, maybe you prefer to understand it intuitively

If you omit X_{2i} from the estimation, then the estimate of β_1 will be biased

The reason for this is that the estimator $\hat{\beta}_1$ is doing two jobs at the same time:

- it captures the direct effect of X_{1i} on Y
  (this is what you *want* to capture; it's the effect β_1)
- but it also captures the indirect effect that X_{2i} has through its covariance with X_{1i}
  (this creates the bias)

Multiple Regression Model

Juergen Meinecke

Roadmap

Omitted Variables and Causal Effects

Example: Mozart Effect

Do kids who learn to play an instrument become more successful later in life than kids who do not learn to play an instrument?

To answer that question, you have available a data set on, say, 10,000 adults and you observe their salaries and whether they learned an instrument when they were kids

You may want to consider the following model:
(suppressing i-subscripts for convenience)

$$\texttt{Salary} = \beta_0 + \beta_1 \texttt{Instrument} + w,$$

where `Instrument` is a dummy variable which is equal to 1 if a person learned an instrument as a kid

What’s wrong with estimating this model?

The problem is that you are omitting lots of other things in this model

Ideally, you would like to include the following explanatory variables in addition to `Instrument`

- person's aptitude and intelligence
- person's education
- parent's socio-economic status

All of these are likely correlated with the regressor `Instrument`, therefore your estimate of β_1 is biased

You cannot give your estimate $\hat{\beta}_1$ of β_1 a *causal interpretation*

Instead, the estimate $\hat{\beta}_1$ is a convolution of the actual effect β_1 that you are after and the indirect effects of aptitude, intelligence, education and parent's socio-economic status that enter through the covariance with `Instrument`

If, for example, you find this model more realistic:

$$\texttt{Salary} = \beta_0 + \beta_1 \texttt{Instrument} + \beta_2 \texttt{Educ} + u,$$

where `Educ` captures the child's education attainment

Remember the bias term is $\beta_2 \cdot \mathrm{Cov}(X_{1i}, X_{2i}) / \mathrm{Var}\,(X_{1i})$

Here, quite plausibly: $\beta_2 > 0$ and $\mathrm{Cov}(X_{1i}, X_{2i}) > 0$ (why?)

Consequently, the simple regression (in which we omit `Educ`) **overstates** the effect of `Instrument` on `Salary`

When `Educ` is omitted, the regressor `Instrument` captures two effects:

- the direct effect of `Instrument` on `Salary`
- the indirect effect of `Educ` on `Salary` through the correlation of `Educ` with `Instrument`

An estimation like the one just described confuses *causation* with *correlation*

Ideally, we want to know the *causal effect* of `Instrument` on `Salary`

But this an unattainable goal for us unless we are able to include a comprehensive list of explanatory variables

How could we get a true *causal effect*?

The only way of getting this would be by conducting a *randomized controlled experiment*, for example:

Of 10,000 randomly chosen babies we randomly select 5,000 and make sure they will learn an instrument as kids (treatment group) and for the other 5,000 we make sure that they will not learn an instrument (control group). 40 years later we collect data on their salaries and then we run the regression of salaries on the instrument dummy variable.

Seems like a bit of a stretch, doesn't it?
(not sure if the ANU Ethics in Research committee would approve this research proposal)

Although the example is preposterous, it neatly shows what we mean by causal effect

By comparing the salaries between the treatment and control groups you immediately learn about the effect of playing an instrument as a kid

Because instrument-playing was randomly assigned, there is no cross-contamination from other regressors

In this example, we do not need to control for aptitude, intelligence, education or parent’s socio-economic status because we randomize everything

Aptitude, intelligence, education or parent’s socio-economic status will be uncorrelated with instrument-playing *by design*

This then leads to an intuitive definition of causal effect

Definition

A **causal effect** is the effect measured in an ideal randomized controlled experiment.

This definition of causal effect is very soft and imprecise

But it offers you a lot of intuition already

In practice, we rarely have ideal randomized controlled experiments in economics

Whenever we do not have a randomized controlled experiments, we have to think hard about the issue of causation versus correlation

We always have to ask: Is my estimate measuring something causal? Or is it only a correlation?

Bottom line: If you are absolutely confident that you included an exhaustive list of regressors then you can (almost) be confident that your estimate measures a causal effect

Multiple Regression Model

Juergen Meinecke

Roadmap

Omitted Variables and Causal Effects

- Example: Return to Education

Labor economists are interested in the so-called return to education: *By how much does a university degree increase your expected earnings?*

Ideally, at age 18 you have a choice: go get a job and earn money or go to university and increase your human capital

A university is like a bank: you bring your human capital and receive a positive annual rate of return on your investment

Estimating the return to education seems so easy: just compare earnings of university graduates to earnings of high-school graduates

What is wrong with this idea?

To estimate that rate of return, labor economists have considered the following multiple linear regression model:

$$\text{Earnings}_i = \beta_0 + \beta_1 \text{Education}_i + \beta_2 \text{Experience}_i + u_i$$

where Experience_i denotes a person's work experience

To estimate the coefficients, all you need is a standard household survey with observations on people's earnings, education and work experience

In Australia, good candidates are the census data and the Household, Income and Labour Dynamics in Australia (HILDA) panel data set (that started in 2001 and is ongoing on a yearly basis)

Here the results when you use the 2009 wave of HILDA

Python Code (output edited)

```
> formula = 'loghrlwage ~ educ + exper'
> reg1 = smf.ols(formula, data=df, missing='drop').fit(cov_type='HC1', use_t=False)
> print(reg1.summary())

OLS Regression Results
==============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept       2.3374      0.097     24.081      0.000       2.147       2.528
educ            0.0744      0.007     10.965      0.000       0.061       0.088
exper           0.0029      0.001      2.323      0.020       0.000       0.005
==============================================================================
```

The annual return to education is estimated to equal 7.4%

Is this a large number?

When you bring $100 to the bank, they will pay you a measly rate of return of 1.10% p.a. (Westpac eSaver standard variable rate; in fairness: they do offer a 5-months teaser top up rate of 3.15%)

When you invest $100 in the stock market, you made an annual rate of return of 6.2% (past decade, pre-coronavirus, quick look up on the internet)

The return to education therefore seems quite large

Congratulations:
you have made an excellent investment decision going to uni!

But: what are we omitting here?

Reconsider the estimation equation

$$\text{Earnings}_i = \beta_0 + \beta_1 \text{Education}_i + \beta_2 \text{Experience}_i + u_i$$

What other variables should be included on the rhs?

Does the omission lead to upward or downward bias in the estimate of 7.4%? In other words, is the actual return lower or higher than this?

An important research paper by Leigh and Ryan (published in the academic journal *Economics of Education Review* in 2008) has conducted a slightly more sophisticated econometric analysis of the returns to education in Australia

Here is what they found:

- Using a multiple regression model, the return to education is estimated to equal 13%
- This estimate is large and statistically significant
- But it suffers from ovb
- Using sophisticated econometric methods, they mitigate ovb and estimate a return to education of around 10%

But unobserved factors may still remain and create ovb

In a fantasy world, how could we estimate the actual causal effect of education on earnings?

We would conduct the following RCT:

- Randomly choose 10,000 high-school students who just finished high-school
- Randomly select 5,000 and tell them that they cannot go to university but instead have to start working in a job
- The other 5,000 have to go to university
- We wait 20 years and then compare the average earnings between the two groups

This would enable us to measure the actual return to education

Multiple Regression Model

Juergen Meinecke

Roadmap

- Selected Topics
    - Frisch-Waugh Theorem

Theorem (Frisch-Waugh Theorem)

Consider the model $Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_i$. Now pursue the following three step regression:

1. *Regress X_{1i} on $X_{2i}, X_{3i}, \ldots, X_{ki}$*
   and let $\widetilde{X}_{1i}$ denote the residuals from this regression
2. *Regress Y_i on $X_{2i}, X_{3i}, \ldots, X_{ki}$*
   and let $\widetilde{Y}_i$ denote the residuals from this regression
3. *Regress $\widetilde{Y}_i$ on $\widetilde{X}_{1i}$.*

Then the OLS estimate of the coefficient of $\widetilde{X}_{1i}$ in step 3 is equal to the OLS estimate of the coefficient of X_{1i} in the regression of Y_i on $X_{1i}, \ldots, X_{ki}$.

The Frisch-Waugh Theorem provides some more intuition about how to interpret coefficient estimates in multiple regressions

The first two steps generate versions of X_1 and Y that are purged (free) from any correlation with the covariates $X_2, X_3, \dots, X_k$

$\widetilde{X}_1$ captures all the variation in X_1 that is uncorrelated with $X_2, \dots, X_k$

You can view $\widetilde{X}_1$ as a version of X_1 from which the effects of $X_2, \dots, X_k$ have been **partialled** or **netted out**

$\widetilde{Y}$ captures all the variation in Y that is uncorrelated with $X_2, \dots, X_k$

You can view $\widetilde{Y}$ as a version of Y from which the effects of $X_2, \dots, X_k$ have been **partialled** or **netted out**

When regressing $\widetilde{Y}$ against $\widetilde{X}_1$ in step 3, you therefore obtain an estimate that captures the true effect of X_1 on Y in which any correlation through the other regressors $X_2, \dots, X_k$ is shut off

This results in the unencumbered effect of X_1 on Y

And that is exactly what you want to estimate when you run the multiple regression of Y on $X_1, \dots, X_k$

Multiple Regression Model

Juergen Meinecke

Roadmap

Selected Topics

Measures of Fit

There are two regression statistics that provide measures of how well the regression line "fits" the data:

- regression R^2, and
- standard error of the regression (SER)

Main idea: how closely does the scatterplot "fit" around the regression line?

For the multiple regression model, there exists a useful modification for the regression R^2, it's called the *adjusted* R^2

The regression R^2 is the fraction of the sample variation of Y_i that is explained by the explanatory variables $X_{1i}, \dots, X_{ki}$

Total variation in the dependent variable can be broken down as

- total sum of squares (TSS)

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

- explained sum of squares (ESS)

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

- residual sum of squares (RSS)

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

It follows that $TSS = ESS + RSS$

Definition

R^2 is defined by

$$R^2 := \frac{ESS}{TSS}.$$

Corollary

Based on the preceding terminology, it is easy to see that

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

Therefore,

- $R^2 = 0$ means $ESS = 0$ (the regressors $X_{1i}, \dots, X_{ki}$ explain nothing in the variation of the dependent variable Y)
- $R^2 = 1$ means $ESS = TSS$
  (the regressors $X_{1i}, \dots, X_{ki}$ explain all the variation of the dependent variable Y)
- $0 \leqslant R^2 \leqslant 1$
- Important mechanical fact:
  when you add another explanatory variable in the regression,
  then the RSS decreases
  as result, R^2 increases
  bottom line:
  In multiple regression, R^2 increases when you add another regressor

It seems undesirable that R^2 increases when a new explanatory variable is added

If you are unhappy with the fit of your model, you could just throw in a ton of variables and R^2 will mechanically increase

To circumvent that, there is a remedy

Definition

The **adjusted** R^2 is defined by

$$\bar{R}^2 := 1 - \frac{n-1}{n-k-1}\frac{RSS}{TSS}.$$

The adjusted R^2 does not necessarily increase when you add regressors

Three interesting points about $\bar{R}^2$

- $\bar{R}^2 < R^2$
- Adding a regressor has two opposing effects:
  RSS decreases but $(n-1)/(n-k-1)$ increases
  net effect depends on particular application
  $\bar{R}^2$ could go up or down
- $\bar{R}^2$ can be negative

The SER measures the spread of the distribution of u

The SER is (almost) the sample standard deviation of the OLS residuals:

$$SER = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2} = \sqrt{\frac{RSS}{n-k-1}}$$

The SER

- has the units of u, which are the units of Y
- measures the spread of the OLS residuals around the estimated PRF