ECO375 Tutorial 3
Partialling Out / Omitted Variable Bias

Matt Tudball

University of Toronto Mississauga

September 28, 2017

Review: Multiple Regression (MR)

- Let's begin by refreshing our memory of the multiple regression model.
- Suppose we have n observations of the following variables:
  - A dependent variable y_i
  - k independent variables $x_{i1}, x_{i2}, ..., x_{ik}$.
- We suppose that they are related according to the model:
  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i$
- Our objective in multiple regression is to estimate the $k + 1$ parameters $\beta_0, \beta_1, ..., \beta_k$.

Review: MR Assumptions

MLR.1 The model is linear in parameters:
$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i$

MLR.2 We have a random sample $\{(x_{i1}, x_{i2}, ..., x_{ik}, y_i)\}_{i=1}^{n}$ of size n following the model in MLR.1.

MLR.3 None of the independent variables is constant and there is no perfect multicollinearity (i.e. no exact linear relationship among the independent variables.

MLR.4 The error u_i has an expected value of zero conditional on all x_i
$\mathrm{E}(u_i|x_{i1}, x_{i2}, ..., x_{ik}) = 0$ for $i = 1, ..., n$.

MLR.5 The error u_i homoscedastic, i.e. it has the same variance for all x_i
$\mathrm{Var}(u_i|x_{i1}, x_{i2}, ..., x_{ik}) = \sigma^2$

MLR.6 The error u_i is independent of the explanatory variables $x_{i1}, x_{i2}, ..., x_{ik}$ and is normally distributed with mean 0 and variance σ^2
$u_i \sim \mathcal{N}(0, \sigma^2)$

Review: MR First Order Conditions

- As before we want to minimise the sum of square residuals:

$$\min_{\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right)^2$$

- We can now take derivatives of this objective function with respect to $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ and set them equal to 0:

$$0 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right)$$

$$0 = \sum_{i=1}^{n} x_{i1} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right)$$

$$0 = \sum_{i=1}^{n} x_{i2} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right)$$

$$\vdots$$

$$0 = \sum_{i=1}^{n} x_{ik} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right)$$

- We have $k+1$ equations in $k+1$ unknowns, so we can solve for $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ (although this isn't easy to do).

Review: MR Residuals and Predicted Values

- Let's now define the predicted values and residuals:
  - Predicted value: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_k x_{ik}$
    You can think of this as all of the variation in y_i which can be explained by $x_{i1}, x_{i2}, ..., x_{ik}$.
  - Sample residual: $\hat{u}_i = y_i - \hat{y}_i$
    You can think of this as all of the variation in y_i which cannot be explained by $x_{i1}, x_{i2}, ..., x_{ik}$.
- We can rewrite the first order conditions in the previous slide with this more compact notation:
  $0 = \sum_{i=1}^{n} \hat{u}_i$
  $0 = \sum_{i=1}^{n} x_{ij} \hat{u}_i$ for all $j = 1, 2, ..., k$.
- Note that this implies that the sum of sample residuals is always 0 and the sample covariance between x_{ij} and $\hat{u}_i$ is also always 0.

"Partialling Out" Procedure: Set-Up

- What if we are only interested in the estimate for β_1? As we mentioned above it isn't easy to derive an exact expression for $\hat{\beta}_1$ by directly minimising the sum of square residuals.
- We know, however, that it would be easy to find an expression for $\hat{\beta}_1$ if it were coming from a simple regression.
- The idea behind the "partialling out" procedure is to remove the effects of $x_{i2}, x_{i3}, ..., x_{ik}$ and then run a simple regression which returns $\hat{\beta}_1$.

"Partialling Out" Procedure: Steps

- Consider the multiple regression model
  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i$.
- **Goal**: Express the OLS estimator $\hat{\beta}_1$ in the multiple regression model above as an OLS estimator in a simple regression model.
- Steps:
  1. Regress x_1 on x_2, x_3, ..., x_k and calculate the residual $\hat{r}_1$.
     $x_{i1} = \hat{\alpha}_0 + \hat{\alpha}_2 x_{i2} + ... + \hat{\alpha}_k x_{ik} + \hat{r}_{i1}$
     $\hat{r}_1 = x_{i1} - \hat{x}_{i1}$
  2. Regress y on $\hat{r}_1$.
     $y_i = \hat{\lambda}_0 + \hat{\lambda}_1 \hat{r}_{i1} + \hat{\epsilon}_i$
  3. The resulting slope estimate $\hat{\lambda}_1$ is always equal to $\hat{\beta}_1$.

$$\hat{\beta}_1 = \hat{\lambda}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{r}_{i1} - 0)}{\sum_{i=1}^{n}(\hat{r}_{i1} - 0)^2} = \frac{\sum_{i=1}^{n} y_i \hat{r}_{i1}}{\sum_{i=1}^{n} \hat{r}_{i1}^2} \tag{1}$$

"Partialling Out" Procedure: Intuition

- Recall that $\hat{r}_1$ can be interpreted as the variation in x_1 that cannot be explained by $x_2, x_3, ..., x_k$.
- In other words, $\hat{r}_1$ is the part of x_1 that is uncorrelated with $x_2, x_3, ..., x_k$.
- Thus, $\hat{r}_1$ is what's left of x_1 after the effects of all the other explanatory variables have been "partialled out".
- So $\hat{\beta}_1$ is obtained by a regression of y on $\hat{r}_1$, the variation that is unique to x_1.

Running A Regression: `regress`

- Let's take a moment to learn how to run a regression in Stata. You can do it with the command `regress`.
- Suppose we want to regress log hourly wages on education, experience and tenure using the WAGE1.dta dataset. We can type the following command:

```
regress lwage educ exper tenure
```

Source	SS	df	MS	Number of obs	=	526
				F(3, 522)	=	80.39
Model	46.8741776	3	15.6247259	Prob > F	=	0.0000
Residual	101.455574	522	.194359337	R-squared	=	0.3160
				Adj R-squared	=	0.3121
Total	148.329751	525	.28253286	Root MSE	=	.44086

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.092029	.0073299	12.56	0.000	.0776292	.1064288
exper	.0041211	.0017233	2.39	0.017	.0007357	.0075065
tenure	.0220672	.0030936	7.13	0.000	.0159897	.0281448
_cons	.2843595	.1041904	2.73	0.007	.0796756	.4890435

Post-Estimation Prediction: `predict`

- After you have run a regression, you are able to make predictions from it using the command `predict`.
- Suppose we want to calculate the predicted values of log wage from our previous regression, then we would type:
  `predict lwage_hat, xb`.
  This will save a new variable called lwage_hat containing the predicted values.
- Suppose we want to calculate the sample residuals from the previous regression, then we would type:
  `predict res, residuals`.
  This will save a new variable called res containing the sample residuals.
- **Note:** This is a post-estimation command so it will only work if you use it after running a regression. If you use it beforehand, Stata will give you the error `last estimates not found`.

In-Class Exercise 1

- If you haven't already, load the WAGE1.dta dataset into Stata. You can download it from my website (matthewtudball.com) under the Tutorial 2 heading.
  1. Run the multiple regression of log wage on education, experience and tenure. Find and record the coefficient for education $\hat{\beta}_1$.
  2. Run the multiple regression of education on experience and tenure. Save the sample residuals from this regression as a new variable called *res*.
  3. Run the simple regression of log wage on *res*. Find and record the slope coefficient $\hat{\lambda}$.
  4. Compare $\hat{\beta}_1$ and $\hat{\lambda}$. Are they equal? Is that to be expected?

Omitted Variable Bias: Set-Up

- Omitting a relevant variable from a regression model can bias the estimated coefficients on the included variables.
- Let's assume that the actual population relationship between y and a set of x's is
  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i$
  satisfying the assumptions MLR.1 to MLR.4
- But suppose that we accidentally specify the the model
  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{k-1} x_{ik-1} + u_i$
  whereby we omit a relevant variable x_k.
- Denote the estimator for β_1 obtained in this misspecified model by $\tilde{\beta}_1$.

Omitted Variable Bias: Deriving The Bias

- By the "partialling out" procedure we know that $\tilde{\beta}_1$ can be written as

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} \tilde{r}_{i1} y_i}{\sum_{i=1}^{n} \tilde{r}_{i1}^2} \tag{2}$$

where $\tilde{r}_{i1}$ are the sample residuals from the regression $x_{i1} = \alpha_0 + \alpha_2 x_{i2} + \alpha_3 x_{i3} + ... + \alpha_{k-1} x_{ik-1} + r_i$

- To derive the bias in the estimator $\tilde{\beta}_1$ we can begin by plugging the *true* y_i into (2).

$$\begin{aligned}
\tilde{\beta}_1 &= \frac{\sum_{i=1}^{n} \tilde{r}_{i1} y_i}{\sum_{i=1}^{n} \tilde{r}_{i1}^2} \\
&= \frac{1}{\sum_{i=1}^{n} \tilde{r}_{i1}^2} \sum_{i=1}^{n} \tilde{r}_{i1} (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_k x_{ik} + \hat{u}_i) \\
&= \frac{1}{\sum_{i=1}^{n} \tilde{r}_{i1}^2} (\hat{\beta}_0 \sum_{i=1}^{n} \tilde{r}_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} \tilde{r}_{i1} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} \tilde{r}_{i1} x_{i2} \\
&\qquad + ... + \hat{\beta}_k \sum_{i=1}^{n} \tilde{r}_{i1} x_{ik} + \sum_{i=1}^{n} \tilde{r}_{i1} \hat{u}_i)
\end{aligned} \tag{3}$$

Omitted Variable Bias: Deriving The Bias

- Recall that from the first order conditions of the regression of x_1 on $x_2, x_3, ..., x_{k-1}$:
  $0 = \sum_{i=1}^{n} \tilde{r}_{i1}$
  $0 = \sum_{i=1}^{n} x_{ij}\tilde{r}_{i1}$ for all $j = 2, ..., k-1$.
- Also note that $0 = \sum_{i=1}^{n} \tilde{r}_{i1}\hat{u}_i$. This follows from the first order conditions of the regression of y on x_1, x_2, ..., x_k which require that $\sum_{i=1}^{n} x_{1i}\hat{u}_i = 0$. Note that:

$$\begin{aligned}
\sum_{i=1}^{n} x_{1i}\hat{u}_i &= \sum_{i=1}^{n}(\tilde{x}_{1i} + \tilde{r}_{1i})\hat{u}_i \\
&= \sum_{i=1}^{n}(\tilde{\alpha}_0 + \tilde{\alpha}_2 x_{i2} + ... + \tilde{\alpha}_{k-1}x_{ik-1} + \tilde{r}_{1i})\hat{u}_i \\
&= \tilde{\alpha}_0 \sum_{i=1}^{n} \hat{u}_i + \tilde{\alpha}_2 \sum_{i=1}^{n} x_{i2}\hat{u}_i + ... + \tilde{\alpha}_{k-1} \sum_{i=1}^{n} x_{ik-1}\hat{u}_i + \\
&\qquad \sum_{i=1}^{n} \tilde{r}_{1i}\hat{u}_i \\
&= \sum_{i=1}^{n} \tilde{r}_{1i}\hat{u}_i = 0
\end{aligned}$$

- Thus we have shown that $\sum_{i=1}^{n} \tilde{r}_{1i}\hat{u}_i = 0$.

Omitted Variable Bias: Deriving The Bias

- Using what we learned in the previous slide we can simplify the expression for $\tilde{\beta}_1$ in equation (3):

$$\begin{aligned}\tilde{\beta}_1 &= \frac{1}{\sum_{i=1}^{n} \tilde{r}_{i1}^2}(\hat{\beta}_0 \sum_{i=1}^{n} \tilde{r}_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} \tilde{r}_{i1} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} \tilde{r}_{i1} x_{i2} \\ &\quad + \ldots + \hat{\beta}_k \sum_{i=1}^{n} \tilde{r}_{i1} x_{ik} + \sum_{i=1}^{n} \tilde{r}_{i1} \hat{u}_i) \\ &= \frac{1}{\sum_{i=1}^{n} \tilde{r}_{i1}^2}(\hat{\beta}_1 \sum_{i=1}^{n} \tilde{r}_{i1} x_{i1} + \hat{\beta}_k \sum_{i=1}^{n} \tilde{r}_{i1} x_{ik})\end{aligned}$$

- By a very similar approach to the derivation in the previous slide we can show that $\sum_{i=1}^{n} \tilde{r}_{i1} x_{i1} = \sum_{i=1}^{n} \tilde{r}_{i1}^2$
- Thus,

$$\begin{aligned}\tilde{\beta}_1 &= \hat{\beta}_1 + \hat{\beta}_k \frac{\sum_{i=1}^{n} \tilde{r}_{i1} x_{ik}}{\sum_{i=1}^{n} \tilde{r}_{i1}^2} \\ &= \hat{\beta}_1 + \hat{\beta}_k \tilde{\delta}_1\end{aligned}$$

where $\tilde{\delta}_1$ is the "partialling out" coefficient for x_1 from the regression of x_k on x_1, x_2, ..., x_{k-1}.

Omitted Variable Bias: Intuition

- The bias in $\tilde{\beta}_1$ is
  $\text{Bias}(\tilde{\beta}_1) = \text{E}(\tilde{\beta}_1 - \beta_1) = \text{E}(\hat{\beta}_1 + \hat{\beta}_k\tilde{\delta}_1 - \beta_1) = \beta_1 + \beta_k\tilde{\delta} - \beta_1 = \beta_k\tilde{\delta}.$
- We can think of $\tilde{\delta}_1$ as the effect of x_1 on x_k and β_k as the effect of x_k on y.
- So the bias in $\tilde{\beta}_1$ is coming from the indirect effect of x_1 on y through x_k.
- It also follows that there are two cases in which $\tilde{\beta}_1$ is unbiased:
  1. $\tilde{\delta}_1 = 0$: x_1 and x_k are uncorrelated.
  2. $\beta_k = 0$: x_k and y are uncorrelated.

	$\text{Cov}(x_1, x_k) > 0$	$\text{Cov}(x_1, x_k) < 0$
$\beta_k > 0$	Positive bias	Negative bias
$\beta_k < 0$	Negative bias	Positive bias

In-Class Exercise 2

- Load the WAGE2.dta dataset into Stata. You can download it from my website (matthewtudball.com) under the Tutorial 3 heading.
- Suppose that the true relationship is
  $ln(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 age_i + \beta_4 IQ_i + u_i$
  but we misspecify our model to exclude IQ (which could be thought of as natural ability).
  1. Run the misspecified regression of log wage on education, experience and age. Record the coefficient on education $\tilde{\beta}_1$. Will $\tilde{\beta}_1$ be biased? If so, in what direction?
  2. Run the regression of education on experience and age. Save the residuals as a new variable called *res*.
  3. Run the regression of IQ on *res*. Record the coefficient $\tilde{\delta}_1$.
  4. Now run the correctly specified regression of log wage on education, experience, age and IQ. Record the coefficient on IQ $\hat{\beta}_4$ and education $\hat{\beta}_1$.
  5. What does the product $\tilde{\delta}_1 \ \hat{\beta}_4$ estimate? Is it consistent with your initial prediction? Verify that $\tilde{\beta}_1 = \hat{\beta}_1 + \tilde{\delta}_1 \ \hat{\beta}_4$.