Week 4: Endogeneity, Training Exercises

Coursera/Erasmus U., Econometric Methods and Applications

Anthony Nguyen

Training Exercise 4.2

Notes:

- This exercise uses the datafile `TrainExer42`and requires a computer.
- The dataset `TrainExer42` is available on the website.

Questions

In this exercise we reconsider the example from lecture 4.1 where an analyst models sales of ice cream over time as a function of price and where price is possibly endogenous due to strategic behavior of the salesperson. In this case the salesperson knows that when a particular event is organized, demand tends to be high. Therefore she may set a high price when there is such an event.

We consider the following data generating process

$$\text{Sales} = 100 - 1 \cdot \text{Price} + \alpha\text{Event} + \varepsilon_1\text{Price} = 5 + \beta\text{Event} + \varepsilon_2$$

where Event is a 0/1 dummy variable indicating whether an event took place at a point in time. However, when trying to estimate the price coefficient the analyst does not have the Event dummy variable and simply regresses Sales on a constant and Price.

The dataset `TrainExer42` contains sales and price data for different values of α and β. For each scenario the same simulated values for ε_1 and ε_2 were used. Specifically, the data contains 4 price series and 16 sales series. Price variables "PriceB" give the price assuming that $\beta = B$ for $B = 0, 1, 5, 10$. Sales variables "SalesA_B" give the sales for $\alpha = A$ and $\beta = B$, where A also takes the values $0, 1, 5, 10$.

(a) First consider the case where the event only directly affects price ($\alpha = 0$). Estimate and report the price coefficients under all 4 scenarios for β and calculate the R^2 for all these regressions. Do the estimated price coefficients signal any endogeneity problem for these values of α and β? Can you also explain the pattern you find for the R^2.

(b) Repeat the exercise above, but now consider the case where the event only directly affects sales, that is, set $\beta = 0$ and check the results for the four different values of α.

(c) Finally consider the parameter estimates for the cases where the event affects price *and* sales, that is, look at $\alpha = \beta = 0, 1, 5, 10$. Can you see the impact of endogeneity in this case?

Answers

(a) First consider the case where the event only directly affects price ($\alpha = 0$). Estimate and report the price coefficients under all 4 scenarios for β and calculate the R^2 for all these regressions. Do the estimated price coefficients signal any endogeneity problem for these values of α and β? Can you also explain the pattern you find for the R^2.

```r
mod1a <- lm(SALES0_0~PRICE0, data = TrainExer42)
mod2a <- lm(SALES0_1~PRICE1, data = TrainExer42)
mod3a <- lm(SALES0_5~PRICE5, data = TrainExer42)
mod4a <- lm(SALES0_10~PRICE10, data = TrainExer42)

stargazer(mod1a, mod2a, mod3a, mod4a, header = FALSE, keep.stat = "rsq")
```

Table 1:

	Dependent variable:			
	SALES0_0	SALES0_1	SALES0_5	SALES0_10
	(1)	(2)	(3)	(4)
PRICE0	-0.976^{***}			
	(0.032)			
PRICE1		-0.966^{***}		
		(0.030)		
PRICE5			-0.973^{***}	
			(0.017)	
PRICE10				-0.985^{***}
				(0.010)
Constant	99.862^{***}	99.808^{***}	99.833^{***}	99.890^{***}
	(0.161)	(0.156)	(0.100)	(0.068)
R^2	0.794	0.808	0.930	0.977
Note:				*p<0.1; **p<0.05; ***p<0.01

When α is set to zero, the regression coefficients are all close to the true value of -1, so price is not endogenous, as the event does not influence sales directly.

The R^2 increases for higher values of β. This is due to the fact that for higher β values, more of the variation in sales can be explained. In other words, for higher β values, the variation in sales increases, and this increase is perfectly explained by the price.

(b) Repeat the exercise above, but now consider the case where the event only directly affects sales, that is, set $\beta = 0$ and check the results for the four different values of α.

```r
mod1b <- lm(SALES0_0~PRICE0, data = TrainExer42)
mod2b <- lm(SALES1_0~PRICE0, data = TrainExer42)
mod3b <- lm(SALES5_0~PRICE0, data = TrainExer42)
mod4b <- lm(SALES10_0~PRICE0, data = TrainExer42)

stargazer(mod1b, mod2b, mod3b, mod4b, header = FALSE, keep.stat = "rsq")
```

Table 2:

	Dependent variable:			
	SALES0_0	SALES1_0	SALES5_0	SALES10_0
	(1)	(2)	(3)	(4)
PRICE0	-0.976^{***}	-0.969^{***}	-0.942^{***}	-0.909^{***}
	(0.032)	(0.039)	(0.106)	(0.201)
Constant	99.862^{***}	99.948^{***}	100.294^{***}	100.727^{***}
	(0.161)	(0.197)	(0.539)	(1.027)
R^2	0.794	0.718	0.243	0.076
Note:				*p<0.1; **p<0.05; ***p<0.01

Again, here we can see that all of the regression coefficients are close to the true value of -1, thus we can say that Price is not endogenous, as the Event only affects Sales and not Price.

Therefore, the omission of the EVENT variable, does not lead to a correlation between the error term and Price.

From the regression results, we can also see that the R^2 term drops significantly for higher values of α. At a high value of α, a lot of variation in Sales is due to the Event, however, this variation is not captured in the regression because we only regressed Sales on a Constant and Price. This is also the reason that the estimate for α as 10 is relatively small (0.08), as it reflects the relatively large estimation uncertainty.

(c) Finally consider the parameter estimates for the cases where the event affects price *and* sales, that is, look at $\alpha = \beta = 0, 1, 5, 10$. Can you see the impact of endogeneity in this case?

```r
mod1c <- lm(SALES0_0~PRICE0, data = TrainExer42)
mod2c <- lm(SALES1_1~PRICE1, data = TrainExer42)
mod3c <- lm(SALES5_5~PRICE5, data = TrainExer42)
mod4c <- lm(SALES10_10~PRICE10, data = TrainExer42)

stargazer(mod1c, mod2c, mod3c, mod4c, header = FALSE, keep.stat = "rsq")
```

Here, on the diagonal of the regression summary, we *can* see consequences of endogeneity. If α and β are both non-zero, the omission of the EVENT dummy will lead to correlation between the error term in the regression in PRICE.
As a consequence of this correlation, the estimate can be completely off. For instance, we can see in the case where α and β is equal to 10 the estimate is almost zero (-0.09).

Table 3:

	Dependent variable:			
	SALES0_0	SALES1_1	SALES5_5	SALES10_10
	(1)	(2)	(3)	(4)
PRICE0	-0.976^{***}			
	(0.032)			
PRICE1		-0.874^{***}		
		(0.036)		
PRICE5			-0.273^{***}	
			(0.033)	
PRICE10				-0.085^{***}
				(0.021)
Constant	99.862^{***}	99.458^{***}	96.515^{***}	95.515^{***}
	(0.161)	(0.187)	(0.197)	(0.146)
R^2	0.794	0.706	0.214	0.064
Note:			*p<0.1; **p<0.05; ***p<0.01	