Week 4: Endogeneity, Training Exercises

Coursera/Erasmus U., Econometric Methods and Applications

Anthony Nguyen

Training Exercise 4.4

Notes:

- This exercise uses the datafile `TrainExer44`and requires a computer.
- The dataset `TrainExer44` is available on the website.

Questions

In this exercise we study the gasoline market and look at the relation between consumption and price in the USA. We will use yearly data on these variables from 1977 to 1999. Additionally, we have data on disposable income, and some price indices. More precisely we have

- GC: log real gasoline consumption;
- PG: log real gasoline price index;
- RI: log real disposable income;
- RPT: log real price index of public transport;
- RPN: log real price index of new cars;
- RPU: log real price index of used cars.

We consider the following model:

$$GC = \beta_1 + \beta_2 PG + \beta_3 RI + \varepsilon$$

(a) Give an argument why the gasoline price may be endogenous in this equation.

(b) Use 2SLS to estimate the price elasticity (β_2). Use a constant, RI, RPT, RPN, and RPU as instruments.

(c) Perform a Sargan test to test whether the five instruments are correlated with ε. What do you conclude?

Answers

(a) Give an argument why the gasoline price may be endogenous in this equation.

For example, given that the demand for gasoline in the U.S. influences the global gasoline market, it is therefore likely that a high demand for gasoline in the U.S. leads to an increase in the global market price for gasoline. In other words, consumption (GC) and price (PG) are determined simultaneously, and hence, we suspect that gasoline price may be endogenous in this equation.

(b) Use 2SLS to estimate the price elaticity (β_2). Use a constant, RI, RPT, RPN, and RPU as instruments.

Perform first stage regression of PG by instrumental variables:

```
mod1 <- lm(PG~RI+RPT+RPN+RPU, data=TrainExer44)

mod1_pred <- predict(mod1)

summary(mod1)
```

```
##
## Call:
## lm(formula = PG ~ RI + RPT + RPN + RPU, data = TrainExer44)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.160161 -0.036024 -0.002891  0.034824  0.175059
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.7410     0.8337   9.285 1.40e-09 ***
## RI           -2.2984     0.2471  -9.303 1.35e-09 ***
## RPT          -0.8080     0.1912  -4.225 0.000277 ***
## RPN          -3.5279     0.3520 -10.023 3.06e-10 ***
## RPU           0.2331     0.1831   1.273 0.214765
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07269 on 25 degrees of freedom
## Multiple R-squared:  0.8868, Adjusted R-squared:  0.8687
## F-statistic: 48.97 on 4 and 25 DF,  p-value: 1.797e-11
```

Next, perform the second stage of the OLS by regressing GC by RI and the predicted values from the first stage:

```
# add 1st stage predicted values to data frame
TrainExer44 <- TrainExer44 %>% mutate(PG_FIT = mod1_pred)

# regress GC~RI+PG_FIT
mod2 <- lm(GC~RI+PG_FIT, data=TrainExer44)

summary(mod2)
```

```
##
## Call:
## lm(formula = GC ~ RI + PG_FIT, data = TrainExer44)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.054154 -0.025003 -0.008033  0.012243  0.090373
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.01370    0.13386   37.46  < 2e-16 ***
## RI           0.56466    0.04050   13.94 7.46e-14 ***
## PG_FIT      -0.54445    0.04618  -11.79 3.72e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03833 on 27 degrees of freedom
## Multiple R-squared:  0.9668, Adjusted R-squared:  0.9644
## F-statistic: 393.3 on 2 and 27 DF,  p-value: < 2.2e-16
```

(c) Perform a Sargan test to test whether the five instruments are correlated with ε. What do you conclude?

In a Sargan test, we check whether the two-stage least square (2SLS) residuals are correlated with the instruments. If so, this is a sign that the instruments may directly influence the dependent variable and are, therefore, non valid.

For additional notes on how to conduct this test in R, please consult the following link.

```
# Save the residuals from second stage of 2SLS
res_2SLS <- residuals(mod2)

# Regress residuals by instruments
mod3 <- lm(res_2SLS~RI+RPT+RPN+RPU, data=TrainExer44)

# Calculate Sargan value by multiplying R^2 by number of observations (i.e. rows)
Sargan <- summary(mod3)$r.squared * nrow(TrainExer44)

# Determine p-value of Sargan test using chi_sq distribution
Sargan_pval <- 1-pchisq(Sargan, 3)

print(paste("Sargan value =", round(Sargan,3)))
```

```
## [1] "Sargan value = 1.228"
```

```
print(paste("Sargan p-value =", round(Sargan_pval, 3)))
```

```
## [1] "Sargan p-value = 0.746"
```

The null hypothesis in the Sargan test is that our instruments are *not* correlated with the error term, and here, the p-value is sufficiently large enough that we cannot reject the null, meaning our instruments are good.