Applied Microeconometrics - Assignment 1

Walter Verwer & Bas Machielsen

8/31/2021

1. Explain why first differencing the equation does not solve the endogeneity problem of lagged consumption.

The first difference specification is:

$$(\log C_{it} - \log C_{it-1}) = \beta_1 \cdot (\log p_{it} - \log p_{it-1}) + \beta_2 \cdot (\log inc_{it} - \log inc_{it-1}) +$$
$$\beta_3 \cdot (\log ilop_{it} - \log ilop_{it-1}) + \beta_4 + \beta_5 \cdot (\log C_{it-1} - \log C_{it-2}) + u_{it} - u_{it-1}$$

First difference estimation is just OLS estimation with transformed data. For the OLS estimator (in general) to be consistent and unbiased, we need $\text{Cov}(X, U) = 0$, where X is the matrix containing all regressors. In the context of our transformed data, we need $\text{Cov}(\Delta X, \Delta U) = 0$. One of the variables in ΔX is $\Delta \log C_{it-1}$. If we evaluate the covariance between $\Delta \log C_{it-1}$ and ΔU_{it}, we find that:

$$\text{Cov}(\Delta \log C_{it-1}, \Delta U_{it}) =$$
$$\text{Cov}(\beta \Delta X_{it-1} + \beta_5 \Delta \log C_{it-2} + \Delta u_{it-1}, \Delta u_{it}) =$$
$$\text{Cov}(\Delta U_{it-1}, \Delta U_{it}) \neq 0$$

Since we observe that the exogeneity assumption is violated, we can conclude that first differencing the equation does not solve the endogeneity problem of lagged consumption.

2. Anderson & Hsiao propose a specific instrumental variable procedure for the model. Write down and perform the associated first stage regression. Comment on its outcomes.

We have to keep in mind that the first-stage regression contains all the exogenous regressors X from the second stage regression, plus the instrument, C_{it-2}. Hence, the first-stage model is:

$$\log C_{it-1} \widehat{- \log} C_{it-2} = \beta_1 \cdot (\log p_{it} - \log p_{it-1}) + \beta_2 \cdot (\log inc_{it} - \log inc_{it-1}) +$$
$$\beta_3 \cdot (\log ilop_{it} - \log ilop_{it-1}) + \beta_4 + \beta_5 \cdot (\log C_{it-2}) + u_{it-1} - u_{it-2}$$

And the predicted values are to be used as follows in the second-stage regression:

$$(\log C_{it} - \log C_{it-1}) = \beta_1 \cdot (\log p_{it} - \log p_{it-1}) + \beta_2 \cdot (\log inc_{it} - \log inc_{it-1}) +$$
$$\beta_3 \cdot (\log ilop_{it} - \log ilop_{it-1}) + \beta_4 + \beta_5 \cdot (\widehat{C_{it-1} - C_{it-2}}) + u_{it} - u_{it-1}$$

Using the data, we find the following first-stage regression (table 1):

```
## Create the first and second differences
dataset <- dataset %>%
    group_by(region) %>%
    mutate(across(contains("log"),
                ~ .x - dplyr::lag(.x), .names = "l1_{.col}"),
          across(starts_with("log"),
                ~ dplyr::lag(.x) - dplyr::lag(.x, 2), .names = "l2_{.col}"),
          level_quantity = dplyr::lag(logquantity, 2))

## Run the first-stage regression
first_stage_reg <- lm(formula = "l2_logquantity ~ l1_logprice + l1_logincome + l1_logillegal +
    level_quantity",
    data = dataset)
```

Whereas the F-statistic is acceptable (higher than 10), it is not *much* higher than 10, leaving questions about the relevance of the instrument. Indeed, the instrument seems to be lacking statistical relevance, and thus predictive power. The coefficient on level quantity is only -0.0150557 and insignificant at the 10% level. This means that consumption is C_{it-2} does not predict differences $C_{it-1} - C_{it-2}$ well, meaning there is no clear relationship between absolute consumption and (near-)future increases/decreases of consumption.

3. Estimate the specification above using the Anderson & Hsiao approach. Comment on the underlying assumptions, tabulate the results and comment on the outcomes.

```
# Use a package to estimate Anderson-Hsiao
dataset2 <- plm::pdata.frame(dataset, c("region", "year"))

anderson_hsiao <- plm(l1_logquantity ~ l1_logprice + l1_logincome +
                        l1_logillegal + l2_logquantity |
                        l1_logprice + l1_logincome + l1_logillegal +
                        level_quantity,
                   data=dataset2,
                   model="pooling"
          )

# Compare with Manual 2SLS
dataset <- modelr::add_predictions(dataset, first_stage_reg) %>%
        rename("c_instrumented"=pred)

manual_2sls <- lm(data=dataset,
    formula = l1_logquantity ~ l1_logprice + l1_logincome + l1_logillegal + c_instrumented)

stargazer(first_stage_reg, anderson_hsiao, manual_2sls,
          label = "tab:reg", header=FALSE, model.names = FALSE,
          column.sep.width="-5pt",
          dep.var.labels=c("$C_{it-1} - C_{it-2}$",
                            "$C_{it} - C_{it-1}$",
                            "$C_{it} - C_{it-1}$"),
          column.labels = c("First-Stage", "A-H", "Manual 2SLS"),
              omit.stat = c("ll", "ser", "rsq"))
```

The table is displayed below. The estimates from models (2) and (3) in tabel 1 are the same. Only the variance of the 2SLS-estimator is off.

4. Describe the Arellano & Bond GMM estimator for this model.

In general the Arellano & Bond GMM estimator aims to use lagged values of endogenous regressors as an instrument. All possible moment conditions are for $t = 2, \ldots, T$ and $k = 2, \ldots, t$, applied to the current model given by:

$$\mathbb{E}[\log(C_{it-k})(u_{it} - u_{it-1})]$$

2

Table 1:

| | $C_{it-1} - C_{it-2}$ | $C_{it} - C_{it-1}$ | |
	First-Stage	A-H	Manual 2SLS
	(1)	(2)	(3)
l1_logprice	−0.617***	0.022	0.022
	(0.089)	(0.556)	(0.338)
l1_logincome	−0.836***	1.878**	1.878***
	(0.219)	(0.762)	(0.463)
l1_logillegal	−0.005	−0.029	−0.029**
	(0.013)	(0.018)	(0.011)
level_quantity	−0.015		
	(0.009)		
l2_logquantity		1.470*	
		(0.851)	
c_instrumented			1.470***
			(0.517)
Constant	0.086	−0.002	−0.002
	(0.061)	(0.023)	(0.014)
Observations	308	308	308
Adjusted R^2	0.157	0.275	0.398
F Statistic (df = 4; 303)	15.264***	76.600***	51.812***

Dependent variable:

Note: *p<0.1; **p<0.05; ***p<0.01

3

5. Estimate the model parameters using the Arellano & Bond estimator, tabulate the results and discuss the parameter estimates.

```
# Here, we use the _normal_ specification as default and not the first difference:
# The package will transform the data accordingly
arellano_bond <- pgmm(data = dataset2,
                      logquantity ~ logincome + logprice +
                        logillegal + lag(logquantity, 1) + as.numeric(year)
                      | lag(logquantity, 2:99), transformation = 'd', effect = 'individual')

stargazer(arellano_bond, label = "tab:reg_ab", header=FALSE, model.names = FALSE,
          column.sep.width="-5pt",
          column.labels = c('Arellano-Bond'))
```

Table 2:

	Dependent variable:
	logquantity
	Arellano-Bond
logincome	1.662***
	(0.206)
logprice	−0.422***
	(0.043)
logillegal	−0.024**
	(0.011)
lag(logquantity, 1)	0.681***
	(0.030)
as.numeric(year)	−0.018***
	(0.006)
Observations	22
Note:	*p<0.1; **p<0.05; ***p<0.01

6. What is in your estimate for the short-run and the long-run price elasticity of opium?

7. Now estimate the model parameters using the system estimator (Blundell & Bond). Tabulate results, compute the elasticities (as in 6.).

```
blundell_bond <- pgmm(data = dataset2,
                      logquantity ~ logincome + logprice +
                        logillegal + lag(logquantity, 1) + as.numeric(year)|
                        lag(logquantity, 2:99) + lag(l1_logquantity, 2:99), transformation = 'ld',
                        effect = 'individual')

stargazer(blundell_bond, label = "tab:reg_bb", header=FALSE, model.names = FALSE,
          column.sep.width="-5pt",
          column.labels = c('Blundell-Bond'))
```

8. Which parameter estimates do you prefer? Explain why. Are there remaining problems with your preferred estimates?

Table 3:

	Dependent variable:
	logquantity
	Blundell-Bond
logincome	0.287***
	(0.017)
logprice	−0.546***
	(0.041)
logillegal	0.040***
	(0.008)
lag(logquantity, 1)	0.890***
	(0.011)
as.numeric(year)	0.036***
	(0.005)
Observations	22
Note:	*p<0.1; **p<0.05; ***p<0.01

`sargan(arellano_bond)`

Sargan test

data: logquantity ~ logincome + logprice + logillegal + lag(logquantity, . . . chisq = 22, df = 104, p-value = 1 alternative hypothesis: overidentifying restrictions not valid

`sargan(blundell_bond)`

Sargan test

data: logquantity ~ logincome + logprice + logillegal + lag(logquantity, . . . chisq = 22, df = 241, p-value = 1 alternative hypothesis: overidentifying restrictions not valid Conducting the Sargan test on both models, we observe that for both models the null hypothesis is rejected. This means that for both models we find that the model is misspecified.