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:

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

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

$$
\begin{aligned}
&\text{Cov}(\Delta Y_{it-1}, \Delta U_{it}) = \\
&\text{Cov}(\beta \Delta X_{it-1} + \rho \Delta Y_{it-2} + \Delta U_{it-1}, \Delta U_{it}) = \\
&\text{Cov}(\Delta U_{it-1}, \Delta U_{it}) \neq 0
\end{aligned}
$$

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 $\mathbb{X}$ from the second stage regression, plus the instrument, C_{it-2}. Hence, the first-stage model is:

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

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

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

Using the data, we find the following first-stage regression:

```r
## 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. 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.

```r
# 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 -1 |
                          l1_logprice + l1_logincome + l1_logillegal +
                          level_quantity -1,
                      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 - 1)

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.

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

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.).

Table 1:

	Dependent variable:		
	$C_{it-1} - C_{it-2}$	$C_{it} - C_{it-1}$	
	First-Stage	A-H	Manual 2SLS
	(1)	(2)	(3)
l1_logprice	-0.617^{***}	0.037	0.037
	(0.089)	(0.540)	(0.323)
l1_logincome	-0.836^{***}	1.893^{**}	1.893^{***}
	(0.219)	(0.756)	(0.452)
l1_logillegal	-0.005	-0.029	-0.029^{***}
	(0.013)	(0.018)	(0.011)
level_quantity	-0.015		
	(0.009)		
l2_logquantity		1.500^{*}	
		(0.799)	
c_instrumented			1.500^{***}
			(0.478)
Constant	0.086		
	(0.061)		
Observations	308	308	308
Adjusted R^2	0.157	0.274	0.419
F Statistic	15.264^{***} (df = 4; 303)	80.780^{***}	56.549^{***} (df = 4; 304)
Note:			*p<0.1; **p<0.05; ***p<0.01

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