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:

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

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:

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

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):

```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. 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.

```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 |
                          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. Nevertheless, the results show a point estimate that is comparable in magnitude, and both of the point estimates are significantly different from zero. The assumptions underlying the approach are (i) no autocorrelation in the error terms, implying that the autoregressive order is correctly specified, and (ii) weak exogeneity, implying that contemporaneous error terms are unrelated to past values. Thirdly, and a less strict assumption is instrument relevance: the lagged level-endogenous variable should be a relevant instrument, meaning with sufficient power to predict the (contemporaneous) first-differences $Y_{it} - Y_{it-1}$. The interpretation of the estimates is in terms of elasticities of the differences. For example, a percent increase in the price difference between two time periods is associated with a 2,2% increase in consumption between two periods. This result is however not significant at commonly used significance levels. An interesting observation that can be made is the strong positive effect of income changes on opium

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^{***}	0.022	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^{***}
Note:			*p<0.1; **p<0.05; ***p<0.01

consumption. This effect is significant at the 5% level for model (2) and significant at the 1% level for model (3). However, the standard error for model (3) is likely to be off. A final observation that can be made is the strong persistence we find for the 2 period lag of opium consumption, denoted by l2_logquantity. This effect is found to be significant at the 10% level by using the package and at the 1% level using the manual 2SLS. The estimate of l2_logquantity tells us that a percent increase in opium consumption of two periods in the past is associated with a 150% increase in the change of opium consumption two periods later in the future.

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:

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

When there are more moment conditions than identifiable parameters, the A&B estimator has to deal with a overidentification. It does so by employing a weighing matrix of the moment conditions. The ideal weighing matrix minimizes the variance, and is usually obtained by generating a sensible estimate (a first-stage estimate with homoskedastic errors implies a Toeplitz-band structure with 2 on the diagonals, 1 on the off-diagonals), which is then used to estimate the parameters, which are then used to derive new estimates and a new matrix. This two-step procedure should lead to efficient standard errors.

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')

arellano_bond2 <- pgmm(data = dataset2,
                          logquantity ~ logincome + logprice +
                            logillegal + lag(logquantity, 1) + as.numeric(year)
                          | lag(logquantity, 2:5), transformation = 'd', effect = 'individual')
```

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

Because the regression equation is in logs, the short-term price elasticity is simply the coefficient belonging to log Price, which is -0.4217413, meaning that a price increase of 1% means a consumption decrease of 0.42%, which is not small. The long-run price elasticity can be found by calculating the long-run multiplier of that coefficient: $LRM = \frac{\beta_{\log(\text{Price})}}{1-\beta_{\log(\text{Consumption})_{it-1}}}$. Concretely, this means that the long-run multiplier in the model using the Arellano-Bond estimator is -1.320012.

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')

blundell_bond2 <- pgmm(data = dataset2,
                          logquantity ~ logincome + logprice +
                            logillegal + lag(logquantity, 1) + as.numeric(year)|
                            lag(logquantity, 2:5) + lag(l1_logquantity, 2:5), transformation = 'ld',
                            effect = 'individual')

no_obs <- list(arellano_bond, arellano_bond2, blundell_bond, blundell_bond2) %>%
  map(~ summary(.x)) %>%
  map_dbl(~ .x$fitted.values %>%
        length())
```

```
stargazer(arellano_bond, arellano_bond2, blundell_bond, blundell_bond2,
          label = "tab:reg_bb", header=FALSE, model.names = FALSE,
          column.sep.width="-5pt",
          add.lines=list(c("Observations", no_obs)),
           omit.stat = c("ll", "ser", "n"),
          column.labels = c("Arellano-Bond", "Arellano-Bond", "Blundell-Bond", "Blundell-Bond"))
```

Table 2:

	Dependent variable:			
	logquantity			
	Arellano-Bond	Arellano-Bond	Blundell-Bond	Blundell-Bond
	(1)	(2)	(3)	(4)
logincome	1.662***	1.672***	0.287***	0.314***
	(0.206)	(0.242)	(0.017)	(0.019)
logprice	−0.422***	−0.461***	−0.546***	−0.569***
	(0.043)	(0.044)	(0.041)	(0.037)
logillegal	−0.024**	−0.022**	0.040***	0.046***
	(0.011)	(0.010)	(0.008)	(0.009)
lag(logquantity, 1)	0.681***	0.641***	0.890***	0.875***
	(0.030)	(0.034)	(0.011)	(0.012)
as.numeric(year)	−0.018***	−0.019***	0.036***	0.036***
	(0.006)	(0.006)	(0.005)	(0.005)
Observations	308	308	638	638
Note:				*p<0.1; **p<0.05; ***p<0.01

The short-term elasticity is -0.546 for the model in which we used all lags possible (model (3)). For robustness, we used a smaller amount of instruments and find a very comparable estimate of -0.569 (model (4)). Both these results are significant at the 1% level. The long-run elasticity for model (3) is -4.9801619. Model (4) yields a similar long-run elasticity of -4.5489627.

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

```
ab_sargantest <- sargan(arellano_bond)
bb_sargantest <- sargan(blundell_bond)
```

Conducting the Sargan test on both models, we observe that for both models the null hypothesis is not rejected (for the Arellano-Bond model, we have a statistic of 22 and a p-value of 1, and for the Blundell-Bond estimator, we have a statistic of 22 and a p-value of 1 This means that for both models we find no evidence to conclude that the models are misspecified.

We prefer the Blundell-Bond model estimates. The reason is that the system estimator is robust for estimating coefficients of variables that are very persistent through time. Opium consumption is likely to be a variable of such type.

We believe that the following problems remain. First, we believe that our estimates are not causally interpretable. This is because there could be variables that are correlated with opium consumption which are

not captured in the model. If there would be a source of randomization this could possibly be overcome. Second, our estimates seem to be not very robust to our model choice, hence the model choice matters a lot, and this could indicate underlying problems.