Applied Microeconometrics, Assignment 1: Dynamic Panel Data Models

Sindri Engilbertsson (584872), Ilse van der Voort (584098)

September 2021

To answer the questions we have made extensive use of the lecture slides and recommended readings. Data analyses are done in R and Stata.

Q1: Explain why first differencing the equation does not solve the endogeneity problem of lagged consumption.

First differencing gives the following equation:

$$\begin{aligned}\log(\mathrm{C}_{i,t}) - \log(\mathrm{C}_{i,t-1}) =& \beta_1(\log(\mathrm{P}_{i,t}) - \log(\mathrm{P}_{i,t-1})) + \beta_2(\log(\mathrm{I}_{i,t}) - \log(\mathrm{I}_{i,t-1})) + \beta_3(\log(\mathrm{O}_{i,t}) - \log(\mathrm{O}_{i,t-1})) + \\ & \beta_4(t-(t-1)) + \beta_5(\log(\mathrm{C}_{i,t-1}) - \log(\mathrm{C}_{i,t-2})) + U_{i,t} - U_{i,t-1}\end{aligned}$$

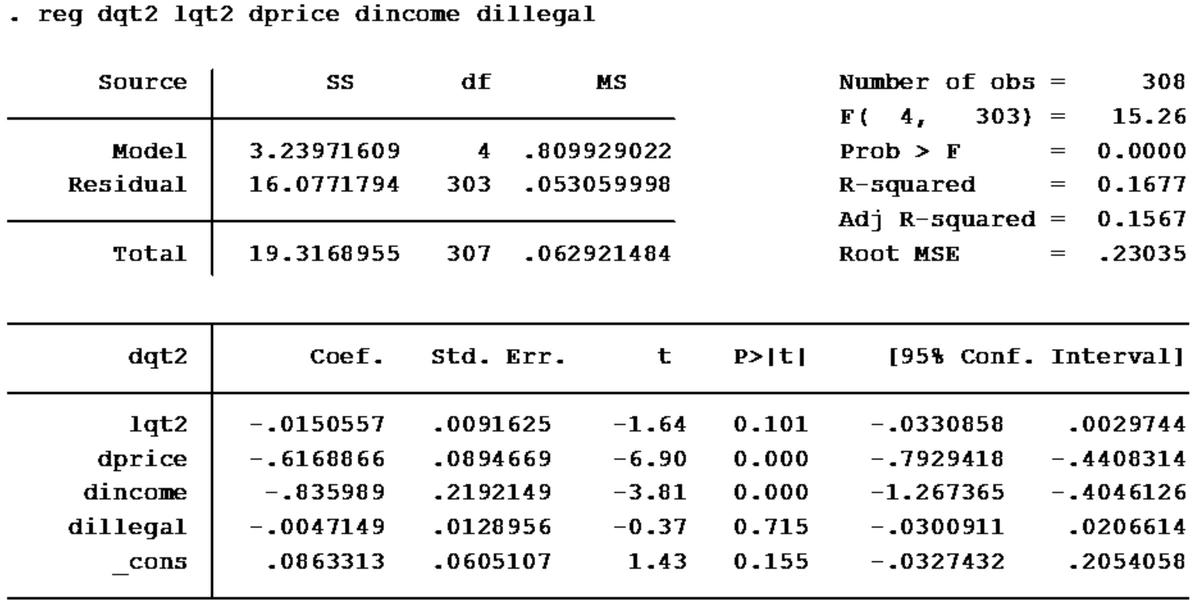
It can be seen that in this equation endogeneity persists as $E[U_{i,t} - U_{i,t-1} | \log(\mathrm{C}_{i,t-1}) - \log(\mathrm{C}_{i,t-2})] \neq 0$. More specifically, this originates from the correlation between $\log(\mathrm{C}_{i,t-1})$ and $U_{i,t-1}$. The equation shows that first-differencing eliminates fixed effects, but the estimator remains biased and inconsistent, even if $T \to \infty$, due to endogeneity.

Q2: Anderson and Hsiao propose a specific instrumental variable procedure for the model. Write down and perform the associated first stage regression. Comment on its outcomes.

For Anderson and Hsiao's 2SLS estimation method to be used, we must first assume that the errors, $U_{i,t}$, are not serially correlated. If there is no serial correlation, we can use the variable $\log(\mathrm{C}_{i,t-2})$ as an instrument for $(\log(\mathrm{C}_{i,t-1}) - \log(\mathrm{C}_{i,t-2}))$, as it is correlated to the regressor, but uncorrelated with the error term $U_{i,t} - U_{i,t-1}$. In addition, assuming the model is correctly specified in the assignment, the exclusion restriction holds. The first stage regression then becomes:

$$\begin{aligned}\log(\mathrm{C}_{i,t-1}) - \log(\mathrm{C}_{i,t-2}) =& \gamma_0 + \gamma_1(\log(\mathrm{P}_{i,t}) - \log(\mathrm{P}_{i,t-1})) + \gamma_2(\log(\mathrm{I}_{i,t}) - \log(\mathrm{I}_{i,t-1})) + \\ & \gamma_3(\log(\mathrm{O}_{i,t}) - \log(\mathrm{O}_{i,t-1})) + \gamma_4(\log(\mathrm{C}_{i,t-2})) + V_{i,t}\end{aligned}$$

Because all first differences of the time variable are equal to one, this variable is eliminated from the equation. Differencing furthermore eliminated the fixed effects. The results of the first-stage regression are displayed below:



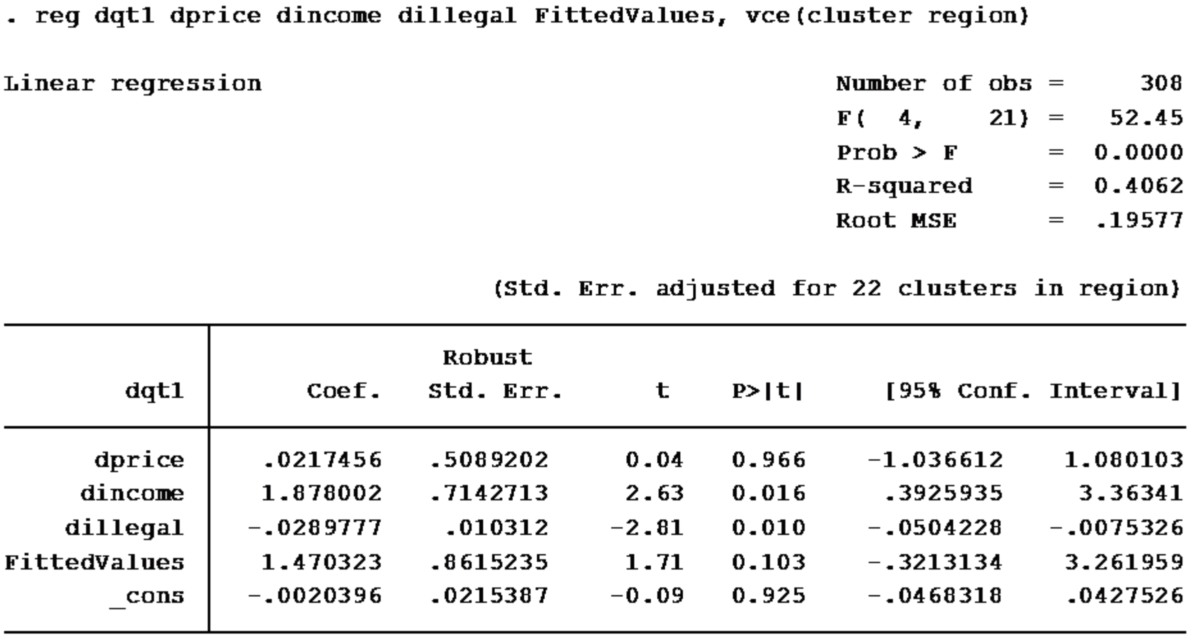
```
. reg dqt2 lqt2 dprice dincome dillegal

      Source |       SS       df       MS              Number of obs =     308
-------------+------------------------------           F(  4,   303) =   15.26
       Model |  3.23971609     4  .809929022           Prob > F      =  0.0000
    Residual |  16.0771794   303  .053059998           R-squared     =  0.1677
-------------+------------------------------           Adj R-squared =  0.1567
       Total |  19.3168955   307  .062921484           Root MSE      =  .23035

------------------------------------------------------------------------------
        dqt2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lqt2 |  -.0150557   .0091625    -1.64   0.101    -.0330858    .0029744
      dprice |  -.6168866   .0894669    -6.90   0.000    -.7929418   -.4408314
     dincome |   -.835989   .2192149    -3.81   0.000    -1.267365   -.4046126
    dillegal |  -.0047149   .0128956    -0.37   0.715    -.0300911    .0206614
       _cons |   .0863313   .0605107     1.43   0.155    -.0327432    .2054058
------------------------------------------------------------------------------
```

We see that the instrument, lqt2, is not significant in the first stage regression at the 10 percent level. This indicates that it might be a weak instrument and that log($C_{i,t-2}$) does not affect log($C_{i,t-1}$) − log($C_{i,t-2}$), while controlling for the other variables. The F-statistic of the joint significance, including the control variables, equals 15.26.

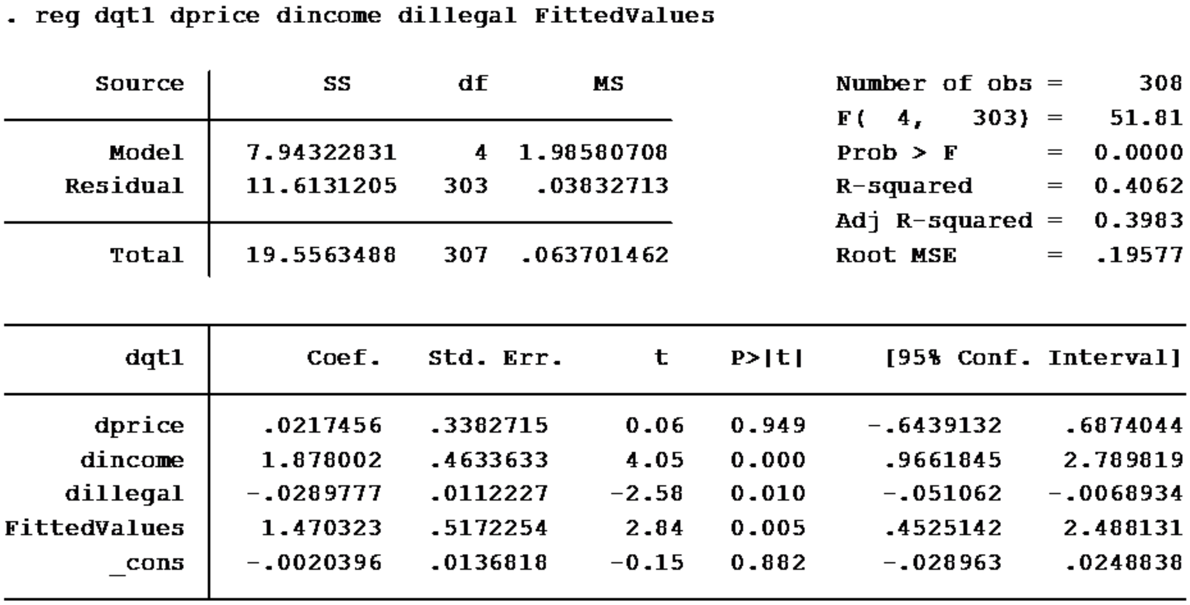
Q3: Estimate the specification above using the Anderson and Hsiao approach. Comment on the underlying assumptions, tabulate the results and comment on the outcomes.

From the aforementioned equation we use the fitted values to estimate the second stage. As mentioned before, there should not be any serial correlation in the errors $U_{i,t}$. Furthermore, the validity, relevance, and exclusion restrictions for IV regression should hold. We argue that the instrument is relevant because of the correlation between log($C_{i,t-2}$) and (log($C_{i,t-1}$) − log($C_{i,t-2}$)). Nevertheless, the insignificance of the instrument in the first stage regression leads us to believe it is a rather weak instrument. We argue that the validity condition holds, because there is no correlation between log($C_{i,t-2}$) and $U_{i,t} - U_{i,t-1}$. Lastly, we argue that the exclusion restriction holds by assumption that the original model is correctly specified.

We use the fitted values of the first stage regression (FittedValues) in our second-stage regression. We estimate the model with standard errors clustered by region and with normal standard errors. The results are displayed below:



```
. reg dqt1 dprice dincome dillegal FittedValues, vce(cluster region)

Linear regression                                      Number of obs =     308
                                                       F(  4,     21) =   52.45
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4062
                                                       Root MSE      =  .19577

                                   (Std. Err. adjusted for 22 clusters in region)
------------------------------------------------------------------------------
             |               Robust
        dqt1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      dprice |   .0217456   .5089202     0.04   0.966    -1.036612    1.080103
     dincome |   1.878002   .7142713     2.63   0.016     .3925935     3.36341
    dillegal |  -.0289777    .010312    -2.81   0.010    -.0504228   -.0075326
FittedValues |   1.470323   .8615235     1.71   0.103    -.3213134    3.261959
       _cons |  -.0020396   .0215387    -0.09   0.925    -.0468318    .0427526
------------------------------------------------------------------------------
```



```
. reg dqt1 dprice dincome dillegal FittedValues

      Source |       SS       df       MS              Number of obs =     308
-------------+------------------------------           F(  4,    303) =   51.81
       Model |  7.94322831     4  1.98580708           Prob > F      =  0.0000
    Residual |  11.6131205   303   .03832713           R-squared     =  0.4062
-------------+------------------------------           Adj R-squared =  0.3983
       Total |  19.5563488   307  .063701462           Root MSE      =  .19577

------------------------------------------------------------------------------
        dqt1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      dprice |   .0217456   .3382715     0.06   0.949    -.6439132    .6874044
     dincome |   1.878002   .4633633     4.05   0.000     .9661845    2.789819
    dillegal |  -.0289777   .0112227    -2.58   0.010     -.051062   -.0068934
FittedValues |   1.470323   .5172254     2.84   0.005     .4525142    2.488131
       _cons |  -.0020396   .0136818    -0.15   0.882     -.028963    .0248838
------------------------------------------------------------------------------
```

The results show that the fitted values are significant at a 1 percent level with an estimated coefficient of approximately 1.47, when we do not use clustered standard errors in the second stage regression. This indicates that an increase in the lagged difference in consumption ($\log(\mathrm{C}_{i,t-1} - \log(\mathrm{C}_{i,t-2}))$, increases the current difference in consumption ($\log(\mathrm{C}_{i,t} - \log(\mathrm{C}_{i,t-1}))$. When we use clustered standard errors, the estimate is no longer significant at at 10 percent level.

In both models, the difference in the logarithm of illegal opium has a significant small negative effect at a 10 percent level. The difference in the logarithm of price does not impact the change in consumption in either model. The change in the logarithm of income has a significant positive effect on the difference of consumption, in the model without clustered standard errors, but no effect in the model with clustered standard errors.

Q4: Describe the Arellano and Bond GMM estimator for this model.

To get the Arellano and Bond estimator we start again with the first difference equation from question 1. We also still assume absence of autocorrelation in the error terms $U_{i,t}$.

The instruments that satisfy the moment conditions $E[Z_{i,t}(U_{i,t} - U_{i,t-1})] = 0$ are all lagged variables of consumption. All possible moment conditions are then $E[\log(\mathrm{C}_{i,t-k}(U_{i,t} - U_{i,t-1})] = 0$, where $t = 2, ..., T$ and $k = 2, ..., T$. Here, there are 15 time periods, so $T = 15$. Replacing $U_{i,t}$ and $U_{i,t-1}$ with the model equations, gives:

$$E[\log(\mathrm{C}_{i,t-k}(\log(\mathrm{C}_{i,t}) - \log(\mathrm{C}_{i,t-1}) + ... - ... - \gamma(\log(\mathrm{C}_{i,t-1}) - \log(\mathrm{C}_{i,t-2}))] = 0$$

In order to estimate γ, Arellano and Bond use GMM. Note that we eliminated several variables from the moment condition as specified above to clearly indicate γ. Because the dataset includes more than 3 time-periods (including $t = 0$), there is overidentification. To get the GMM estimator, we need to minimise the following:

$$\left(\sum_{i=1}^{N} Z_i' \Delta U_i\right)' W_N \left(\sum_{i=1}^{N} Z_i' \Delta U_i\right)$$

Here, Z_i is the matrix that indicates the possible instruments, which are the lagged consumption variables as specified before. This means that, contrary to the previous questions, now not only the second lag of the logarithm of consumption ($\log(\mathrm{C}_{i,t-2})$) is used as an instrument for ($\log(\mathrm{C}_{i,t-1}) - \log(\mathrm{C}_{i,t-2})$), but also further lags. The weighting matrix W_N that Arellano and Bond use assumes homoskedasticity of the errors $U_{i,t}$. The weighting matrix then becomes:

$$W_N = \left(\frac{1}{N}\sum_{i=1}^{N} Z_i' J_N Z_i\right)^{-1}$$

Here, J_N is the same as the matrix specified in the lecture slides.

The estimator $\hat{\gamma}$ that is obtained from minimisation is consistent and asymptotically efficient (assuming homoskedasticity holds).

Q5: Estimate the model parameters using the Arellano and Bond estimator, tabulate the results and discuss the parameter estimates.

For this question we use the plm-package function pgmm, described here. The code is in the accompanying Rmd file, the estimator is displayed in table 1, along with the Blundell & Bond estimator from question 7. The function automatically performs a Sargan test where we cannot reject the null-hypothesis that the instrumental variables are valid and that the model is correctly specified.

All the coefficients, except for illegal opium, are significant at a 1% level. Illegal opium is significant at a 10% level. The table shows that an increase of the price in opium leads to less consumption, a 1% price increase correlating with a -0.42% decrease in consumption. The income effect is greater than 1, indicating that a 1% increase in income leads to a 1.650% increase in consumption. This is possibly because people substitute illegal opium for legal opium when they earn more, increasing the official consumption. The estimated coefficient for illegal consumption is negative. Indicating that more illegal opium intercepted results in less consumption. A 1% increase in illegal opium intercepted leading to 0.024% less legal opium consumed. This might be due to more illegal opium being intercepted being a result of more illegal opium being produced and sold, but this coefficient is not significant at the 5% level in any case. The consumption seems to slowly decrease with time, with a 1 year change leading to a 2% decrease in consumption. Finally, we see that a 1% increase in consumption last period leads to a 0.661% increase in consumption today, unsurprisingly as opium is an addictive substance.

Q6: What is in your estimate for the short-run and the long-run price elasticity of opium?

In the short-run, the coefficient for logprice gives us the price elasticity of opium, while in the long run there is no change between periods so we have that: $C = \alpha P + \beta C \rightarrow C/P = \alpha/(1-\beta)$. Now comparing the estimates based on our different estimates, using Anderson and Hsiao we find that the short-run elasticity equals 0.0217 but is statistically insignificant at a 10 percent level, the long-run elasticity is $\frac{0.217}{1-1.470} \approx -0.0462$. For Arellano and Bond's estimator the short-run elasticity is -0.420 and it is statistically significant at a 1 percent level, and the long-run price elasticity equals $\frac{-0.420}{1-0.661} \approx -1.239$. An increase in the price by 1% according to these estimates then decreases consumption by approximately 0.420%. A permanent increase in prices by 1% decreases consumption by approximately 1.239%.

Q7: Now estimate the model parameters using the system estimator (Blundell and Bond). Tabulate results, compute the elasticities (as in 6.).

Again we turn to the pgmm function and again we cannot reject the Sargan test, null-hypothesis that the instrumental variables are valid and that the model is correctly specified. The results are in Table 1 below.

We see that according to the Blundell-Bond estimator, all of the coefficients are significant at the 1% level. The short-run price elasticity, β_1, equals -0.551, meaning that a 1% increase in price is correlates to a 0.551% decrease in consumption. The income effect is considerably weaker than for the Arellano-Bond estimator, with a 1% increase in the income index correlating with a 0.295% increase in consumption. For the *logillegal* coefficient, contrary to the Arellano-Bond estimator, there is a positive correlation, with a 1% increase in the amount of illegal opium intercepted being correlated with a 0.44% increase in the consumption of legal opium. Also contrary to Arellano and Bond, we see that consumption of legal opium increases with time in our sample, by 3.6%. Finally, unsurprisingly seeing as opium is an addictive drug, a 1% increase in consumption last period is expected to be followed by a 0.884% increase in consumption in the following period.

Table 1:

	Dependent variable:	
	$log(C_{it}))$	
	Arellano-Bond	Blundell-Bond
	(1)	(2)
logprice	-0.420^{***}	-0.551^{***}
	(0.050)	(0.050)
logincome	1.650^{***}	0.295^{***}
	(0.219)	(0.030)
logillegal	-0.024^{*}	0.044^{***}
	(0.014)	(0.012)
as.numeric(year)	-0.020^{***}	0.036^{***}
	(0.007)	(0.006)
lag(logquantity)	0.661^{***}	0.884^{***}
	(0.041)	(0.019)
Observations	22	22
Note:	*p<0.1; **p<0.05; ***p<0.01	

In calculating the elasticities, we use the same method as in question 6. The short-run elasticity equals -0.551 and is statistically significant at a 1 percent level. The long-run price elasticity equals $\frac{-0.551}{1-0.884} \approx -4.750$. In the long-run, a permanent change in prices leads to a decrease in consumption for both models. A permanent rise in prices by 1% decreases consumption by 4.750% in the Blundell and Bond model and by 1.239% in the Arellano and Bond model.

Q8: Which parameter estimates do you prefer? Explain why. Are there remaining problems with your preferred estimates?

To find the best estimates, there are several things to take into account. First, there is a trade-off between the number of instruments we use. The Blundell and Bond estimator differs from Arellano and Bond in that it adds additional moments, based on the level equation. The additional moment conditions are then $E[\Delta \mathrm{C}_{i,t-1}(\eta_i + U_{i,t})] = 0$. However, adding instruments is not necessarily a good thing, because it can increase the small sample IV bias.

Second, another trade-off includes the number of lagged variables used as instruments. The difference between Anderson and Hsiao and Arellano and Bond, is that the latter adds further lagged variables as instruments. However, the further away the lag, the weaker the instrument is likely to be. Adding weak instruments can increase the IV bias.

Third, a problem occurs when the estimate for β_5 in the equation of Q1 is close to 1. This means that in the first-stage regression, the estimated coefficient for the instrument is insignificant and the instrument becomes irrelevant. Thus, γ_4 in the equation from Q2 is insignificantly different from zero. In our first-stage regression in Q2 we indeed see that our instrument is insignificant. This leads us to believe that this might be a valid concern. $\log(\mathrm{C}_{i,t-2})$ would then no longer be a relevant instrument. The system estimator of Blundell and Bond performs better than the other two methods through its additional information on levels instead of differences.

Due to the insignificant estimate of the instrument in Q2, we argue that Blundell and Bond give the preferred estimates. We argue that the issues of adding weak instruments is not a great concern because the Sargan test performed in Q7 clearly shows no reason to assume the model is not specified incorrectly. For the estimates to hold and the Sargan test to be valid, we need the absence of autocorrelation in the errors. The estimates might also be more precise if several instruments (e.g. specific lags) are eliminated, when these specific instruments are weak and may produce more noise than information.