Empirical exercise – The two-stage least squares estimator

1. Aim of the exercise

In this exercise we consider the instrumental variable method to tackle endogeneity. The empirical context is as follows. Using data on cigarette consumption, cigarette prices, and tax on cigarettes, we investigate how the instrumental variable method could help us to analyse the causal effect of cigarette price on cigarette consumption. The data is from *Jonathan Gruber, 2001. Tobacco at the crossroads: the past and future of smoking regulation in the United States, Journal of Economic Perspectives, 15(2), 193-212.* It consists of annual data for the 48 continental U.S. states for years 1985 and 1995. In this exercise we pool the observations, and ignore the time dimension of the data. 'state' indicates the state. 'year' is the year. 'cigarcons' is the quantity consumed, and it is measured by annual per capita cigarette sales in packs per fiscal year, as derived from state tax collection data. 'cigartax' is the general sales tax. It is the average tax, in cents per pack, due to the broad-based state sales tax applied to all consumption goods for the fiscal year. 'cigartaxspecific' is the average excise taxes for fiscal year, including sales taxes. 'cigarprice' is the average retail cigarette price per pack during the fiscal year, including sales taxes. 'income' is the state personal income (total, nominal). 'cpi' is the Consumer Price Index (U.S.). 'pop' is the state population. All prices, income, and taxes are deflated by the Consumer Price Index and thus are in real dollars.

2. Load the data

Load the data to the memory, define the number of observations, and create the constant term.

```
clear;
load 'M:\exercisetsls.mat';
N = size(cigarprice,1);
x_0 = ones(N,1);
```

3. Produce descriptive statistics

Examine the correlation coefficients between the variables of interest using the `corrcoef` function. The function returns as output `R1` and `P1` arguments, which contain the correlation coefficient, and the p value corresponding to the test of the null hypothesis of no correlation. Inspect all the correlations.

```
A = [cigarcons cigarprice cigartax];
[R1,P1] = corrcoef(A);
```

4. An endogenous variable and an instrumental variable

The empirical question of interest is how to decrease a smoker's demand for cigarettes. Increasing the price of cigarettes could be one way. To analyze this relation, we consider the model

$$cigarcons = \beta_0 + \beta_1 cigarprice + u,$$

where *cigarcons* is the number of packs of cigarettes sold per capita in the respective state of the US, and *cigarprice* is the average real price per pack of cigarettes including all taxes. The problem with this model is that *cigarprice* is probably correlated with u. To demonstrate the correlation, assume a model that fully explains *cigarprice* with no error:

$$cigarprice = \alpha_0 + \alpha_1 cigartax + \alpha_2 cigarcons.$$

cigartax is the sales tax on cigarettes measured in dollars per pack. It is a relevant determinant of *cigarprice* because tax is part of the price, and if the tax on cigarettes increase, the price of the cigarette must increase. The correlation matrix also showed a positive and significant correlation between the two variables. Hence, we can motivate that the following holds:

$$\text{Cov}\left[cigartax, cigarprice\right] \neq 0.$$

Furthermore, *cigarcons* is a relevant determinant of *cigarprice* because if the demand for cigarettes increase, the price must increase. The correlation matrix demonstrated a significant and strong correlation between the two variables. Plug the latter model in the former model to obtain

$$cigarcons = \beta_0 + \beta_1(\underbrace{\alpha_0 + \alpha_1 cigartax + \alpha_2 cigarcons}_{cigarprice}) + u.$$

Pay attention to the two components of *cigarprice*, *cigartax* and *cigarcons*. Suppose there is a change, say an increase, in the error term, u. This will increase the demand, that is, the dependent variable *cigarcons*.

Consider the first component of *cigarprice*. The increase in *cigarcons* is not likely to affect *cigartax* because there is no immediate reason we could think of that it should. Hence, we can motivate that the following condition holds:

$$\text{Cov}\left[cigartax, u\right] = 0.$$

Consider the second component of *cigarprice*. The increase in *cigarcons* obviously affects *cigarprice* through *cigarcons* because if the demand increases price must increase. Given this economic argument, we suspect that u and *cigarprice* are correlated, or in econometric parlance, *cigarprice* is *endogenous*. What happens is that price affects demand, as stated in the very first model, but we know that demand affects price too. This is *reverse causation*, or in econometric parlance, *simultaneity*. Hence, we can state our problem as:

$$\text{Cov}\left[cigarcons, u\right] \neq 0.$$

Note that we motivate simultaneity by an economic argument, that there is interaction between price and quantity. We did not motivate any econometric reason for why *caigarprice* would be correlated with u. Later we will carry out a statistical test to see if this is the case.

Consider again the very first model. We wanted to estimate the effect of *cigarprice* on *cigarcons*. The discussion above showed that the estimate of β_1 will be biased because *cigarprice* is probably correlated with u, indirectly through *cigarcons*. The famous conditional mean zero assumption is violated: $\text{E}\left[u \mid cigarprice\right] \neq 0$. We are in trouble, but not by

all means. There are two sources of variation in *cigarprice*. One source comes from *cigartax*, and the other from *cigarcons*. We can leave the problematic variation in *cigarcons* aside in the error, but use the variation in *cigartax* to estimate the effect of *cigarprice* on *cigarcons*. *cigartax* then becomes our instrument, and it seems to be a 'relevant' instrument because we argued that $\text{Cov}\left[cigartax, u\right] = 0$, and it seems to be a 'valid' instrument because we argued that $\text{Cov}\left[cigarcons, u\right] \neq 0$. We can estimate the following model:

$$cigarcons = \beta_0 + \beta_1 \underbrace{(\alpha_0 + \alpha_1 cigartax)}_{cigarprice} + \tilde{u}.$$

Compare this model to the very first model. We use *cigartax* as a proxy for *cigarprice*. But we are also aware that the error term contains a determinant of *cigarprice*, which is *cigarcons*, so that the error is not denoted by u anymore, but by $\tilde{u}$. We note this for later reference.

There is a question hanging in the air. Why do not we directly replace the endogenous variable *cigarprice* with the proxy variable *cigartax*, and regress *cigarcons* on *cigartax*? Endnote 1 explains the reason.

We can estimate the model in two stages. In the first stage we estimate

$$cigarprice = \alpha_0 + \alpha_1 cigartax + e,$$

and obtain the predicted $\alpha_0 + \alpha_1 cigartax$. In the second stage we estimate

$$cigarcons = \beta_0 + \beta_1 \underbrace{(\hat{\alpha}_0 + \hat{\alpha}_1 cigartax)}_{\widehat{cigarprice}} + \tilde{u}$$

where $\widehat{cigarprice}$ denotes predicted *cigarprice*.

To carry out the two stage least squares estimation, we will use the routine in the function file `exercisefunctiontslsrobust.m`. Take your time and inspect the content of this function in detail. It takes three arguments `y`, `X` and `Z` which need to be defined in our script file. The first argument defines the dependent variable, the second argument defines the endogenous variable, and the third argument defines the instrumental variable. More than one variable can be specified in `X` and `Z`. Study in the function file how these arguments are used to calculate the coefficient estimates. The function also considers robust standard errors for the usual reason that the errors of the IV regression might be prone to heteroskedasticity.

The IV estimate of `cigarprice` suggests that a one dollar increase in the price of a pack of cigarette decreases the annual consumption by about one pack (–1.0195), on average. Using the robust standard errors, we can conclude that `cigarprice` is significant at the 0.01 level.

```
y = cigarcons;
X = [x_0 cigarprice];
Z = [x_0 cigartax];
S041 = exercisefunctiontslsrobust(y,X,Z);
```

5. The forbidden regression

We should not carry out the TSLS estimation by considering ordinary least squares estimation in two stages. This leads to wrong predictions of the dependent variable after the second stage,

and cannot be used to calculate residuals, which then cannot be used to calculate standard errors for the coefficient estimates.

To see this, use the routine at the end of the section to carry out ordinary least squares estimation in two stages. Be aware that you are using only `exercisefunctionlssrobust` for these calculations. Observe that the predictions stored in `S052.y_hat` are not the same as those stored in `S041.y_hat_st`. That is, in the second stage of the TSLS routine, in `exercisefunctionlssrobust.m` we calculate the predictions using `X*LSS.B_hat` where `X = [x_0 S051.y_hat]`, while in function `exercisefunctiontslsrobust.m` we calculate the predictions using `X*LSS.B_hat_st` where `X = [x_0 cigarprice]`. That is, in the former case the predicted `cigarprice` from the first stage are used to obtain the predictions after the second stage, which is not correct, whereas in the latter case just the `cigarprice` is used to obtain the predictions after the second stage, which is correct. The latter is what we did in Section 4. However, note that `LSS.B_hat` and `LSS.B_hat_st` are the same. This means that using `X = [x_0 S051.y_hat]` for calculating the second stage estimates is correct, but using it for calculating the second stage predictions is not correct.

Econometric software packages consider a routine like in `exercisefunctiontslsrobust.m` instead of a routine like in `exercisefunctionlssrobust.m` to obtain the correct predictions after the second stage.

```
y = cigarprice;
X = [x_0 cigartax];
S051 = exercisefunctionlssrobust(y,X);
y = cigarcons;
X = [x_0 S051.y_hat];
S052 = exercisefunctionlssrobust(y,X);
```

6. Testing the relevance of the instrument

The first condition the instrument must satisfy is that it is relevant, i.e. it is correlated with the endogenous variable, so that $\text{Cov}\left[cigartax, cigarprice\right] \neq 0$. If the instrument is not relevant, the TSLS estimator will be biased and have a non-normal sampling distribution, even in large samples. The consequence is that we cannot make statistical inference.

We can test whether the instrument we use is relevant. Consider the first stage regression using the routine at the end of the section. The t statistic in the output is testing the hypothesis that the coefficient on the instrument is zero. The p-value is significant at the 0.01 level. Therefore, we reject the null hypothesis that the instrument is not relevant. Or, the R-squared in the output states that the variation in the sales tax of cigarettes explains by itself 49 percent of the variation in cigarette prices. Hence, *cigartax* seems to be a good proxy for *cigarprice*.

```
y = cigarprice;
X = [x_0 cigartax];
S061 = exercisefunctionlssrobust(y,X);
```

7. Testing the exogeneity of the instrument, and failing to do so

The second condition the instrument must satisfy is that it is uncorrelated with the error so that $\text{Cov}[cigartax, u] = 0$. This condition implies that the instrument should affect the dependent variable only through the endogenous variable and not true to error term.

Let us see if we can test this condition. Review the starting model

$$cigarcons = \beta_0 + \beta_1 cigarprice + u.$$

We suspect that *cigarprice* is endogenous, and we consider using *cigartax* as an instrumental variable. We want to test if it is a valid instrument i.e. if $\text{Cov}[cigartax, u] = 0$ is satisfied. We could regress u on *cigartax*, and test if *cigartax* is insignificant. We do not observe u but we can estimate it with the residuals from the stated regression model. There is one problem. The OLS estimates of β_0 and β_1 in the starting model are biased because we suspect that *cigarprice* is endogenous. Therefore we will use the IV estimates of β_0 and β_1 to calculate the residuals. The testing procedure is as follows. First, obtain the IV estimates of β_0 and β_1. Second, obtain the residuals. Third, regress the residuals on the instrumental variable. Finally, observe the t statistic to check if the instrumental variable is significant.

Consider the routine at the end of the section to carry out this test. Observe in the output that the coefficient estimate of `cigartax` is virtually zero. The R-squared is zero. In fact, these results are not driven by the data at hand, but by the construction of the econometric model. Endnote 2 explains the reason. This means that whenever we have a single instrumental variable for an endogenous variable, we always obtain a value of zero for the coefficient estimate of the instrumental variable, and therefore we cannot test the exogeneity of that instrumental variable. However, the reasoning of the test is valid, and we can use it if we have more than one instrument. This is the topic of Section 12.

How do we then motivate that $\text{Cov}[cigartax, u] = 0$ holds? We must rely on economic arguments. Indeed we gave one reason, during our simultaneity discussion, that *cigarcons* would not affect *cigartax* because there is no direct reason we could think that it should.

```
y = cigarcons;
X = [x_0 cigarprice];
Z = [x_0 cigartax];
S071 = exercisefunctiontslsrobust(y,X,Z);
y = S071.u_hat_st;
X = [x_0 cigartax];
S072 = exercisefunctionlssrobust(y,X);
```

8. Testing the endogeneity of *cigarprice*

We suspected that *cigarprice* is correlated with u. This was due to simultaneity that we motivated by an *economic* argument. However, we never tested if indeed $\text{Cov}[cigarprice, u] \neq 0$. If *cigarprice* is not correlated with u, then our economic argument is not supported by our data. Why then did not we test for endogeneity from the outset? After all, if we do not have endogeneity, we would not need to find an instrument and carry out the TSLS estimation. The reason is that we can use the TSLS procedure to check if there is endogeneity. That is, we can compare the OLS and TSLS coefficient estimates to see if they are statistically different from each other. That is, first regress `cigarcons` on `cigarprice`. Obtain the coefficient vector and call it `B`. Second, carry out the TSLS regression, i.e. regress `cigarcons` on `cigarprice`

predicted by `cigartax`. Obtain the coefficient vector and call it `b`. Compare `B` and `b`. There does not appear a substantial difference between the two coefficient estimates from the two regression models. `H_s` formally tests whether the OLS and TSLS coefficient estimates are statistically far from each other. This is the Durbin-Wu-Hausman test. The p-value of the test (`H_p`) suggests that `cigarprice` is not endogenous.

There is still another test of endogeneity. Consider the starting model:

$$cigarcons = \beta_0 + \beta_1 cigarprice + u.$$

Next, consider the first stage regression:

$$cigarprice = \alpha_0 + \alpha_1 cigartax + e.$$

In the starting model *cigarprice* is the endogenous variable. The latter model is settled to instrument this endogenous variable. Given the terms of the latter model, in the first model, the only way *cigarprice* is correlated with u is if e is correlated with u because we know that *cigartax* is not supposed to be correlated with u. Obtain the residual, $\hat{e}$, and consider it in the model for cigarcons: $cigarcons = \beta_0 + \beta_1 cigarprice + \beta_2 \hat{e} + \epsilon$. If it turns out that $\hat{\beta}_2$ is statistically significantly different from 0, then we conclude that *cigarprice* is endogenous. One interpretation of this test is that we check whether the error term in the starting model is just u, or if it is a combination of $\beta_2 \hat{e}$ and ϵ. Another interpretation is the following. Realize that $\hat{e}$ represents *cigarprice* where the exogenous variation in it due to *cigartax* is partialled out. Hence we consider part of *cigarprice* that is potentially endogenous, and check if it is significant. Viewed from this perspective, we can see that the coefficient estimates in the last equation will be equal to the TSLS coefficient estimates, because we in effect are accounting for the variation in *cigarprice* that is endogenous. The last three lines of the syntax at the end of the section proves this. Study the routine of the second test at the end of the section. Observe that the coefficient of the `residauls` is not significant. Once again, we conclude that `cigarprice` is not endogenous.

In both tests, we fail to reject the null hypothesis that *cigarprice* is exogenous. This means that the OLS estimator of β_1 in the starting model is not biased. We do not need to instrument *cigarprice*. In fact, if *cigarprice* is exogenous, TSLS estimates are less efficient than the OLS estimates of the starting model. A comparison of `B_hat_VCE` from OLS estimation with `b_hat_VCE` from IV estimation seems to confirm this.

```
% Test one
y = cigarcons;
X = [x_0 cigarprice];
S081 = exercisefunctionlssrobust(y,X);
y = cigarcons;
X = [x_0 cigarprice];
Z = [x_0 cigartax];
S082 = exercisefunctiontslsrobust(y,X,Z);
B = S081.B_hat;
b = S082.B_hat_st;
B_hat_VCE = S081.B_hat_VCE_robust;
b_hat_VCE = S082.B_hat_VCE_robust_st;
H_s = ((b-B)'/(b_hat_VCE-B_hat_VCE))*(b-B);
```

```
H_p = chi2cdf(H_s,1,'upper');
% Test two
y = cigarprice;
X = [x_0 cigartax];
S083 = exercisefunctionlssrobust(y,X);
y = cigarcons;
X = [x_0 cigarprice S083.u_hat];
S084 = exercisefunctionlssrobust(y,X);
e = 0.00000001;
a = abs(S082.B_hat_st(1,1)-S084.B_hat(1,1)) <= e;
b = abs(S082.B_hat_st(2,1)-S084.B_hat(2,1)) <= e;
```

9. Augment the instrumental variable regression with an exogenous variable

Review the IV model

$$cigarcons = \beta_0 + \beta_1(\underbrace{\hat{\alpha}_0 + \hat{\alpha}_1 cigartax}_{\widehat{cigarprice}}) + \tilde{u}.$$

We have given an economic argument that *cigartax* is not endogenous because it does not suffer from *simultaneity* through *cigarcons*. This argument meant that *cigartax* is an exogenous variable and hence a valid instrument. However, simultaneity is only one of the causes of endogeneity. *cigartax* can also be endogenous if it is correlated with an exogenous variable that is left in $\tilde{u}$. It will then suffer from omitted variable bias. For example, it is believed that sales tax is negatively related to income because the government will tax people living in rich regions through their income rather than their consumption. That is, sales tax could be lower in high income states than in low income states. Hence, *cigartax* could be correlated with income. In the data we observe real per capita state income, documented in the variable *income*. We can check the correlation of *cigartax* with *income* using the the command syntax `[R2,P2] = corrcoef(cigartax,income)`. There appears a significant correlation. To avoid omitted variable bias, we can include *income* in our regression. The augmented model becomes

$$cigarcons = \beta_0 + \beta_1(\underbrace{\hat{\alpha}_0 + \hat{\alpha}_1 cigartax + \hat{\alpha}_2 income}_{\widehat{cigarprice}}) + income + \tilde{u}.$$

Estimate the IV model augmented with *income*. In the code presented at the end of the section, note that we include `income` both in `X` and `Z`. This shows that `income` is an instrument too but it is not excluded from the main regression equation. What is excluded is `cigartax`. Hence, `income` is an 'included instrument' and `cigartax` is an 'excluded instrument'. We observe that the coefficient estimate of `cigarprice` changes only slightly, implying that it does not suffer, or suffers less, from omitted variable bias (through *cigartax*).

```
income = income/1000000;
[R2,P2] = corrcoef(cigartax,income);
y = cigarcons;
X = [x_0 cigarprice income];
Z = [x_0 cigartax income];
```

```
S091 = exercisefunctiontslsrobust(y,X,Z);
```

10. An endogenous variable and two instrumental variables

Finding an instrumental variable is difficult, but sometimes we may have multiple instrumental variables at our disposal. In the data we observe that some states apply a tax that is specific to tobacco products. This is documented in the $cigartaxspecific$ variable. For the same reasons that apply to $cigartax$, we can consider $cigartaxspecific$ as a proxy for $cigarprice$. Augment the model so that our instrument becomes a linear combination of the two instrumental variables and the income. The model becomes

$$cigarcons = \beta_0 + \beta_1(\underbrace{\hat{\alpha}_0 + \hat{\alpha}_1 cigartax + \hat{\alpha}_3 cigartaxspecific + \hat{\alpha}_4 income}_{\widehat{cigarprice}}) + income + \tilde{u}.$$

Estimate the model by TSLS, and compare the estimation results with the results when we had $cigartax$ as the only instrumental variable. That is, compare the output stored in `S101` with the output stored in `S091`. In particular, pay attention to that the standard errors of the coefficient estimates of the second stage become smaller as we consider the additional instrumental variable. It seems this reflects that the generalised instrumental variable (GIV) estimator is at least as efficient as the instrumental variable estimator.

```
y = cigarcons;
X = [x_0 cigarprice income];
Z = [x_0 cigartax cigartaxspecific income];
S101 = exercisefunctiontslsrobust(y,X,Z);
```

11. Testing the relevance of the two instruments

Consider the routine presented at the end of the section to check if `cigarprice` is correlated with `cigartax` and `cigartaxspecific` so that these two variables are relevant instruments for `cigarprice`. In particular, check the joint significance of the two instruments with the F statistic, the individual significance of each variable with the t statistic, and also the explanatory power of the two variables with the R-squared, which all are informative.

```
y = cigarprice;
X = [x_0 cigartax cigartaxspecific income];
O111 = exercisefunctionlssrobust(y,X);
y = cigarprice;
X = [x_0 income];
S112 = exercisefunctionlssrobust(y,X);
F_s = ((S112.RSS-S111.RSS)/2)/(S111.RSS/(N-4));
F_p = fcdf(2,N-4,F_s,'upper');
```

12. Testing the exogeneity of the two instruments

Review the starting model

$$cigarcons = \beta_0 + \beta_1 cigarprice + u.$$

We suspect that *cigarprice* is endogenous, and we have two instrumental variables. We want to test if the two instrumental variables are indeed exogenous. That is, we want to test if $\text{Cov}[cigartax, u] = 0$ 'and' $\text{Cov}[cigartaxspecific, u] = 0$ are satisfied. We can regress u on *cigartax* and $cigartaxspecific$, and test if they are jointly not significant. The test implies that we can only test if at least one of the stated two conditions hold. We do not observe u but we can estimate it with the residuals from the starting model. However, there is one problem. The OLS estimates of β_0 and β_1 in this model are biased because we suspect that *cigarprice* is endogenous, and the residuals from this model are not reliable. Therefore we will use the IV estimates of β_0 and β_1.

The test is carried out as follows. First, carry out the IV regression. Second, obtain the residuals from this model. Third, regress the residuals on the instrumental variables and the exogenous variable *income*. That is, the model to be estimated is $\hat{u} = \alpha_0 + \alpha_1 cigartax + \alpha_2 cigartaxspecific + \alpha_3 income + \epsilon$. Finally, compute $N * R^2$, which has a chi-squared distribution asymptotically with degrees of freedom equal to the number of instruments less the number of endogenous variables. Alternatively, use the F statistic to test the null hypothesis $H_0 : \alpha_1 = 0, \alpha_2 = 0$, against the alternative $H_1 : \alpha_1 \neq 0$ and/or $\alpha_2 \neq 0$. The p-value of the test suggests that at least one of the instruments is exogenous. This is the Sargan statistic for testing whether the overidentifying restrictions are valid. There are other statistics for the same test. Hansen's J statistic is one of them. We do not cover them here.

```
y = cigarcons;
X = [x_0 cigarprice income];
Z = [x_0 cigartax cigartaxspecific income];
O121 = exercisefunctiontslsrobust(y,X,Z);
y = O121.u_hat_st;
X = [x_0 cigartax cigartaxspecific income];
S122 = exercisefunctionlssrobust(y,X);
overid_s = N*S122.R2_c;
overid_p = chi2cdf(overid_s,1,'upper');
```

† Endnotes

1. We do not directly regress *cigarcons* on *cigartax*. That is, we do not directly estimate the model

$$cigarcons = \alpha_0 + \alpha_1 cigtax + u,$$

but we estimate

$$cigarcons = \beta_0 + \beta_1 \underbrace{(\alpha_0 + \alpha_1 cigtax)}_{cigprice} + \tilde{u}.$$

This is because we are interested in the effect of *cigarprice* on *cigarcons* measured by β_1. We are not interested in the effect of *cigartax* on *cigcons* that is measured by α_1 in the former model. In fact, *cigartax* should not have a direct effect on *cigarcons* but have an effect through *cigarprice*.

2. Consider the residuals from the second step of the test:

$$\hat{u} = cigcons - \hat{\beta}_1 cigprice$$

To make the algebra easier, the constant is ignored. $\hat{\beta}_1$ is the IV estimate from the first step of the test. Since we did not include a constant, it can be shown that $\hat{\beta}_1$ is equal to

$$\hat{\beta}_1 = (cigtax^{'}\ cigprice)^{-1}\ cigtax^{'}\ cigcons.$$

Plug this term in the residual equation, and multiply both sides of the equation with $cigartax$ which will give

$$cigtax^{'}\ \hat{u} = cigtax^{'}\ cigcons - (cigtax^{'}\ cigprice)^{-1}\ cigtax^{'}\ cigcons\ cigtax^{'}\ cigprice.$$

The terms cancel out, and the right hand side becomes zero. This means that by construction the residuals will always have a zero correlation with the instrumental variable, if this instrumental variable is the only instrumental variable for the endogenous variable. Therefore, we cannot test if the instrumental variable has explanatory power for the residuals; which was the idea behind the test we are considering.