R300 – Advanced Econometric Methods

PROBLEM SET 5 - SOLUTIONS

Due on Mon. November 16, 2020

1. Consider the classical linear regression model with two scalar regressors x_i and z_i. Set up the LM statistic for the null that the coefficient on z_i is zero. This is a test for an omitted variable.

The model is

$$y_i = x_i\beta + z_i\gamma + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2).$$

We already know that the information matrix is block diagonal so there is no need to consider the contribution of not knowing σ^2 here. The scores for the regression slopes β, γ are

$$-\sum_i x_i(y_i - x_i\beta - z_i\gamma)/\sigma^2, \qquad -\sum_i z_i(y_i - x_i\beta - z_i\gamma)/\sigma^2,$$

respectively. Evaluation in the constrained estimator gives

$$-\sum_i x_i\hat{\varepsilon}_i/\hat{\sigma}^2 = 0, \qquad -\sum_i z_i\hat{\varepsilon}_i/\hat{\sigma}^2,$$

where

$$\hat{\varepsilon}_i = y_i - x_i\hat{\beta}, \qquad \hat{\beta} = \frac{\sum_i x_i y_i}{\sum_i x_i^2}, \qquad \hat{\sigma}^2 = \frac{\sum_i \hat{\varepsilon}_i^2}{n}.$$

The part of the information matrix relating to the slope coefficients is estimated as

$$\frac{1}{\hat{\sigma}^2}\begin{pmatrix} \sum_i x_i^2 & \sum_i x_i z_i \\ \sum_i x_i z_i & \sum_i z_i^2 \end{pmatrix}$$

and its inverse equals

$$\frac{\hat{\sigma}^2}{\sum_i x_i^2 \sum_i z_i^2 - (\sum_i x_i z_i)^2}\begin{pmatrix} \sum_i z_i^2 & -\sum_i x_i z_i \\ -\sum_i x_i z_i & \sum_i x_i^2 \end{pmatrix}.$$

The LM statistic therefore becomes

$$\frac{(\sum_i z_i\hat{\varepsilon}_i)^2}{\hat{\sigma}^2}\frac{\sum_i x_i^2}{\sum_i x_i^2 \sum_i z_i^2 - (\sum_i x_i z_i)^2} = n\left(\frac{\sum_i z_i\hat{\varepsilon}_i}{\sqrt{\sum_i z_i^2}\sqrt{\sum_i \hat{\varepsilon}_i{}^2}}\right)^2\left(\frac{1}{1 - \left(\frac{\sum_i x_i z_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i z_i^2}}\right)^2}\right)$$

which we can compactly write as

$$n\frac{\hat{\rho}^2}{1-\hat{\varrho}^2}$$

for $\hat{\rho}$ the (uncentered) sample correlation between z_i and $\hat{\varepsilon}_i$ and $\hat{\varrho}$ the (uncentered) sample correlation between z_i and x_i

---

2. Let x_i be binary with success probability $\theta \in (0,1)$. For a sample of size n write down the LR, LM, and Wald test for the null $\theta = \theta_0$ and two-sided alternative $\theta \neq \theta_0$.

The MLE of θ is $\hat{\theta} = \overline{x}$. By the sample mean theorem its asymptotic variance is $\theta(1-\theta)$. The Wald statistic thus is

$$n\,\frac{(\hat{\theta}-\theta_0)^2}{\hat{\theta}(1-\hat{\theta})}.$$

The likelihood function for Bernoulli is

$$\prod_i \theta^{x_i}\,(1-\theta)^{1-x_i} = \theta^{n\overline{x}}\,(1-\theta)^{n-n\overline{x}} = \theta^{n_1}\,(1-\theta)^{n_0}$$

where n_1 and n_0 are the number of success and failures in the sample, respectively. The log-likelihood thus is

$$n_1\log(\theta) + n_0\log(1-\theta).$$

The LR statistic becomes

$$-2\big((n_1\log(\theta_0)+n_0\log(1-\theta_0))-(n_1\log(\hat{\theta})+n_0\log(1-\hat{\theta}))\big) = 2n_1\log\left(\frac{\hat{\theta}}{\theta_0}\right)+2n_0\log\left(\frac{1-\hat{\theta}}{1-\theta_0}\right).$$

The score for Bernoulli is

$$\frac{n_1}{\theta}-\frac{n_0}{1-\theta} = \frac{n_1(1-\theta)-n_0\theta}{\theta(1-\theta)} = \frac{n_1-n\theta}{\theta(1-\theta)} = n\frac{(n_1/n)-\theta}{\theta(1-\theta)} = n\frac{\hat{\theta}-\theta}{\theta(1-\theta)}$$

The variance of the score is $n/\theta(1-\theta)$. so that the LM statistic becomes

$$n\frac{\hat{\theta}-\theta_0}{\theta_0(1-\theta_0)}\left(\frac{\theta_0(1-\theta_0)}{n}\right)n\frac{\hat{\theta}-\theta_0}{\theta_0(1-\theta_0)} = n\,\frac{(\hat{\theta}-\theta_0)^2}{\theta_0(1-\theta_0)}.$$

This differs from the Wald statistic only in how the information is estimated (under the null here).

---

3. Consider the regression

$$y_i = x_i\beta + \varepsilon_i,$$

with ε_i mean-zero, homoskedastic, and independent of (scalar) x_i. Suppose, further, for simplicity that you know the variance of the errors. Consider a situation where each observation can be categorized into one of G mutually-exclusive groups; so $i \in g$ for one $g = 1, \ldots, G$ (An example is student i in classroom g or firm i in sector g.) Rather than (y_i, x_i), we observe averages at the group level, i.e, $\overline{y}_g = |g|^{-1} \sum_{i \in g} y_i$ and $\overline{x}_g = |g|^{-1} \sum_{i \in g} x_i$ for example, average test scores within a classroom.

(i) Show that the error in the regression

$$\overline{y}_g = \overline{x}_g\beta + \overline{\varepsilon}_g$$

is heteroskedastic, in general.

(ii) Under what condition(s) on the groups will $\overline{\varepsilon}_g$ be homoskedastic?

(iii) Does least-squares applied to the averaged regression lead to an unbiased estimator?

The following two questions are optional (as we have not quite covered the relevant material yet) but encouraged.

(iv) Is ordinary least-squares (semiparametrically) efficient here?

(v) If your answer to (iv) is negative can you give an alternative? (Think about constructing an estimator derived from a moment condition that restores the information equality. Check Gauss-Markov and weighted least squares.)

(i) As

$$\overline{\varepsilon}_g = |g|^{-1} \sum_{i \in g} \varepsilon_i$$

and the ε_i are i.i.d. $(0, \sigma^2)$ we have

$$\operatorname{var}(\overline{\varepsilon}_g) = \frac{\sigma^2}{|g|}.$$

This varies with group size $|g|$.

(ii) This will happen if and only if all groups are of the same size.

(iii) This is immediate, as the average error is independent of the average regressors.

(iv) With heteroskedasticity the Gauss-Markov theorem no longer applies. The estimator is

$$\hat{\beta} = \beta + \frac{\sum_g \overline{x}_g \overline{\varepsilon}_g}{\sum_g \overline{x}_g^2}.$$

Its variance is

$$\text{var}(\hat{\beta}) = \sigma^2 \frac{\sum_g \frac{\overline{x}_g^2}{|g|}}{\left(\sum_g \overline{x}_g^2\right)^2}.$$

(v) The best unbiased estimator in this problem is the weighted least-squares estimator

$$\check{\beta} = \arg\min_b \sum_g |g| \left(\overline{y}_g - \overline{x}_g b\right)^2$$

which satisfies

$$\check{\beta} = \beta + \frac{\sum_g |g| \, \overline{x}_g \overline{\varepsilon}_g}{\sum_g |g| \, \overline{x}_g^2}.$$

Its variance is

$$\text{var}(\check{\beta}) = \sigma^2 \frac{\sum_g |g| \, \overline{x}_g^2}{\left(\sum_g |g| \, \overline{x}_g^2\right)^2} = \frac{\sigma^2}{\sum_g |g| \, \overline{x}_g^2}.$$

---

4. We have gathered survey data on health expenditure. For 1691 individuals we have access to the following variables

- eheal: expenditure on health-related products and services during the year (in British pound Sterling);
- income: total income during the year (in British pound Sterling);
- sex: gender dummy that equals one when individual is male and zero when female.

We presume that the conditional mean function takes the following form:

$$E(\text{eheal}|\text{income}, \text{sex}) = \beta_0 + \beta_1 \, \text{income} + \beta_2 \, \text{sex} + \beta_3 \, (\text{income} \times \text{sex}).$$

The coefficients in this model were estimated, and the results are given in Figure 1 below. The first column (Coef.) contains the point estimates, the second column (Std. Err.) provides the appropriate standard errors for statistical inference on these coefficients.

(i) Write down the equation for the conditional mean for males and for females separately.

(ii) Derive the average marginal effect of income on health expenditure for both males and females.

(iii) Formulate and test the hypothesis that the expected change in health expenditure resulting from a change in income is the same for males and females. That is, define H_0 and H_1, say which test procedure you use, and present your conclusion.

(iv) I tested the joint hypothesis H_0 : "$\beta_2 = 0$ and $\beta_3 = 0$" against "H_1 : H_0 is false". Interpret this null hypothesis in words.

v) The p-value of the test from the previous question was zero up to the fourth decimal digit. What does this mean?

Figure 1: Estimation results for Question 4

Source	SS	df	MS		
Model	338619013	3	112873004		
Residual	4.2621e+09	1687	2526417.94		
Total	4.6007e+09	1690	2722299.45		

Number of obs	=	1691			
F(3, 1687)	=	44.68			
Prob > F	=	0.0000			
R-squared	=	0.0736			
Adj R-squared	=	0.0720			
Root MSE	=	1589.5			

eheal	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	.0391367	.0068661	5.70	0.000	.0256697	.0526036
sex	-240.5175	145.4127	-1.65	0.098	-525.7259	44.69081
income_sex	-.038963	.0083186	-4.68	0.000	-.055279	-.0226471
_cons	869.0945	109.2712	7.95	0.000	654.773	1083.416

(i) We have

$$E(\text{eheal}|\text{income}, \text{sex} = 1) = \beta_0 + \beta_1 \,\text{income}$$
$$E(\text{eheal}|\text{income}, \text{sex} = 0) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\,\text{income}$$

(ii) The average marginal effect for males is β_1 and for females is $\beta_1 + \beta_3$.

(iii) The null is that the marginal effects from (ii)—i.e., the slopes of the regression line—are the same for males and females. Thus, $H_0 : \beta_3 = 0$. The alternative is that the slopes are different. Hence, $H_1 : \beta_3 \neq 0$. The table contains the t-statistic for this null (-4.68).

The p-value is zero up to 3 decimal digits. Consequently we reject the null in favor of the alternative for all conventional significance levels. The regression slopes are not the same.

(iv) This null states the the regression line is the same for males and females. Both the slope and intercept are equal under the null.

(v) This again means that we can reject the null that the regression lines are the same.