Assignment Part I

G. J. Bouwens 2701442 g.j.bouwens@student.vu.nl
R. F. Kenny 2691815 r.f.kenny@student.vu.nl
S. V. van Put 2673309 s.v.van.put@student.vu.nl
A. G. Roepstorff 2781130 a.g.roepstorff@student.vu.nl

Computational Methods in Econometrics 2022

September 30, 2022

Introduction

In this assignment, we intend to test whether autocorrelation is present in the regression residuals of a parametric time series regression model. The observations and regressors are given, so do not need to be simulated.

To set up a test, a rejection region needs to be classified under the null hypothesis. There are many options for the value of the null, but the choice affects the distribution of the test statistic, which depends on an underlying distribution. Yet, the value that we choose for our null does not matter for the distribution, as the test statistic is a pivot under the null.

Whenever a model has no autocorrelation, the correlation parameter ρ will have a value of 0. So we proceed by testing $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$ using the Durbin-Watson test. which is $d = 2(1 - \rho)$. Then, under the null hypothesis that we use, d is equal to the value 2.

The statistic for d is:

$$\hat{d} = \frac{\sum_{t=2}^{n}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=2}^{n}\hat{\varepsilon}_{t-1}^2}$$

Exercise 1

The test statistic, $\hat{d}$ above, is a pivot under the null. As the cdf $G_n(\cdot, F) = G_n(\cdot)\forall F \in M$ (F=underlying population, M=parametric model in this case). So a distribution G_n does not depend on the underlying population F. Such that F does not need to be known as the statistic's distribution does not depend on it. So, we can derive $G_n(\cdot)$, and then derive $P(T_n \in R_T|H_0) = P(T_n \in R_T|G_n)$. As $\varepsilon_t \sim NID(0, \sigma^2)$ holds under the null, it follows that $\hat{d} \sim N(0, 1)$ independent of $\frac{(n-k)\hat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-k}$. So $\hat{d} \sim t_{n-k}$. So as $\varepsilon_t \sim NID(0, \sigma^2)$, the residuals will be simulated from the t-distribution, and not N(0,1) (as not IID).

In derivation, using $\hat{\varepsilon} = M_x\varepsilon$ & $\varepsilon = \sigma u$ & r_t = t'th row of M_x

$$d = \frac{\sum_{t=2}^{n}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=2}^{n}\hat{\varepsilon}_{t-1}^2} = \frac{\sum_{t=2}^{n}(r_t\varepsilon_t - r_t\varepsilon_{t-1})^2}{\sum_{t=2}^{n}(\sigma^2 u'_{t-1}M_x u_{t-1})^2} = \frac{\sum_{t=2}^{n}\sigma^4(r_t u_t - r_t u_{t-1})^2}{\sum_{t=2}^{n}\sigma^4(u'_{t-1}M_x u_{t-1})^2} = \frac{\sum_{t=2}^{n}(r_t u_t - r_t u_{t-1})^2}{\sum_{t=2}^{n}(u'_{t-1}M_x u_{t-1})^2}$$

As you can see, this last results does not depend on the parameters from $\frac{(n-k)\hat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-k}$, and hence, the text above applies to our assignment.

Exercise 2

Use Monte Carlo to get the rejection region R_d and p-value for $\alpha = 0.1$ (10% significance) by use of the Durbin-Watson test statistic, given X and y.

$$R_d = (0, c_1) \cup (c_2, 4) \text{ with } c_1 = t^*_{1-(\alpha/2)(B+1)} \text{ and } c_2 = t^*_{1+(\alpha/2)(B+1)}$$

$$p_{MC} = 2min\{\frac{r_l + 1}{B+1}, \frac{r_u + 1}{B+1}\}$$

, where r_u is the number of simulated test statistics that are bigger than or equal and r_l the number of simulated test statistics that are smaller than or equal to the actual test statistic.

$$R_d = (0, 1.577) \cup (2.545, 4)$$

This rejection region is quite good as the actual d-statistic (derived by using X and y) is 1.944, which lies close to the mean of the simulated d-statistic, namely 2.024.

$$p_{MC} = 2min\{\frac{5797 + 1}{B+1}, \frac{4203 + 1}{B+1}\} = 2min\{0.580, 0.420\}$$

Pseudocode:

1. load the given datasets X (regressors) and y (observables).
2. define the global and local (starting) values.
   (a) We choose N (number of iterations and thus datapoints) as 10000. As the datasets, Y_0 and Y_1 are simulated this many times also. Also because, when B tends to inf, the empirical cdf a.s. tends to come close to the theoretical cdf.
   (b) We choose B (number of test statistics) as 9999. Such that the test becomes exact when adding 1 and multiplying by α.
   (c) We choose the k as 9. As X consists of 9 regressors.
   (d) We set a random seed, such that the randomized residuals remain the same for each simulation. (mean is zero as given in the assignment). We simulate from the t_{n-k} distribution.
   (e) We set n (number of (rows) observations in X and y) to 100.
3. find $\hat{d}$ by use of Monte Carlo simulation:
   (a) simulate $\varepsilon^*_1 ... \varepsilon^*_n \sim t_{n-k}$
   (b) calculate the DW statistic as $\hat{d} = \frac{\sum_{t=2}^n (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=2}^n \hat{\varepsilon}_{t-1}^2}$,with $\hat{\varepsilon^*_n} = M_x \varepsilon$
   (c) repeat a and b for B times and store the generated values for the DW statistic: $DW^*_1 ... DW^*_B$ in an ordered array.
4. find the approximated rejection region $R_d = (0, c_1) \cup (c_2, 4)$ by taking the $t^*_{1-(\alpha/2)(B+1)}$ and $t^*_{1+(\alpha/2)(B+1)}$'th DW test statistic for c_1 and c_2, respectively. (Make sure that the $(\alpha/2)(B+1)$ part is an integer to get an exact test.

Exercise 3

Does Y_0 ($\rho = 0$ so H_0) fit this in the rejection region? (for 10% significance)

Given an average of around 1.79 and the $R_d = (0, 1.577) \cup (2.545, 4)$ it seems likely the value will be inside the rejection region. In our study 12.14% of the test statistics we calculated from data simulated under the null lie in the rejection region. This means our test is not totally accurate, but can still be considered useful for inference. We expect better results with a larger n.

Pseudocode:

1. load the given datasets X (regressors) and Y_0 (simulated observables under H_0).
2. use the same global and local (starting) values.
3. find $\hat{d}$ by use of Monte Carlo simulation:
   (a) simulate $\varepsilon_1^* ... \varepsilon_n^* \sim t_{n-k}$
   (b) calculate $betas = A \cdot X' \cdot \varepsilon_i^*$
   (c) calculate all $\hat{\varepsilon}_n^* = Y_0 - X \cdot betas$
   (d) calculate the DW statistic as $\hat{d} = \frac{\Sigma_{t=2}^{n}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\Sigma_{t=2}^{n}\hat{\varepsilon}_{t-1}^2}$,with $\hat{\varepsilon}_n^*$ as in c.
   (e) repeat a until d for all Y_0 its N columns and store the generated values for the DW statistic: $DW_1^* ... DW_B^*$ in an ordered array.
4. compare all generated DW with the R_d found in exercise 2.
5. give the percentage that falls out of this region (should be around 10%).

Exercise 4

The Durbin-Watson test has the following $(1-\alpha)$ Confidence Interval:

$$[1 - \frac{\hat{d_{up}}}{2}, 1 - \frac{\hat{d_{low}}}{2}]$$

$$[-0.273, 0.212]$$

Exercise 5

Does the set Y_1 ($\rho = 0.2$ so H_1) fit this Rejection region? (for 10% significance)

Given an average of around 1.43 and the $R_d = (0, 1.577) \cup (2.545, 4)$ it seems likely the value will be outside the rejection region. Except for some outliers. In our study 78.22% of the calculated test statistics do not lie in the rejection region. Comparing this to 90% we would like to get, it is evident that our test is not very accurate. It must noted however, that $\rho = 0.2$ is relatively close to the null ($\rho = 0$). Therefore it is unreasonable to expect any test to be correct 90% of the time.

Pseudocode:

1. load the given datasets X (regressors) and Y_1 (simulated observables under H_1).
2. use the same global and local (starting) values.
3. find $\hat{d}$ by use of Monte Carlo simulation:
   (a) simulate $\varepsilon_1^* ... \varepsilon_n^* \sim t_{n-k}$
   (b) calculate $betas = A \cdot X' \cdot \varepsilon_i$
   (c) calculate all $\varepsilon_i = Y_1 - X \cdot betas$
   (d) calculate the DW statistic as $\hat{d} = \frac{\Sigma_{t=2}^{n}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\Sigma_{t=2}^{n}\hat{\varepsilon}_{t-1}^2}$,with $\hat{\varepsilon_n^*}$ as in c.
   (e) repeat a until d for all Y_1 its N columns and store the generated values for the DW statistic: $DW_1^* ... DW_B^*$ in an ordered array.
4. compare all generated DW with the R_d found in exercise 2.
5. give the percentage that falls out of this region (should be around 90%).