MOOC Econometrics

Case project

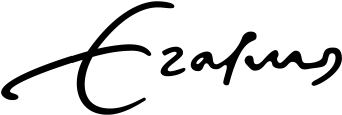
**Notes:**

* See website for how to submit your answers and how feedback is organized.
* This exercise uses the datafile Case GDP and requires a computer.
* The dataset Case GDP is available on the website.

**Goals and skills being used:**

* Get hands-on experience with applying econometric methods.
* Apply techniques and interpret results related to discrete choice models.
* Apply techniques and interpret results using time series models.

**Background**

A good understanding of the macroeconomic cycle with alternating recession and expansion periods (also known as the business cycle) is important for various decision makers. Macroeconomic policy is often based on predictions of this cycle, and such predictions can influence investment decisions of large companies. Central banks and other institutions often publish so-called leading indicators that are helpful to predict the state of the economy. These indicators are based on macroeconomic series like job formation, interest rates, credit, demand, and supply.

In this case project you will predict GDP growth by using quarterly data on a hypothetical economy from 1950 quarter 1 to 2015 quarter 4. The data set contains the GDP of the economy and two leading indicators li1 and li2. In order to evaluate the predictive performance of econometric models, you need to split the data in two parts. As estimation sample you take the period from 1951 to 2010 (240 observations), and as evaluation sample you take the period from 2011 to 2015 (20 observations). The first year of data (1950) is used only to create lags of variables.

The project consists of two parts. In the first part (a-c) you use logit models to predict whether the economic situation improves or declines, and in the second part (d-g) you use time series models to predict the size of the growth rate of the economy.

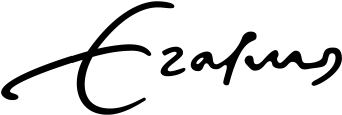
**Data**

The data file Case GDP contains the following variables:

* DATE: Date of the observation;
* GDP: Gross Domestic Product of the economy;
* GDPIMPR: dummy variable indicating whether the GDP has increased (1) or decreased (0);
* LOGGDP: Log of Gross Domestic Product;
* GrowthRate: Relative growth of the economy: GrowthRate*t* = log(GDP*t*) − log(GDP*t*−1);
* li1: First leading indicator;
* li2: Second leading indicator;
* T: Linear trend (where the first observation, for 1950 quarter 1, is defined as 0).

1. The table below summarizes the outcomes of four logit models to explain the direction of economic development (GDPIMPR) for the period 1951 to 2010. Perform three Likelihood Ratio tests to prove both the individual and the joint significance of the 1-quarter lags of li1 and li2, where the alternative hypothesis is always the model with both indicators included.

|  |  |  |  |
| --- | --- | --- | --- |
| Dependent variable: GDPIMPR  Sample size: 240 | |  |  |
| Variable Coeff | Coeff | Coeff | Coeff |
| Constant 0.693 | 0.812 | 0.636 | 0.729 |
| li1(-1) *x* | −0.340 | *x* | −0.372 |
| li2(-1) *x* | *x* | −0.087 | −0.120 |
| Log likelihood −152.763 | −139.747 | −149.521 | −134.178 |

1. It could be that the leading indicators lead the economy by more than 1 quarter. The table below summarizes outcomes of four logit models that differ in the lags of the indicators. For what reason can we use McFadden R2 to select the best lag structure among these four models? Compute the four values of McFadden R2 (with four decimals) and conclude which model is optimal according to this criterion.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dependent variable: GDPIMPR  Sample size: 240 | | |  |  |
|  | 1 | 2 | 3 | 4 |
| Variable | Coeff | Coeff | Coeff | Coeff |
| Constant | 0.729 | 0.731 | 0.746 | 0.749 |
| li1(-1) | −0.372 | −0.366 | *x* | *x* |
| li1(-2) | *x* | *x* | −0.429 | −0.421 |
| li2(-1) | −0.120 | *x* | −0.131 | *x* |
| li2(-2) | *x* | −0.121 | *x* | −0.129 |
| Log likelihood | −134.178 | −134.126 | −130.346 | −130.461 |

1. Use the logit model 3 of part (b) (with li1(-2) and li2(-1)) to calculate the predicted probability of economic growth for each of the 20 quarters of the evaluation sample. Assess the predictive performance by means of the prediction-realization table and the hit rate, using a cut-off value of 0.5. Evaluate the outcomes.
2. Perform the Augmented Dickey-Fuller test on LOGGDP to confirm that this variable is not stationary. Use only the data in the estimation sample and include constant, trend, and a single lag in the test equation (*L* = 1, see Lecture 6.4). Present the coefficients of the test regression and the relevant test statistic, and state your conclusion.
3. Consider the following model: GrowthRate*t* = *α* + *ρ*GrowthRate*t*−1 + *β*1li1*t*−*k*1 + *β*2li2*t*−*k*2 + *εt*. Here the numbers *k*1 and *k*2 denote the lag orders of the leading indicators. Estimate four versions of this model on the estimation sample from 1951 to 2010, by setting *k*1 and *k*2 equal to either 1 or 2. Show that the model with *k*1 = *k*2 = 1 gives the largest value for R2, and present the four coefficients of this model in six decimals.
4. Perform the Breusch-Godfrey test for first-order residual serial correlation for the model in part (e) with *k*1 = *k*2 = 1. Does the test outcome signal misspecification of the model?
5. Use the model in part (e) with *k*1 = *k*2 = 1 to generate a set of twenty one-step-ahead predictions for the growth rates in each quarter of the period 2011 to 2015. Note that the required values of the lagged leading indicators are available for each of these forecasts. Calculate the root mean squared error of these forecasts and present a time series graph of the predictions and the actual growth rates.

Answer:  
  
**(a)** Perform three likelihood ratio tests in Python code:  
  
\* Refer to the definitions of log-likelihood functio and likelihood ratio (slide 4 & 10) in 5.3.  
  
 Results: Logit of full model y ~ (const, li1, li2)

=================================================================

Model: Logit Pseudo R-squared: 0.122

Dependent Variable: GDPIMPR AIC: 274.3565

Date: 2017-12-16 23:19 BIC: 284.7984

No. Observations: 240 Log-Likelihood: -134.18

Df Model: 2 LL-Null: -152.76

Df Residuals: 237 LLR p-value: 8.4833e-09

Converged: 1.0000 Scale: 1.0000

No. Iterations: 5.0000

--------------------------------------------------------------------------------------------------------

Coef. Std.Err. z P>|z| [0.025 0.975]

---------------------------------------------------------------------------------------------------------

const 0.7288 0.1536 4.7454 0.0000 0.4278 1.0298

li1(-1) -0.3719 0.0727 -5.1176 0.0000 -0.5143 -0.2294

li2(-1) -0.1203 0.0377 -3.1936 0.0014 -0.1941 -0.0465

=================================================================

Considering 4 models (R-stylle):  
  
 1) y ~ const  
 2) y ~ const + li1(-1)  
 3) y ~ const + li2(-1)  
 4) y ~ const + li1(-1) + li2(-1)  
  
where y is GDPIMPR.  
  
We can calculate the loglikelihood function values: [-152.763, -139.747, -149.520, -134.178]

Model 2) LR = -2\*(-149.520-(-134.178)) = 30.684  
Model 3) LR = -2\*(-139.747-(-134.178)) = 11.138  
Model 4) LR = -2\*(-152.763-(-134.178)) = 37.170

Corresponding Likelihood Ratio P-Values: [Const+li1=0.000000, Const+li2=0.010874, Const+li1+li2=0.000000]  
  
Since LR > 6, the p-values of model 2) and 4) are small for 1% level, and p-value of model 3) is small for 5% level, we can reject the hypothesis that li1(-1) and li2(-1) are not significant. In other words, the individual and joint of 1-quarter lags of li1 and li2 are significant.

**(b)** McFaddens R^2 = 1 - LL\_Model/LL\_Intercept.  
  
Let’s consider 4 models (R-stylle):  
  
 1) y ~ const + li1(-1) + li2(-1)  
 2) y ~ const + li1(-1) + li2(-2)  
 3) y ~ const + li1(-2) + li2(-1)  
 4) y ~ const + li1(-2) + li2(-2)  
  
Log Likelihoods:

LL\_null=-152.7634, LL\_c11=-134.1782, LL\_c12=-134.1264, LL\_c21=-130.3455, LL\_c22=-130.4611

McFaddens R2:

0.1217, 0.1220, **0.1467**, 0.1460  
  
As shown in the above R2 values, we can see that model 3), that is, model const + li1(-2) + li2(-1) is optimal.

**(c)** Prediction-realization table and hit rate, using a cut-off value of 0.5.

Results: Logit

=================================================================

Model: Logit Pseudo R-squared: 0.147

Dependent Variable: GDPIMPR AIC: 266.6909

Date: 2017-12-17 01:42 BIC: 277.1328

No. Observations: 240 Log-Likelihood: -130.35

Df Model: 2 LL-Null: -152.76

Df Residuals: 237 LLR p-value: 1.8366e-10

Converged: 1.0000 Scale: 1.0000

No. Iterations: 6.0000

-------------------------------------------------------------------------------------------------------

Coef. Std.Err. z P>|z| [0.025 0.975]

-------------------------------------------------------------------------------------------------------

const 0.7457 0.1573 4.7397 0.0000 0.4373 1.0540

li1(-2) -0.4287 0.0763 -5.6175 0.0000 -0.5783 -0.2791

li2(-1) -0.1312 0.0386 -3.3994 0.0007 -0.2068 -0.0556

=================================================================

Prediction Table: [[ 5 2]

[ 3 10]]

Thus, the Prediction Realization Table: [[ 0.25 0.1]

[ 0.15 0.5]]

Hit Rate: 0.25 + 0.5 = 0.75  
The probability of growth = 0.1 + 0.5 = 0.6  
  
The hit rate is 75%, and the probability of growth is 60%. So the economic growth is very likely.

**(d)** ADF Test for Log GDP.

Statistic: -2.51821089277 > -3.5 (the critical value for ADF with trend)  
p-value: 0.318918118591

Confidence Levels: {'1%': -3.9973, '5%': -3.4291, '10%': -3.1380}

Results: Ordinary least squares

===================================================================

Model: OLS Adj. R-squared: 0.393

Dependent Variable: y AIC: -1888.3395

Date: 2017-12-17 02:37 BIC: -1874.4504

No. Observations: 238 Log-Likelihood: 948.17

Df Model: 3 F-statistic: 52.20

Df Residuals: 234 Prob (F-statistic): 7.23e-26

R-squared: 0.401 Scale: 2.0630e-05

---------------------------------------------------------------------------------------------------------------

Coef. Std.Err. t P>|t| [0.025 0.975]

---------------------------------------------------------------------------------------------------------------

**x1** -0.0204 0.0081 -2.5182 0.0125 -0.0364 -0.0044

x2 0.6325 0.0509 12.4338 0.0000 0.5323 0.7328

const 0.0956 0.0375 2.5512 0.0114 0.0218 0.1695

x3 0.0001 0.0000 2.4979 0.0132 0.0000 0.0001

----------------------------------------------------------------------------------------------------------------

Omnibus: 23.009 Durbin-Watson: 2.012

Prob(Omnibus): 0.000 Jarque-Bera (JB): 36.479

Skew: 0.578 Prob(JB): 0.000

Kurtosis: 4.530 Condition No.: 23965

===================================================================

\* The condition number is large (2e+04). This might indicate

strong multicollinearity or other numerical problems.

Note: *[x1, x2, const, x3] = ['LOG GDP lag(-1)', 'Diff LOG GDP lag(-1)', 'Constant', 'Trend']*

From the above regression test results, we can see that p-value = 0.0125 for LOGGDP lag(-1), so it’s significant for 5% level.  
  
From the ADF test, p-value = 0.3189, and the statistic = -2.518 > -3.5 (the critical value for ADF with trend).  
We can’t reject the hypothesis that H0: LOGGDP is non-stationary.  
So LOGGDP is not stationary.  
  
  
**(e)** Estimate 4 version of the following mmodels.  
  
Model (R-stylle):  
Growthrate(t) ~ const + GrowthRate(t-1) + li1(t-k1) + li2(t-k2)  
 k1, k2 in {1,2}

R Squares: [mod11=**0.507975**, mod12=0.477193, mod21=0.507665, mod22=0.477130]

Model mod11(k1=k2=1) has the largest R^2 value. Mod21 (k1=2, k2=1) has second largest R^2 value.

The coefficients of mod11:

const 0.001737

GrowthRate(-1) 0.461579

li1(-1) -0.001023

li2(-1) -0.000149  
  
**(f)** BG test for serial correlation  
  
Breusch Godfrey Test for model mod11 (k1=k2=1) on serial autocorrelation

lm(Lagrange multiplier test statistic) = 0.230366

lmpval(p-value for Lagrange multiplier test) = 0.631253

fval(fstatistic for F test) = 0.225783

fpval(pvalue for F test) = 0.635110  
  
Since the p-value lmpval is 0.63 is very high, we can not rejec the H0: first order residual is not correlated.  
So no serial autocorrelation in 1-order residual.  
  
**(g)** Forecast for the Growth Rate.

Model (R-style):  
Growthrate ~ const + GrowthRate(-1) + li1(-1) + li2(-1)

RMSE=0.003156   
  
As show below, the prediction model fits the actual values well.  
  
