Econometrics: Methods and Applications

Erasmus University Rotterdam

Peer-graded Assignment: Case project

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: $\text{GrowthRate}_t = \log(\text{GDP}_t) - \log(\text{GDP}_{t-1})$
- **`li1`**: First leading indicator
- **`li2`**: Second leading indicator
- **`T`**: Linear trend (where the first observation, for 1950 quarter 1, is defined as 0).

Visualizing the data:

Content

Textbook

Christiaan Heij, Paul de Boer, Philip Hans Franses, Teun Kloek, Herman K. van Dijk (2004): “Econometric Methods with Applications in Business and Economics”, Oxford University Press

- **GDP**: Gross Domestic Product of the economy

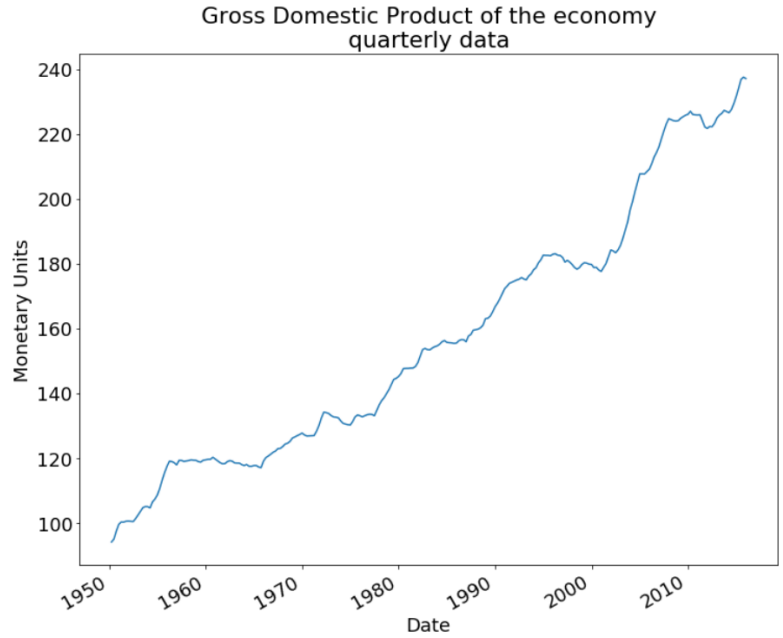

Figure 1 GDP

- **LOGGDP**: Log of Gross Domestic Product
- **GDPIMPR**: dummy variable indicating whether the GDP has increased (1) or decreased (0)

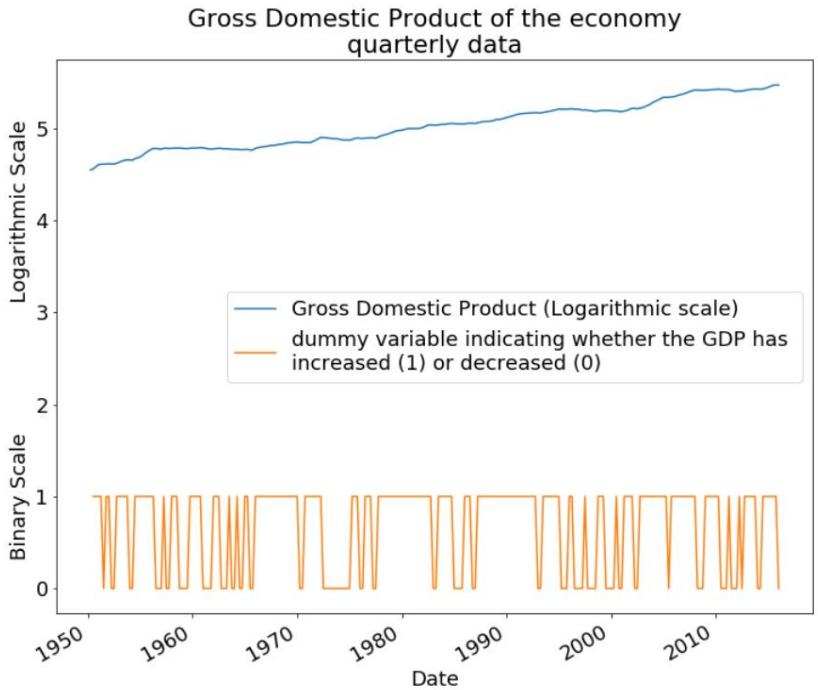

Figure 2 LOGGDP and GDPIMPR

- **`GrowthRate`**: Relative growth of the economy:

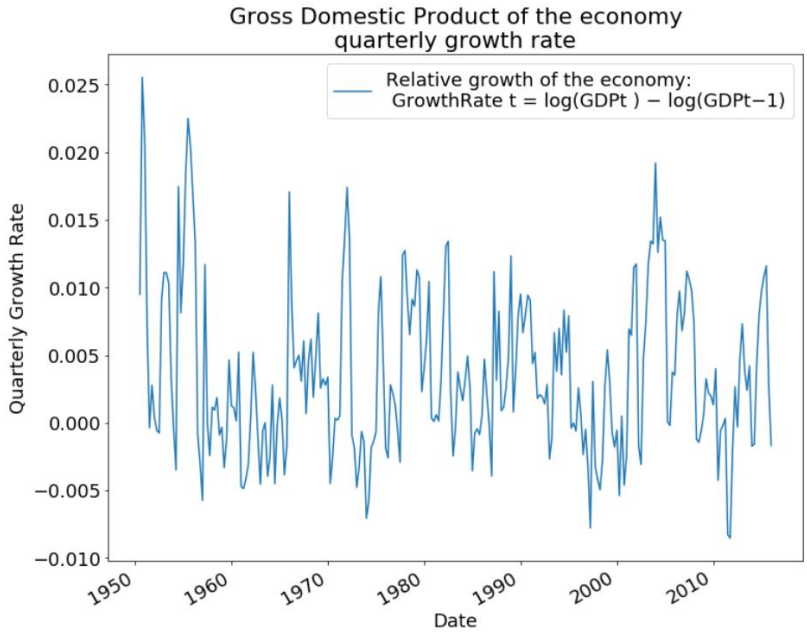

Figure 3 GrowthRate

- **`li1`**: First leading indicator
- **`li2`**: Second leading indicator

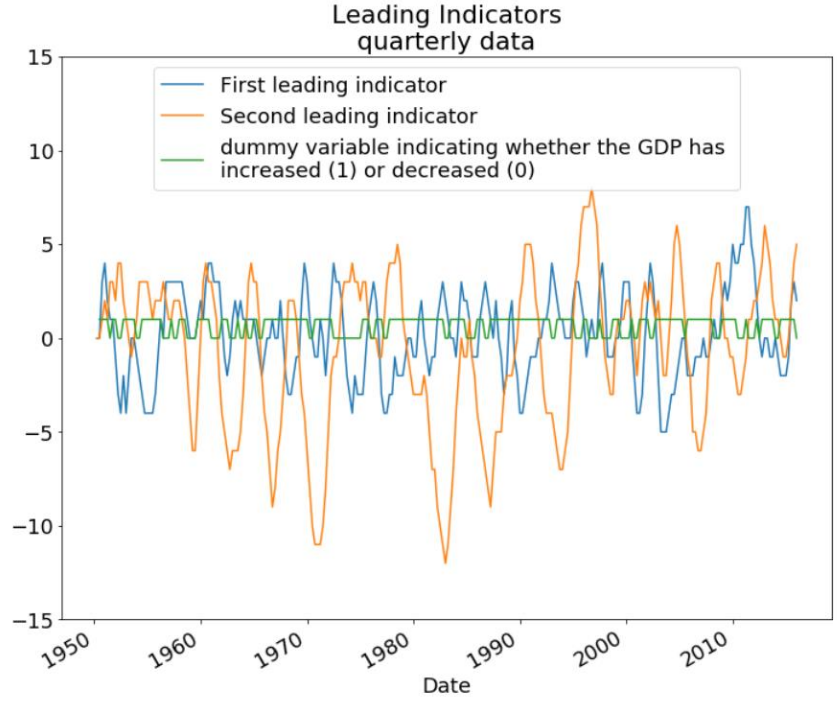

Figure 4 Leading Indicators li1 and li2

Question a) Likelihood Ratio Tests

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

Calculation:

The logit models are non-linear, and the parameters can be estimated by maximum likelihood (Textbook p.447).

Likelihood Ratio test of the 1-quarter lags of **li1** and **li2**

```
                         Results: Logit
================================================================
Model:              Logit            Pseudo R-squared: 0.122
Dependent Variable: GDPIMPR          AIC:              274.3565
Date:               2023-01-06 09:56 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
================================================================
```

Goodness of fit

The significance of individual explanatory variables can be tested by the usual t-test based on (6.10). The sample size should be sufficiently large to rely on the asymptotic expressions for the standard errors, and the t-test statistic then follows approximately the standard normal distribution. (Textbook p.453)

individual significance

$\boldsymbol{li1_{t-1}}$ p-value = 0.0000 → significant 99%-confidence level

$\boldsymbol{li2_{t-1}}$ p-value = 0.0014 → significant 99%-confidence level

joint significance

The overall goodness of fit of the model can be tested by the LR-test on the null hypothesis that all coefficients (except the constant term) are zero—that is,

$$\beta_2 = \ldots = \beta_k = 0$$

This test follows (asymptotically) the $\chi^2(k - 1)$ distribution (Textbook p.453).

Likelihood Ratio tests	p-values	Result
$\boldsymbol{Constant + li1_{t-1}}$	0.0000	significant 99%
$\boldsymbol{Constant + li2_{t-1}}$	0.01087	significant 95%
$\boldsymbol{Constant + li1_{t-1} + li1_{t-1}}$	0.0000	significant 99%

Answer (a)

The LM-Test shows that the two models

- $\boldsymbol{Constant + li1_{t-1}}$
- $\boldsymbol{Constant + li1_{t-1} + li1_{t-1}}$

are significant at the 1%-Level.

The $\boldsymbol{Constant + li2_{t-1}}$ model is significant at the 5%-Level.

Question b) McFadden R^2

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 R^2 to select the best lag structure among these four models? Compute the four values of McFadden R^2 (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

Calculations:

Sometimes one reports measures similar to the R^2 of linear regression models—for instance, McFadden's R^2 defined by

$$R^2 = 1 - \frac{\log(L_1)}{\log(L_0)},$$

,

where L1 is the maximum value of the unrestricted likelihood function and L_0 that of the restricted likelihood function. It follows from (6.7) that $L_0 \leq L_1 < 0$, so that $0 \leq R^2 < 1$ and higher values of R^2 correspond to a relatively higher overall significance of the model. Note, however, that this R^2 cannot be used, for example, to choose between a logit and a probit model, as these two models have different likelihood functions. (Textbook p.453)

logit model	coeff	LL (Loglikelihood)	McFadden's R^2
Model 0	$\boldsymbol{Constant}$	-152.763	
Model 1	$\boldsymbol{Constant + li1_{t-1} + li1_{t-1}}$	-134.178	0.12166
Model 2	$\boldsymbol{Constant + li1_{t-1} + li1_{t-2}}$	-134.126	0.1220
Model 3	$\boldsymbol{Constant + li1_{t-2} + li1_{t-1}}$	-130.346	0.1467
Model 4	$\boldsymbol{Constant + li1_{t-2} + li1_{t-2}}$	-130.461	0.1460

Answer (b)

The model optimal according to the criterion McFadden's R^2 is model 3

$\hat{y}_t$ = Constant + $li1_{t-2}$ + $li1_{t-1}$

Question c) Prediction-realization table and the hit rate

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.

Calculation;

```
                         Results: Logit
================================================================
Model:              Logit            Pseudo R-squared: 0.147
Dependent Variable: GDPIMPR          AIC:              266.6909
Date:               2023-01-06 10:43 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
================================================================
```

Parameters of the Logit model from the training set n=240

β_0 = 0.7457 β_1 = -0.4287 β_2 = -0.1312

Predicted probability of economic growth

$$Pr_t = \frac{e^{(\beta_0+\beta_1 li1(-2)+ \beta_2\ li2(-1)}}{1+ e^{\ (\beta_0+\beta_1 li1(-2) + \beta_2\ Li2(-1))}}$$

$$\Leftrightarrow\ Pr_t = \frac{e^{(0.7457-0.4287\ li1(-2)-\ 0.1312\ li2(-1)}}{1+ e^{\ (0.7457-0.4287\ li1(-2)-\ 0.1312\ Li2(-1))}}$$

Calculate the predicted probability of economic growth for each of the 20 quarters of the evaluation sample:

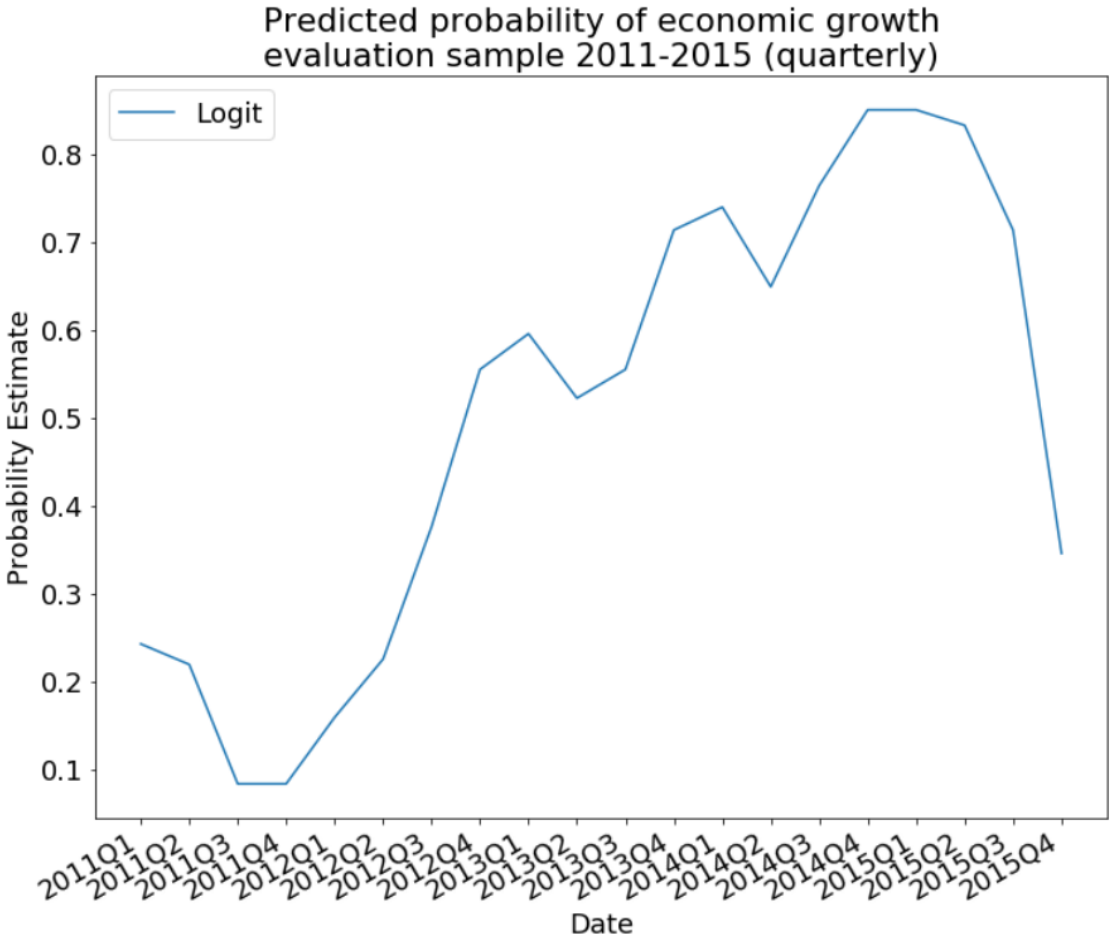

Figure 5 Predicted Probability LOGIT model

rule: $Pr_t \geq 0.5$ ➔ $\hat{y} = 1$ cut-off value of 0.5

The model predicts growth in 12 quarters of the evaluation sample of 20 quarters (out-of-sample forecast $y_{hat} = 1$). Let us now compare the forecast with the actual economic growth in those 20 quarters. The dummy variable GDPIMPR indicates whether the GDP has

increased (1) or decreased (0). We have seen actual growth in 13 quarters.

Prediction-Realization Table	$y_{hat} = 1$ (8/20)	$y_{hat} = 0$ (12/20)
GDPIMPR =1 (13/20)	10	3
GDPIMPR =0 (7/20)	2	5

The hit rate is defined as the fraction of correct predictions in the sample. Formally, let w_i be the random variable indicating a correct prediction—that is, $w_i = 1$ if $y_i = \hat{y}_i$ and, $w_i = 0$ if $y_i \neq \hat{y}_i$; then the hit rate is defined by

$$h = \frac{1}{n}\sum_{i=1}^{n} w_i$$

In the population the fraction of successes is p. If we randomly make the prediction 1 with probability p and 0 with probability (1- p), then we make a correct prediction with probability q = p² + (1 - p) ². (Textbook p.453)

$$h = \frac{1}{20}\sum_{i=1}^{20} w_i$$

$$\Leftrightarrow h = \frac{1}{20}\,15$$

$$\Leftrightarrow h = \frac{3}{4}$$

Answer (c)

The out-of-sample hit ratio of the model is 0.75 The model was correct 75% of the time. It has correctly predicted 10 quarters with economic growth, and 5 quarters with economic contraction (i.e. no growth). However, the model has missed 3 quarters of economic expansion, and it has missed 2 quarters of contraction.

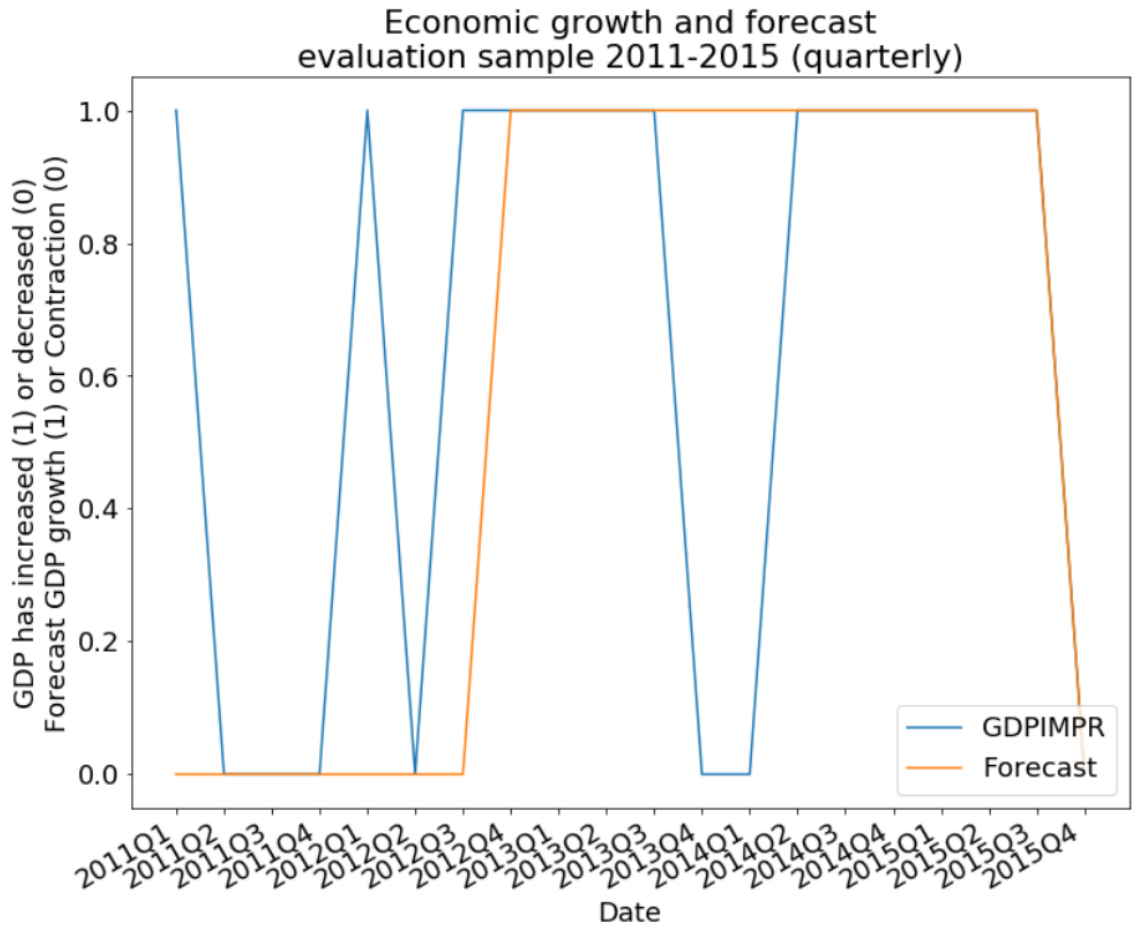

Figure 6 Binary Forecast LOGIT model

Looking at the graph we find that the model forecast fluctuates much less than the actual economic growth. The model forecast based on the two leading indicators is rather sluggish, and it tends to keep its direction for several quarters before it changes direction again.

Question d) Augmented Dickey-Fuller test (Stationarity)

> 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.

Calculations:

Augmented Dickey-Fuller (ADF) test for y_t

$$\text{ADF: } \Delta y_t = \alpha + \beta t + \rho y_{t-1} + \sum_{j=1}^{3} \gamma_j \Delta y_{t-j} + \varepsilon_t$$

A time series is said to be “stationary” if it has no trend, exhibits constant variance over time, and has a constant autocorrelation structure over time.

One way to test whether a time series is stationary is to perform an augmented Dickey-Fuller test, which uses the following null and alternative hypotheses:

- H_0: The time series is non-stationary.
- H_A: The time series is stationary.

- Test with deterministic trend if data clear trend direction:

  $$\Delta y_t = \alpha + \beta t + \rho y_{t-1} + \gamma_1 \Delta y_{t-1} + \ldots + \gamma_L \Delta y_{t-L} + \varepsilon_t$$

  Reject H_0 of non-stationarity if $t_p < 3.5$

Include constant, trend, and a single lag in the test equation:

$$\Delta y_t = \alpha + \beta t + \rho y_{t-1} + \gamma_1 \Delta y_{t-1}$$

Results: Ordinary least squares

Model:	OLS	Adj. R-squared:	0.393
Dependent Variable:	y	AIC:	-1888.3395
Date:	2023-01-06 10:47	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]
LOG GDP lag1	-0.0204	0.0081	-2.5182	0.0125	-0.0364	-0.0044
Diff LOG GDP lag1	0.6325	0.0509	12.4338	0.0000	0.5323	0.7328
Constant	0.0956	0.0375	2.5512	0.0114	0.0218	0.1695
Trend	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.

coefficient of $\log(y_{t-1})$ = 0.0204 std error: 0.0081
t-value: -2.5182
p-value: 0.0125

Rule:
Reject H_0 of non-stationarity if $t_p < 3.5$

$t_p = -2.5182$ for $\log(y_{t-1}$ I) → not stationary

Answer (d)
We do reject H_0 and conclude that the time series y_t is non-stationary. In other words, it has some time-dependent structure and does not have constant variance over time.

Question e) Model selection (AR(1) and AR(2))

Consider the following model:

$$GrowthRate_t = \alpha + \rho GrowthRate_{t-1} + \beta_1 li1_{t-k_1} + \beta_2 li2_{t-k_2} + \varepsilon_t .$$

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 R^2 and present the four coefficients of this model in six decimals.

Calculations:

estimation sample from 1951 to 2010 (n=240)	k_1	k_2
Model 1	1	1
Model 2	2	1
Model 3	1	2
Model 4	2	2

Model 1 with $k_1=k_2= 1$

Dep. Variable:	GrowthRate	**R-squared:**	0.508
Model:	OLS	**Adj. R-squared:**	0.502
Method:	Least Squares	**F-statistic:**	81.22
Date:	Sat, 07 Jan 2023	**Prob (F-statistic):**	3.97e-36
Time:	10:01:48	**Log-Likelihood:**	980.34
No. Observations:	240	**AIC:**	-1953.
Df Residuals:	236	**BIC:**	-1939.
Df Model:	3		
Covariance Type:	nonrobust		

Present the four coefficients of the model with $k_1=k_2= 1$ in six decimals:

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0017	0.000	5.433	0.000	0.001	0.002
GrowthRate(-1)	0.4616	0.048	9.556	0.000	0.366	0.557
li1(-1)	-0.0010	0.000	-7.880	0.000	-0.001	-0.001
li2(-1)	-0.0001	6.42e-05	-2.326	0.021	-0.000	-2.29e-05

Omnibus:	23.834	**Durbin-Watson:**	2.039
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	34.778
Skew:	0.628	**Prob(JB):**	2.81e-08
Kurtosis:	4.379	**Cond. No.**	786.

Answer (e)

estimation sample from 1951 to 2010 (n=240)	k_1	k_2	R^2
Model 1	1	1	0.507975
Model 2	2	1	0.477193
Model 3	1	2	0.507665
Model 4	2	2	0.477130

Model 1 with $k_1=k_2= 1$ gives the largest value for R^2 .

Model 1 with $k_1=k_2= 1$

Coefficients	$GrowthRate_t = \alpha + \rho GrowthRate_{t-1} + \beta_1 li1_{t-1} + \beta_2 li2_{t-1} + \varepsilon_t$.
α	0.001737
ρ	0.461579
β_1	-0.001023
β_2	-0.000149

Question f) Breusch-Godfrey test (Serial correlation)

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?

Calculation

Breusch–Godfrey test for serial correlation of order *p*

- *Step 1: Apply OLS.* Apply OLS in the model $y = X\beta + \varepsilon$ and compute the residuals $e = y - Xb$.
- *Step 2: Perform auxiliary regression.* Apply OLS in the auxiliary regression equation

$$e_i = x_i'\delta + \gamma_1 e_{i-1} + \cdots + \gamma_p e_{i-p} + \omega_i, \quad i = p+1, \cdots, n.$$

- *Step 3: $LM = nR^2$ of the regression in step 2.* Then $LM = nR^2$ where R^2 is the coefficient of determination of the auxiliary regression in step 2. This is asymptotically distributed as $\chi^2(p)$ under the null hypothesis of no serial correlation, that is, if $\gamma_1 = \cdots = \gamma_p = 0$.

(Textbook p.364)

One of the key assumptions in linear regression is that there is no correlation between the residuals, e.g., the residuals are independent. To test for first-order autocorrelation, we can perform a Durbin-Watson test. However, if we'd like to test for autocorrelation at higher orders then we need to perform a **Breusch-Godfrey** test.

This test uses the following hypotheses:

- H_0 (null hypothesis): There is no autocorrelation at any order less than or equal to p.

- H_A (alternative hypothesis): There exists autocorrelation at some order less than or equal to p.

The test statistic follows a Chi-Square distribution χ^2 with p degrees of freedom. If the p-value that corresponds to this test statistic is less than a certain significance level (e.g., 0.05) then we can reject the null hypothesis and conclude that autocorrelation exists among the residuals at some order less than or equal to p.
(Source: https://www.statology.org/breusch-godfrey-test-python/)

Breusch–Godfrey test	$GrowthRate_t = \alpha + \rho GrowthRate_{t-1} + \beta_1 li1_{t-1} + \beta_2 li2_{t-1} + \varepsilon_t$.
Test statistic X^2	0.2304
P-value	0.6313

Answer (f)
We cannot reject the H_0 (null hypothesis) and conclude that there is no autocorrelation at any order less than or equal to 1, because the high P-value of 0.6613 is not significant at a 5%-Level (i.e. 0.6613 > 0.05). However, such a high P-value might signal misspecification of the model.

Question g) Forecasting with AR(1) model

(a) 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 (g)

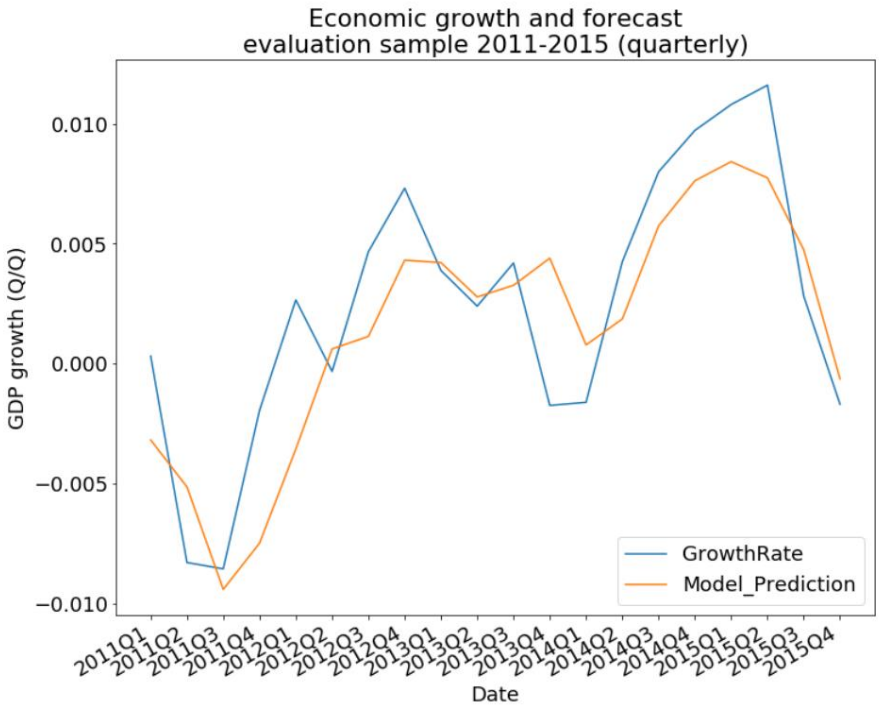

Figure 7 Model Prediction (Out-of-Sample)

Another useful method for model selection is to compare the predictive performance of the models. For this purpose the data set is split in two parts, an 'estimation sample' (used to construct the model) and a 'prediction sample' or 'hold-out sample' for predictive evaluation. So the models are estimated using only the data in the first subsample, and the estimated models are then used to predict the y-values in the prediction sample. Possible evaluation criteria are the root mean squared error (RMSE) and the mean absolute error (MAE). These are defined by

$$\text{RMSE} = \left(\frac{1}{n_f} \sum_{i=1}^{n_f} (y_i - \hat{y}_i)^2 \right)^{1/2}$$

$$\text{MAE} = \frac{1}{n_f} \sum_{i=1}^{n_f} |y_i - \hat{y}_i|,$$

where n_f denotes the number of observations in the prediction sample and $\hat{y}_i$ denotes the predicted values. (Textbook p.280)

Out-of-Sample Predictions	$GrowthRate_t = \alpha + \rho GrowthRate_{t-1} + \beta_1 li1_{t-1} + \beta_2 li2_{t-1} + \varepsilon_t$.
root mean squared error (RMSE)	0.003156
mean absolute error (MAE)	0.052807