# Assignment 1 - Econometrics 2

Department of Economics - University of Copenhagen

Bilstein, Moritz
wmj863@alumni.ku.dk

Hochuli, Raul
ckv730@alumni.ku.dk

Stöckli, Patrick
xqp997@alumni.ku.dk

Sunday 04.10.2020, 23:00

# 1 Introduction

**Main Question** For our assignment we have been given US GDP data on a quarterly basis. We intend to provide an estimate, using a univariate time series model to produce a forecast for the recovery and later development of 2009(3)-2019(4).

**Motivation** In previous economic courses the handled data was primarily considered to be independent and identically distributed (iid). Auto-regressive models relax this assumption to some extent. This is especially interesting as it allows us to work with non-iid data. Furthermore auto-regressive models require no additional observations or explanatory variables. This makes them easy to compute and valuable.

**Econometric model and conclusion** Using an AR(2) model we can estimate a very close fit to the actual GDP observations and also a decent forecast to predict the near future development of US GDP. For mid- or far future developments more sophisticated version such as a vector auto regressive model might be worth considering.

# 2 Description of data

The data set provided for this assignment contains quarterly data for the real US GDP form the first quarter of 1975 to the fourth quarter of 2019. The data has been downloaded as the series GDPC1 from the FRED database maintained by the Federal Reserve Bank of St. Louis[1]. Further there are the following transformations in the data set:

- logGDP = log(GDP$_t$)
- D4logGDP = $\Delta_t$log(GDP$_t$)=log(GDP$_t$)-log(GDP$_{t-4}$)
- LinFor: a conservative forecast of the recovery, assuming that the recovery will be a linear adjustment towards the long-run growth potential of 3% after 2009(2)

In Figure 1 we show that the transformation to D4logGDP is clearly beneficial for our analysis. The time series now looks stationary to an acceptable extent. Despite having an overall stationary appearance, we notice shifts in the GDP which might also imply level changes. Further it is clearly visible that certain periods include considerable outliers, most prominent amongst them would be the subprime crisis (2008), the second oil crisis (1979) and the US-Recession of 1981 and 1982. These have to be kept in mind when estimating our model as they might be a source of potential bias or miss specification.

# 3 Economic theory

**General model** As mentioned in the introduction, uni-variate models are particularly useful for characterizing dependence and for computing simple forecasts. Given the characteristics of the data we select an AR-model. Thereby our model is given with AR(p):

$$y_t = \delta + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \theta_3 y_{t-3+...} + \theta_p y_{t-p} \epsilon_t \tag{1}$$

where our residuals $\epsilon \sim iid(0, \sigma^2)$, and we can observe the $y_{-(p-1)}, y_{-(p-2)}, ..., y_0$ initial values. Using the lag-operator $L$ which has the following properties $Ly_t = y_{t-1}$, we can reformulate the AR(p) model,

[1] https://fred.stlouisfed.org

so that all the $y$ terms can be combined with a single coefficient. Taking equation 1 we can reformulate all the terms containing $y_t$ to:

$$\begin{aligned}(1-\theta_1 L-\theta_2 L^2-...-\theta_p L^p)y_t) &= \delta+\epsilon_t \\ \theta(L)y_t &= \delta+\epsilon_t\end{aligned} \tag{2}$$

Given $|\theta| < 1$ the inverse of the polynomial $\theta(L)$ can be defined as $\theta^{-1}(L)$. Because the polynomial can also have complex solutions, we use the characteristic equation:

$$\theta(z) = 1-\theta_1 z-\theta_2 z^2-...-\theta_p z^p = 0$$

to then factorize the polynomial to obtain the inverse roots: $\phi_1 = z_1^{-1}, \phi_2 = z_2^{-2}, ..., \phi_p = z_p^{-p}$. $\theta(z)$ is now invertible if each of its factors is invertible, meaning if $|\phi_j| < 1$ for all $j = 1, ..., p$. The mentioned inverse polynomial is written as:

$$\theta(z)^{-1} = 1+c_1 z+c_2 z^2+c_3 z^3,+...+,c_p z^p$$

As $\theta(L)$ is invertible under the mentioned conditions, we take the results of equation (2) and reformulate it to:

$$\begin{aligned}y_t &= \theta(L)^{-1}(\delta+\epsilon_t) \\ &= (1+c_1 L+c_2 L^2+...c_p L^p)(\delta+\epsilon_t) \\ &= (1+c_1+c_2+...+c_p)\delta+\epsilon_t+c_1\epsilon_{t-1}+c_2\epsilon_{t-2}+...+c_p\epsilon_{t-p} \\ &\implies \frac{\partial y_t}{\partial \epsilon_t} = 1, \quad \frac{\partial y_t}{\partial \epsilon_{t-1}} = c_1, \quad \frac{\partial y_t}{\partial \epsilon_{t-2}} = c_2, ...\end{aligned} \tag{3}$$

In the last line of equation (3) we see that $c_j$ measures the dynamic impact of a shock on $y_t$ and is therefore also called impulse responses.

### Model assumptions

- **Stationarity**: AR(p) is stationary if $\theta(z)$ is invertible, or put differently if $|\phi_j| < 1$ for $j = 1, 2, ..., p$.
- **Weak dependence**: The stationarity condition implies that the impulse response will decrease over time and eventually die out. Looking at Figure 1, we clearly see that the ACF is decreasing and eventually converging to 0. This observation strengthens the assumption of weak dependence, with decreasing autocorrelation with increasing lags.

**Our model** Our final model is a AR(2) model with two lags and a control vector $\overrightarrow{\gamma}$ which includes dummy variables to capture outliers with extremely large residuals. Those events were mentioned in section two. Excluding this quarters helps our model to receive approximately normally distributed residuals, a vital condition for the consistency of the model and the estimator. Further we include a dummy variable (”Regime”) in our control vector which indicates the period 2001(1)-2009(3). As previously mentioned in the data description, we suspect a level shift in the the time period after 2001(1) and it’s recession. Adjusting for this, we also hope to improve the approximative normality of our model residuals.

$$\begin{aligned}y_t &= \delta+\theta_1 y_{t-1}+\theta_4 y_{t-4} \\ &+\gamma_1\mathbf{1}_{1979(1)}+\gamma_2\mathbf{1}_{1981(4)}+\gamma_3\mathbf{1}_{1982(1)}+\gamma_4\mathbf{1}_{2008(4)}+\gamma_5\mathbf{1}_{Regime}+\epsilon_t\end{aligned} \tag{4}$$

Or in vector notation:

$$y_t = \overrightarrow{\theta}\overrightarrow{x}_t+\overrightarrow{\gamma}\overrightarrow{D}+\epsilon_t, \tag{5}$$

with the vectors:

$$\begin{aligned}&\overrightarrow{\theta} = (\delta, \theta_1, \theta_4), \overrightarrow{x}_t = (1, y_{t-1}, y_{t-4})', \overrightarrow{\gamma} = (\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5),\\ &\overrightarrow{D} = (\mathbf{1}_{1979(1)}, \mathbf{1}_{1981(4)}, \mathbf{1}_{1982(1)}, \mathbf{1}_{2008(4)}, \mathbf{1}_{Regime})'.\end{aligned} \tag{6}$$

Or in matrix notation:

$$\overrightarrow{Y} = X\beta + \overrightarrow{\epsilon} \tag{7}$$

with the vectors and matrices:

$$X = \begin{pmatrix} \overrightarrow{x}'_5 \overrightarrow{D}' \\ \overrightarrow{x}'_6 \overrightarrow{D}' \\ \vdots \\ \overrightarrow{x}'_T \overrightarrow{D}' \end{pmatrix} \text{,a } (T-4) \times 8 \text{ matrix,} \tag{8}$$

$$\begin{aligned}\beta &= (\overrightarrow{\theta}\,\overrightarrow{\gamma})', \text{ an } 8 \times 1 \text{ vector,}\\ \overrightarrow{\epsilon} &= (\epsilon_5, \ldots ,\epsilon_T)',\\ \overrightarrow{Y} &= (y_5, \ldots ,_T)'.\end{aligned} \tag{9}$$

$T$ denotes the length of the time series D4logGDP. As we have a maximum lag of four, we lose the first four observations in our estimation.

**The estimator and its properties** We use the OLS estimator $\hat{\beta}$, which is also the maximum likelihood estimator:

$$\hat{\beta} = (X'X)^{-1}X'\overrightarrow{Y} \tag{10}$$

Under certain conditions the OLS estimator is both consistent and unbiased:

$$\hat{\beta} = (X'X)^{-1}X'(X\beta + \overrightarrow{\epsilon}) = \beta + (X'X)^{-1}X'\overrightarrow{\epsilon}. \tag{11}$$

For unbiasedness the mean of the residuals has to be zero:

$$E(\hat{\beta}|X) = \beta + (X'X)^{-1}X'E(\overrightarrow{\epsilon}|X) = \beta + \overrightarrow{0} \tag{12}$$

With the law of iterated expectations also the unconditional expectation of the estimator is the true coefficient of the linear model:

$$E(\hat{\beta}) = E(E(\hat{\beta}|X)) = \beta \tag{13}$$

For consistency we need the residuals to be independent from the explanatory variables and we need the law of large numbers to apply. It applies if $y_t$ has a well defined mean and if the observations are drawn independently and identically.

$$\hat{\beta} = \beta + (\tfrac{X'X}{T-4})^{-1}\tfrac{X'\overrightarrow{\epsilon}}{T-4} \tag{14}$$

$$\lim_{T\to\infty}(\hat{\beta}) = \beta + E(X'X)^{-1} \begin{pmatrix} E(\epsilon_t) \\ E(\epsilon_t y_{t-1}) \\ E(\epsilon_t y_{t-4}) \\ E(\epsilon_t \mathbf{1}_{1979(1)}) \\ \vdots \\ E(\epsilon_t \mathbf{1}_{Regime}) \end{pmatrix} \tag{15}$$

Just like for unbiasedness we need the conditional mean of the residuals to be 0. With the law of iterated expectations also the unconditional mean is 0. Additionally we need the residuals to be stochastically independent from the explanatory variables. Substituting recursively for $y_{t-1}$ in it's expected value one can see that the residuals must not be correlated.

$$E(\epsilon_t y_{t-1}) = E(\epsilon_t(\delta + \theta_1 y_{t-2} + \theta_4 y_{t-5} + \epsilon_{t-1})) \tag{16}$$

Our estimation will indicate that the residuals in our model are not correlated. In case of the dummy variables, the conditional mean reduces to a non stochastic term and the mean of the residuals. Summarized, we need at least finite second order moments of $y_t$ for the law of large numbers to apply - which is the case if our time series is stationary. We also need the residuals not to be correlated, so that our estimator is consistent. They also have to have a conditional mean, conditioned on $X$, of zero. This is the case if they are independent from the explanatory variables. For the matrix $X'X$ to be invertible there also must not be perfect multicollinearity. That is the case if none of our explanatory variables are perfectly correlated. The correlation between $y_{t-1}$ and $y_{t-4}$ is only 0.406. We also do not have a problem with the dummy variable trap.
For the standard errors to be computed correctly we also need conditional homoscedasticity:

$$\begin{aligned} Var(\hat{\beta}|X) &= E((\hat{\beta} - \beta)\hat{\beta} - \beta'|X) = E((X'X)^{-1}X'\overrightarrow{\epsilon}\overrightarrow{\epsilon}'X(X'X)^{-1}|X) \\ &= \sigma^2 \mathbf{I}_{(T-4\times T-4)}(X'X)^{-1} = \sigma^2(X'X)^{-1} \end{aligned} \tag{17}$$

Homoscedasticity is not strictly necessary as there are standard error estimators that account for heteroscedasticity. In our model the residuals are in fact homoscedastic. As all necessary criteria are fulfilled $\hat{\beta}$ is a consistent estimator with a variance estimated in a correct way. It's asymptotic distribution is normal:

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1}) \tag{18}$$

The variance converges to zero, as $X'X$ grows in $T$.

$$\lim_{x\to\infty}(\sigma^2(X'X)^{-1}) = 0 \tag{19}$$

# 4 Empirical analysis

**Empirical results** Our final forecasting model is displayed in Table 1, the second column. Column 1 and 3 of Table 1 show an intermediate step in our model selection (1) and the comparison to another model group (3), both explained more detailed in the next paragraph.

**Model selection** Looking at the Figure 1 for the ACF and PACF of D4logGDP we see a clear convergence towards zero which indicates weak dependence. The graph shows (nearly) significant PACF terms still within lags of over 10 periods. Following the general-to-specific (GETS) approach we start our model estimation with an AR(13) model.

As the first model is over-specified and contains various insignificant coefficients we use automatic model selection for a fist shrinking, with the results in the first column of Table 1. Seeing that the miss specification tests are in general rejected, this model seems to not fulfill the assumptions of stationarity and weak dependence. Also looking at the inverse roots in Figure 3, we see that they are scattered close to the unit circle. Adjusting for these flaws, we decrease the number of lags (to lower the general size of the inverse roots) and introduce 4 dummies, each canceling out a quarter with a particularly large residual. In each quarter there was an economic shock to the economy of the United States, the second oil crisis in 1979, the recession of 1981-82 and the subprime crisis in 2008. This correction should help improve the distribution of the residuals to be approximately normal distributed.

As already mentioned in the data description, we suspect a level shift in the later half of the observed period. After the first quarter of 2001 the GDP seems to oscillate around a lower level than before. This would violate our assumption of stationarity, as the mean seems to have decreased.

Given this we introduce the dummy regime to account for this shift and reestablish the assumption of stationarity.

Even though we followed the GETS approach, the few significant ACF terms of the time series leave us in doubt about potential MA terms which could further improve the model. Given the first ACF term is the largest, we extend our AR(2) model with an additional MA term to an ARMA(2,1). As the number of AR terms is very low, we are not concerned about potential cancelling roots. At a first glance the MA inclusion seems to be successful as all information criteria show a greater value, and the likelihood score on the other hand increased. However, looking at the miss specification test, we see that all tests have been rejected. Having non-normality in the residuals would transfer the MLE estimator to the QMLE for our ARMA(2,1). However, the combination of heteroscedasticity and autocorrelation make the estimator inconsistent and therefore worse than the simpler AR model.

**Miss specification tests and link to economic theory** Looking at the miss specification tests of our final AR(2) model, we conclude that we consider our estimator to be consistent. The AR model shows no auto correlation, heteroscedasticity and its residuals are considered normal distributed, thereby fulfilling all the necessary criteria for consistency. The inverse roots plotted in Figure 3 are reasonably far away from the unit circle, with all roots having a modulus smaller than 0.88.

**Comparison to linear forecast** We formulate a Diebold-Mariano test to test if our forecast is significantly better than the linear forecast. We use the package "forecast" by R. Hyndman and the included "dm.test"-function[2]. The test reveals, that our forecast seems to be worse for the time span of 42 quarters. We cannot reject that our forecast is less accurate (for both squared errors and absolute value errors) but we can reject that our forecast is more accurate (for absolute value errors, with a significance level of 10%). We think that although our forecast is very accurate for the rest of 2009 and the year 2010 the discrepancy in 2011 is too high for our forecast to be better than the linear forecast. This is emphasized by the sum of squared errors in the first five quarters of forecasting. The sum of squared errors of our forecast is 0.0007071481 and the one of the linear forecast is 0.001280855 and almost twice as large. Until 2011(3) our forecast has a lower errors than the linear forecast.

**Interpretation of the empirical results** The coefficients of our model can be interpreted as following: A one percent increase in the yearly-growth rate in the quarter before increases the yearly-growth rate of the dependent quarter by 0.96%. An one percent increase in the last years yearly-growth rate (D4log(GDP)_4) decreases the yearly-growth rate of the dependent quarter by 0.23%. As expected, our time dummies control for the extraordinary negative shock in the corresponding quarters and the Regime variable takes the level-shift into account. All parameters are highly significant.

# 5 Conclusion

We derive a well thought out linear model to describe the quarterly GDP growth of the USA. It fulfills all requirements for a consistent and precise estimation with the OLS estimator. We compared it to other possible model choices and can conclude that ours is less prone to errors as there is no miss specification and no possible threat of a unit root.

With this model, we are able to compute a forecast that predicts the GDP growth very precise until the end of 2010. Considering forecasts of longer intervalls however, we would advise for a more sophisticated model like a vector auto regression, which is also capable of controlling for other factors.

[2] https://www.rdocumentation.org/packages/forecast/versions/8.13/topics/dm.test

# 6 Appendix

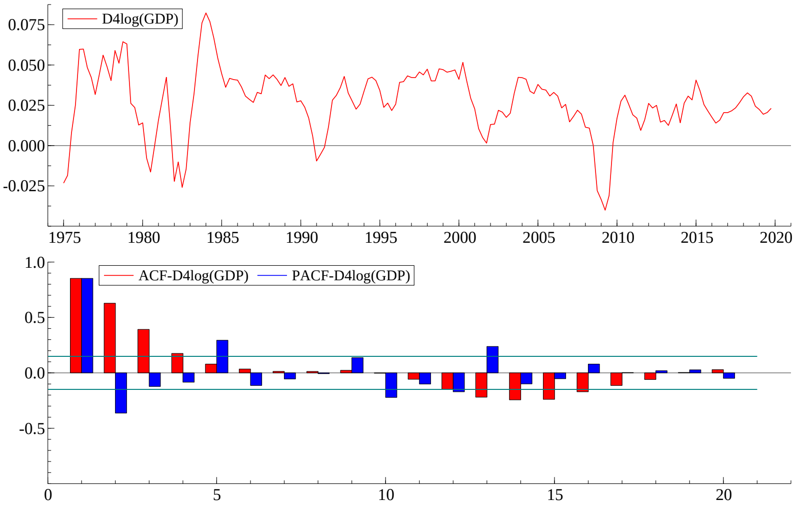

Figure 1: ACF and PACF for the D4logGDP variable

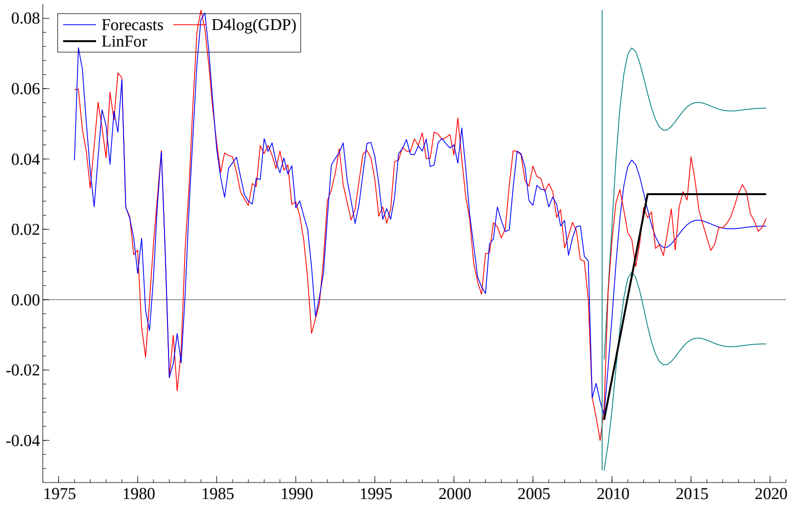

Figure 2: Fitted values of AR(2) model to actual time series, including forecast for period after 2009(3) with 95% confident-bands and the linear forecast

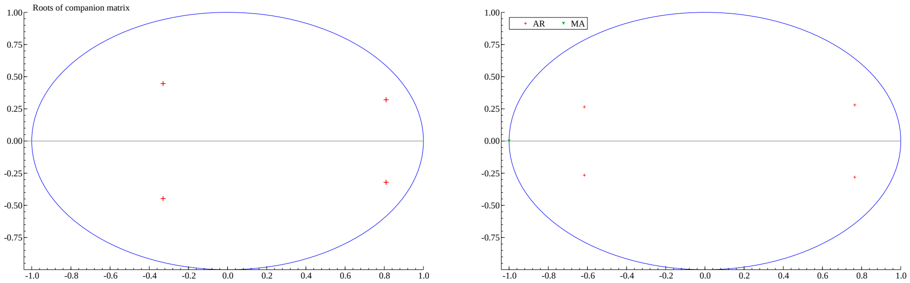

Figure 3: Plot of inverse roots for AR(2) (left) and ARMA(4,1) (right).

| | Automatic Selection | AR | ARMA |
|---|---:|---:|---:|
| Constant | . | 0.01<br>(5) | 0.003<br>(6.49) |
| D4log(GDP)_1 | 1.167<br>(28.8) | 0.960<br>(21.7) | 1.159<br>(55.6) |
| D4log(GDP)_4 | −0.714<br>(−4.65) | −0.234<br>(−5.17) | −0.262<br>(−12) |
| D4log(GDP)_5 | 0.569<br>(3.33) | . | . |
| D4log(GDP)_8 | −0.483<br>(−3.34) | . | . |
| D4log(GDP)_9 | 0.617<br>(4.22) | . | . |
| D4log(GDP)_12 | −0.650<br>(−5.62) | . | . |
| D4log(GDP)_13 | 0.467<br>(5.87) | . | . |
| MA-1 | . | . | −1<br>(−26.3) |
| I:1979(2) | . | −0.031<br>(−21.8) | −0.014<br>(−2.74) |
| I:1981(4) | . | −0.038<br>(−19.5) | −0.019<br>(−1.49) |
| I:1982(1) | . | −0.041<br>(−29.1) | 0.009<br>(0.77) |
| I:2008(4) | . | −0.029<br>(−19.2) | −0.044<br>(−5.35) |
| Regime | . | −0.004<br>(−2.83) | −0.001<br>(−3.18) |
| $\hat{\sigma}$ | 0.00779 | 0.007876 | 0.0118 |
| Log-lik. | 433.105 | 463.075 | 402.368 |
| AIC | -6.818 | -6.792 | -5.856 |
| HQ | -6.753 | -6.722 | -5.768 |
| SC/BIC | -6.659 | -6.619 | -5.640 |
| No autocorr. | [0.05] | [0.24] | [0.00] |
| No hetero. | [0.01] | [0.16] | [0.00] |
| Normality | [0.01] | [0.09] | [0.02] |
| T | 125 | 134 | 134 |
| Sample start | 1978(2) | 1976(1) | 1976(1) |
| Sample end | 2009(2) | 2009(2) | 2009(2) |

Table 1: The table shows estimates of the model in equation (X) with various restrictions imposed. T-ratios in $(\cdot)$ and p-values in $[\cdot]$ for misspecification tests.