ARIMA with Exogenous Variables, Seasonality and Long-Range Dependence

ARMAX

The autoregressive moving average model including exogenous covariates, ARMAX(R,M), extends the ARMA(R,M) model by including the linear effect that one or more exogenous series has on the stationary response series y_t. The general form of the ARMAX(R,M) model is

$$y_t = \sum_{i=1}^{R} \phi_i \, y_{t-i} + \sum_{k=1}^{q} \beta_k \, x_{tk} + \varepsilon_t + \sum_{j=1}^{M} \theta_j \, \epsilon_{t-j}$$

and it has the following condensed form in lag operator notation:

$$\phi(L) y_t = c + x_t' \beta + \theta(L) \varepsilon_t$$

The vector x_t' holds the values of the q exogenous, time-varying predictors at time t, with coefficients denoted β.

ARMAX

ARMAX models have the same stationarity requirements as ARMA models.

The response series is stable if the roots of the homogeneous characteristic equation of $\phi(L) = 0$ lie in the unit circle.

If the response series y_t is not stable, then you can difference it to form a stationary ARIMA model. Do this by specifying the degrees of integration *d.*

ARIMAX

The response series y_t have to be differenced *before* including the exogenous covariates if you specify the degree of integration *d*. In other words, the exogenous covariates enter a model with a *stationary response*. Therefore, the ARIMAX(*p*,*d*,*q*) model is

$$\phi(L)y_t = c^* + x_t'\beta + \theta^*(L)\varepsilon_t$$

where $c^* = c/(1-L)^d$ and $\theta^*(L) = \theta(L)/\left((1-L)\right)^d$. Subsequently, the interpretation of β has changed to the expected effect a unit increase in the predictor has on the *difference* between current and lagged values of the response (conditional on those lagged values).

ARIMAX

Whether the predictor series x_t are stationary have do be assessed in advance. All predictor series that are not stationary have to be differenced during the data preprocessing stage. If x_t is nonstationary, then a test for the significance of β can produce a false negative. The practical interpretation of β changes if you difference the predictor series.

The maximum likelihood estimation for conditional mean models is used. In Matlab either a Gaussian or Student’s t for the distribution of the innovations can be specified .

ARIMAX

Seasonal components can be included in an ARIMAX model which creates a SARIMAX(R,d,M)(R_s,d_s,M_s) model. Assuming that the response series y_t is stationary, the model has the form

$$\phi(L)\Phi(L)y_t = c + x_t'\beta + \theta(L)\Theta(L)\varepsilon_t$$

where *Φ(L)* and *Θ(L)* are the seasonal lag polynomials.

The exogenous covariates are treated as fixed during estimation and inference.

Box-Jenkins Approach

- Determination of ARIMAX(R,d,M) model structure
- Estimation and model selection
- Test for model adequacy

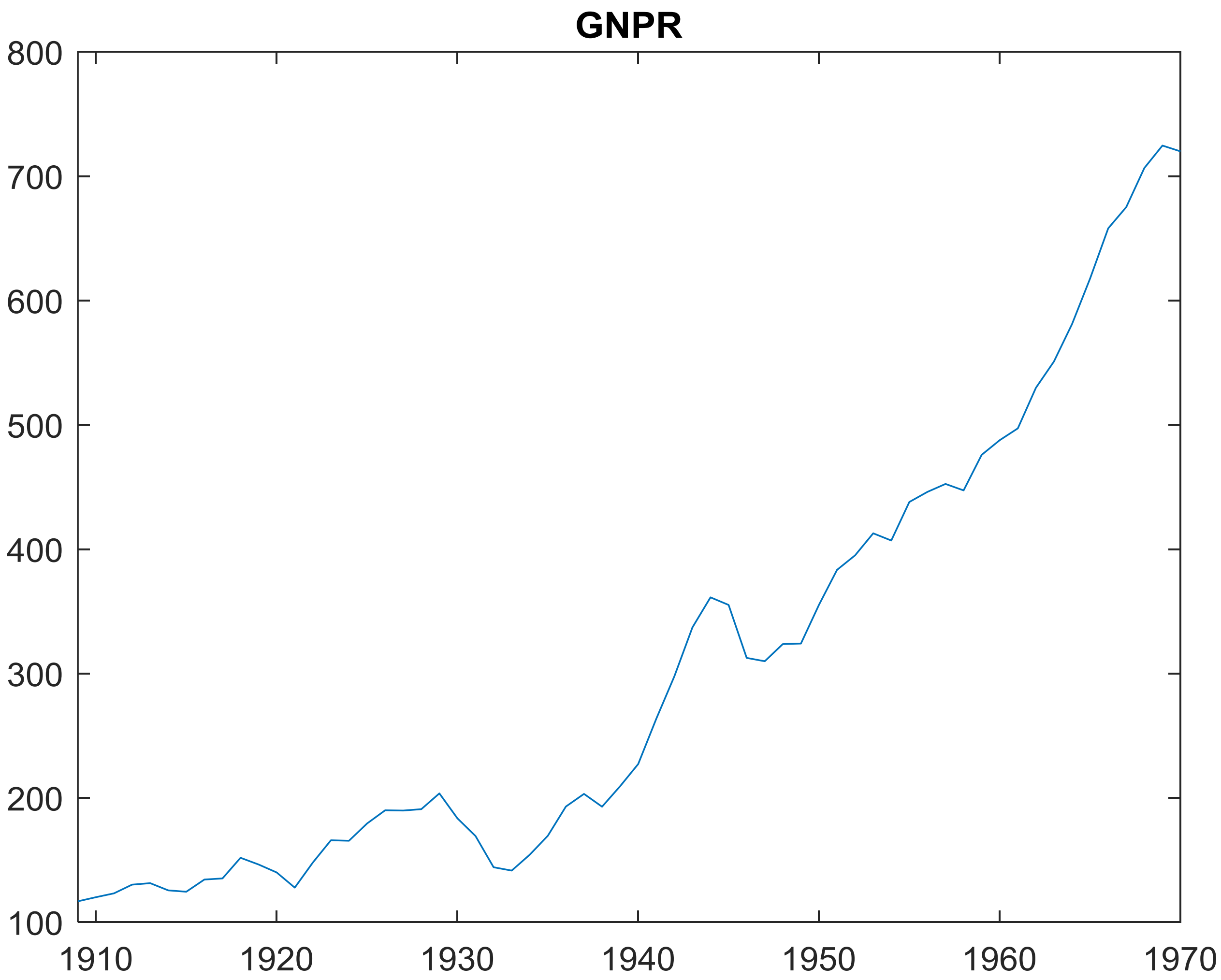
Case 1 - ARIMAX

- **Let’s consider annual time series of the current US real gross national product (GNPR) in USA**
- **The period is 1860 to 1970**
- **The prospective exogenous variables are:**
  - **Current industrial production index (IPI)**
  - **Employment (E)**
  - **Wages (WR)**

Case 1 - ARIMAX

- Among the series, some of the sample start dates begin in different years.
- We have to remove all leading NaNs(not a number) from the data by applying listwise deletion.

Case 1 - ARIMAX



Case 1 - ARIMAX

- Definitely the TS is non-stationary and we transform it to rates, more precisely logarithmic returns
- We do the same transformation with the exogenous variables also

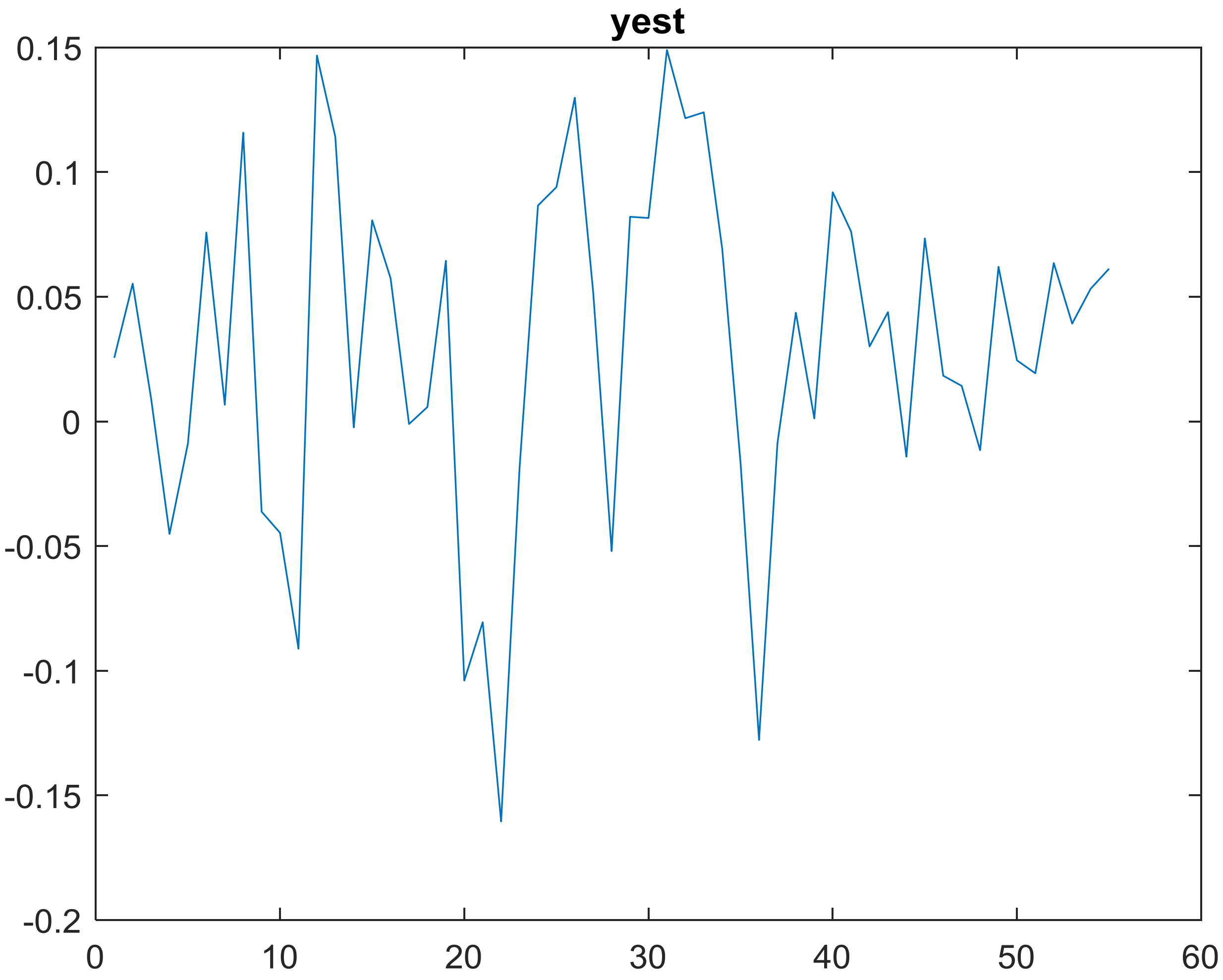
Case 1 - ARIMAX

- **We have to partition the timeline of the sample into pre-sample, estimation, and forecast periods.**
  - **The responses at the end of the estimation period have to be specified as a pre-sample to initialize the autoregressive term for the forecasts.**
  - **The predictor data at the end of the estimation period also have to be specified as a pre-sample to initialize the moving average component for the forecasts.**
  - **We have to include the effects of the predictor variables on the forecasted responses by specifying future predictor data.**

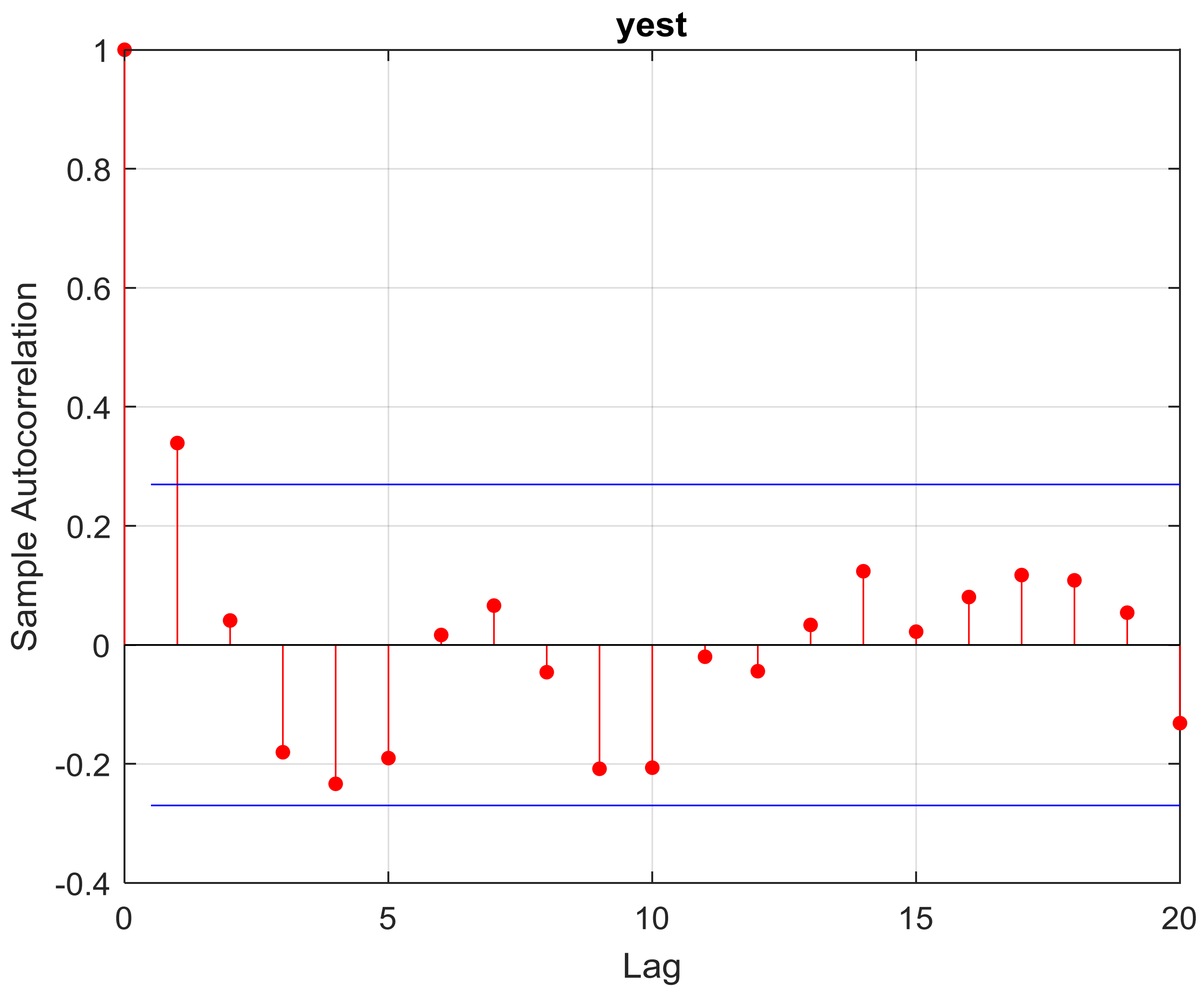
Case 1 - ARIMAX

- All leading NaNs from the data were removed.
- The response and predictor variables were converted to rates.
- The total size of the sample is 61 observations.
- The size of the pre-sample set is 1.
- The size of the estimation set is 55.
- The size of the test set (for forecasting) is 5.

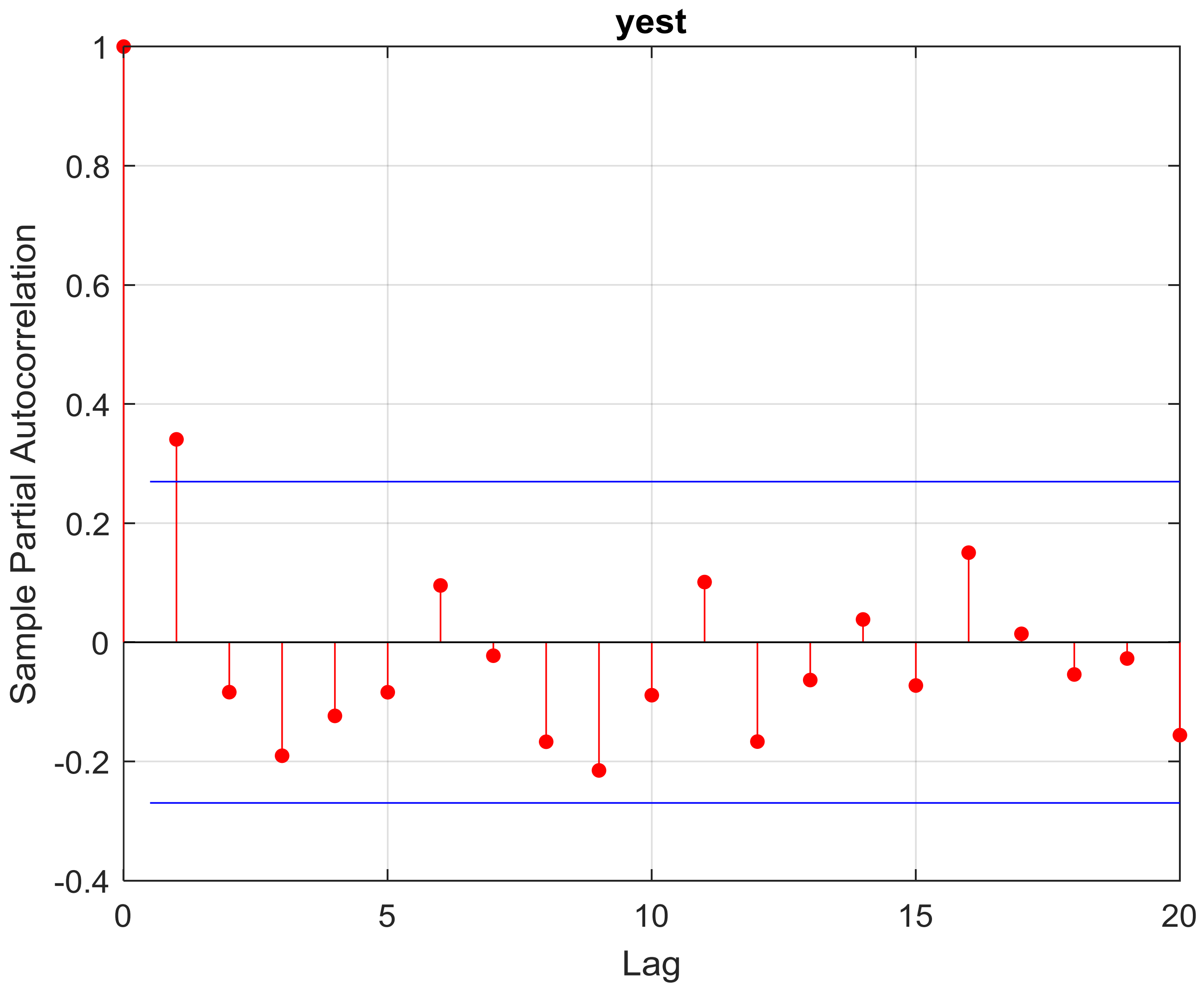
Case 1 - ARIMAX



Case 1 - ARIMAX



Case 1 - ARIMAX



Case 1 - ARIMAX

- **An ARMAX(1,1) model that predicts the current US real gross national product (GNPR) rate with the current industrial production index (IPI), employment (E), and real wages (WR) rates as exogenous variables seems to be appropriate.**

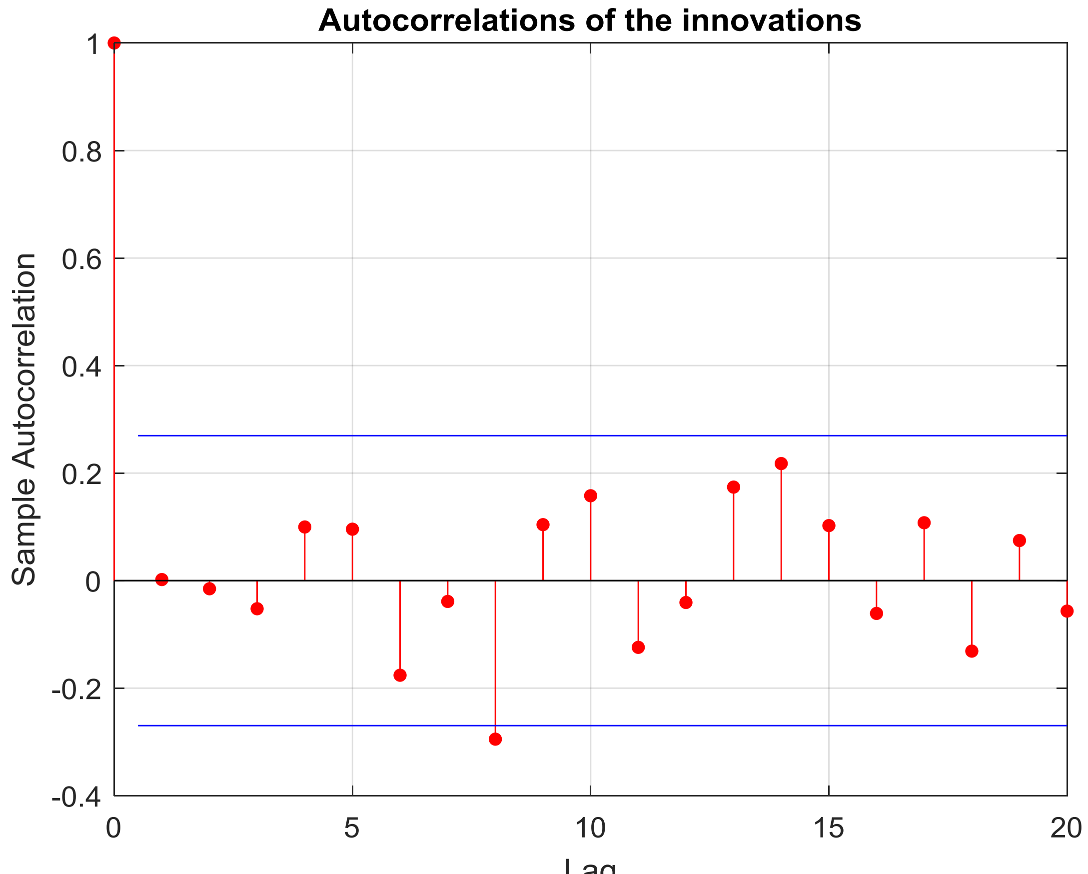
ARIMAX(1,0,1) Model (Gaussian Distribution)

Effective Sample Size: 55
Number of Estimated Parameters: 7
LogLikelihood: 138.574
AIC: -263.147
BIC: -249.096

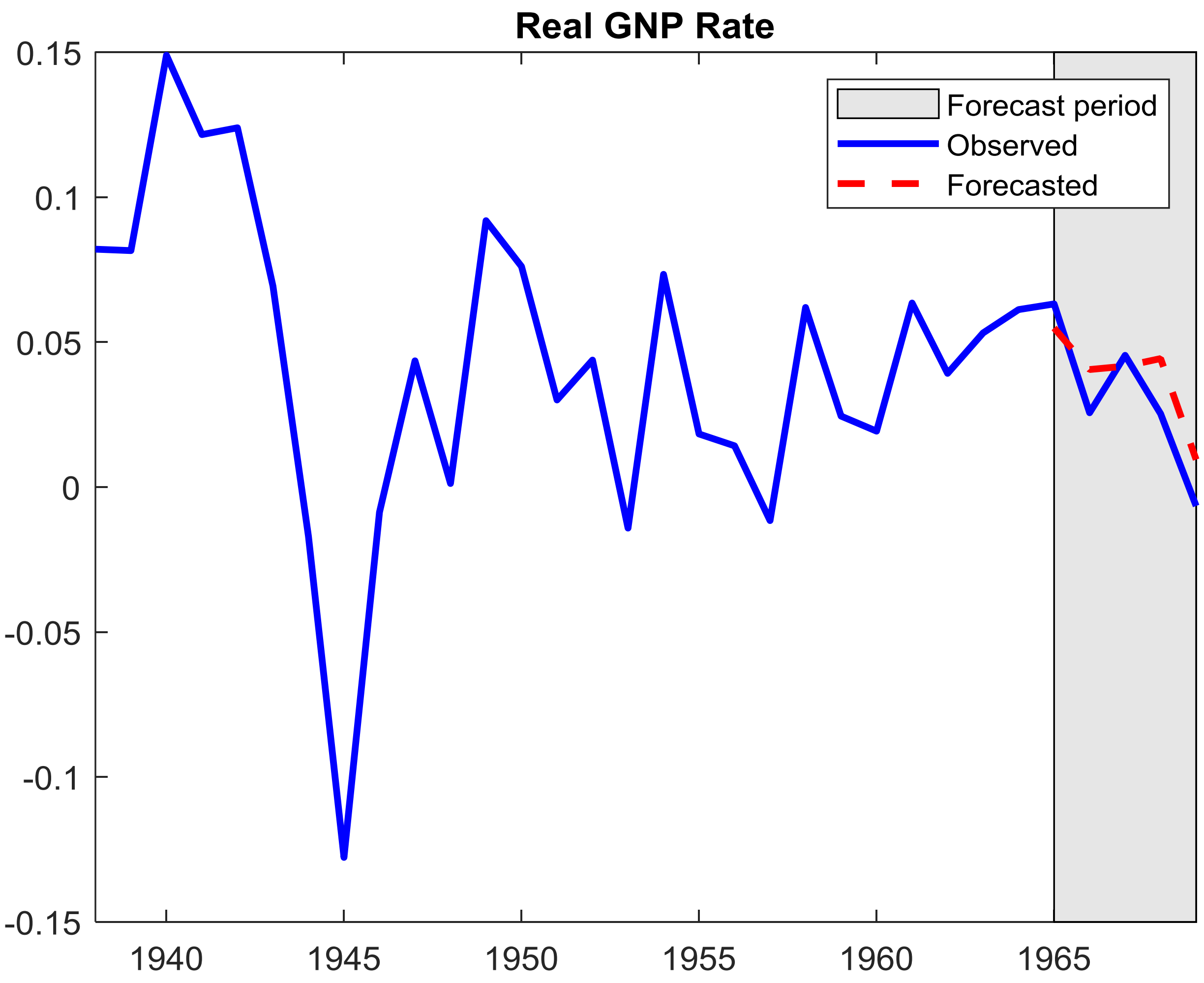
	Value	StandardError	TStatistic	PValue
Constant	0.008646	0.0037667	2.2955	0.021705
AR{1}	-0.057691	0.056367	-1.0235	0.30607
MA{1}	0.22722	0.13491	1.6842	0.092136
Beta(1)	0.07601	0.044092	1.7239	0.084729
Beta(2)	1.4219	0.16355	8.6941	3.4964e-18
Beta(3)	0.06118	0.10459	0.58496	0.55857
Variance	0.0003793	9.5603e-05	3.9684	7.2349e-05

Case 1 - ARIMAX

- The result of the Ljung-Box test is that the innovations can be considered as a white noise.
- The sample autocorrelations lie in the 95% confidence interval



Case 1 - ARIMAX



SARIMA

Many time series collected periodically (e.g., quarterly or monthly) exhibit a seasonal trend, meaning there is a relationship between observations made during the same period in successive years.

In addition to this seasonal relationship, there can also be a relationship between observations made during successive periods.

The seasonality mixes with the correlations among the sequential months/quarters

The multiplicative ARIMA model is an extension of the ARIMA model that addresses seasonality and potential seasonal unit roots.

SARIMA

In lag operator polynomial notation, $L^i y_t = y_{t-i}$. **For a series with periodicity** s**, the multiplicative ARIMA**$(R,d,M)\times(R_s,d_s,M_s)_s$ **is given by**

$$\phi(L)\Phi(1-L)^d(1-L^s)^{d_s}y_t = c + \theta(L)\Theta(L)\varepsilon_t$$

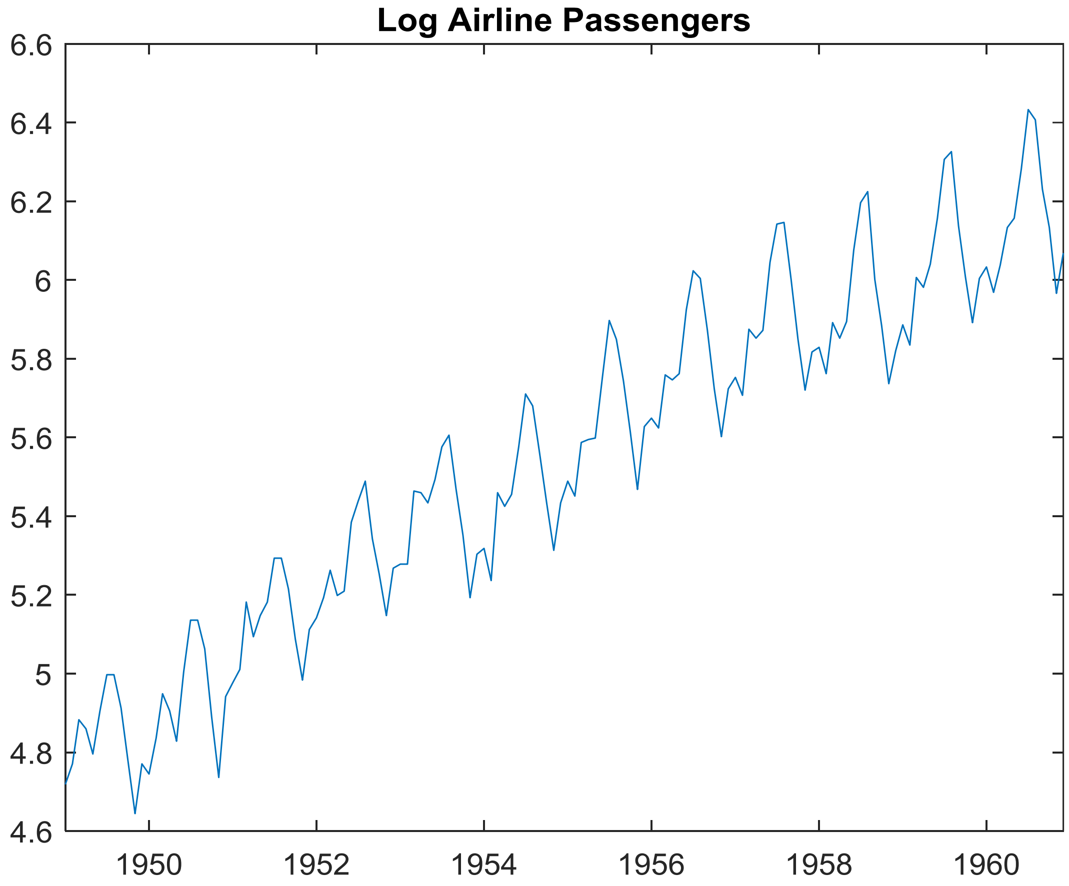
Here we have stable, degree R AR operator $\phi(L) = (1-\phi_1 L - \cdots - \phi_R L_R)$ and $\Phi(L)$ is also stable, degree p_s AR operator of the same form. Similarly, the degree M MA operator is invertible $\theta_M(L) = (1+\theta_1 L + \cdots + \theta_M L_M)$ and $\Theta(L)$ is an invertible, degree M_s MA operator of the same form.

ARIMA with Seasonality

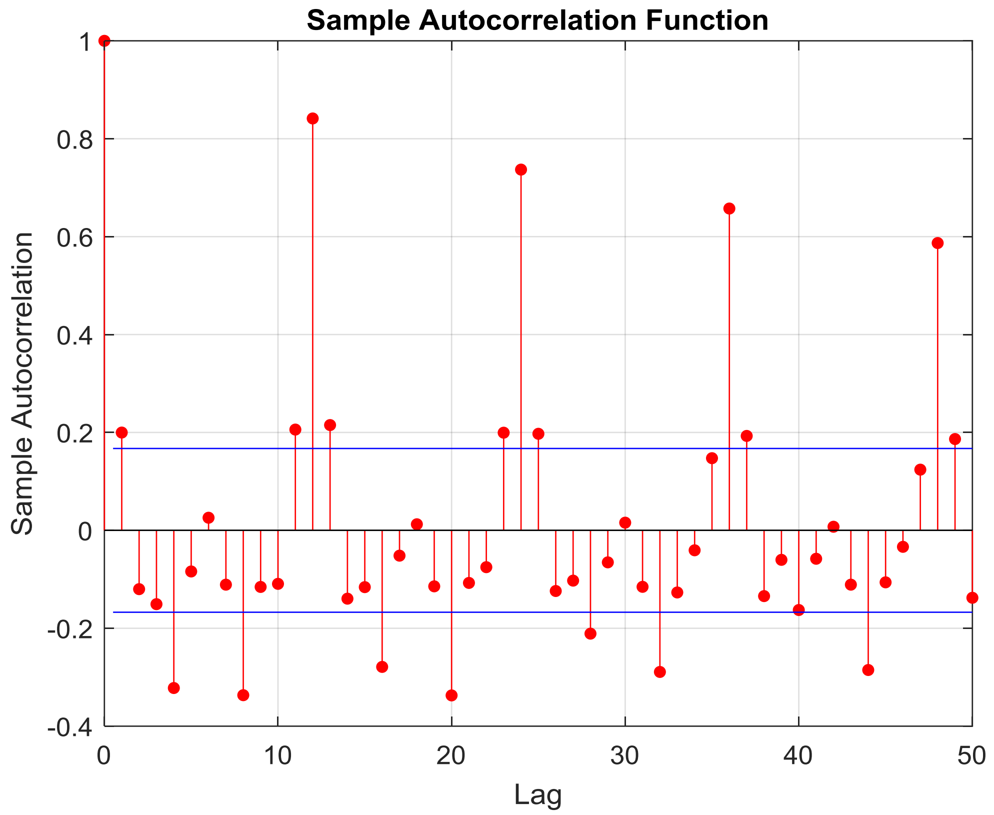
The autocorrelation function for seasonal time series has typical peeks at an interval of 12 (in the case of monthly data) or four (quarterly data).

Case 2 - SARIMA

- **The time series is monthly international airline passenger numbers for USA from 1949 to 1960.**



Case 2 - SARIMA



Case 2 - SARIMA

- We consider the following model:

$$(1-L)(1-L^{12})y_t = (1-\theta_1 L)(1-\Theta_{12}L^{12})\varepsilon_t$$

ARIMA(0,1,1) Model Seasonally Integrated with Seasonal MA(12) (Gaussian Distribution):

	Value	StandardError	TStatistic	PValue
Constant	0	0	NaN	NaN
MA{1}	-0.37716	0.073426	-5.1366	2.7972e-07
SMA{12}	-0.57238	0.093933	-6.0935	1.1047e-09
Variance	0.0013887	0.00015242	9.1115	8.1249e-20

Case 2 – SARIMA and ARMAX

- **An ARIMAX(0,1,1) model and a regression component containing last 12 observations can be also applied to the data.**

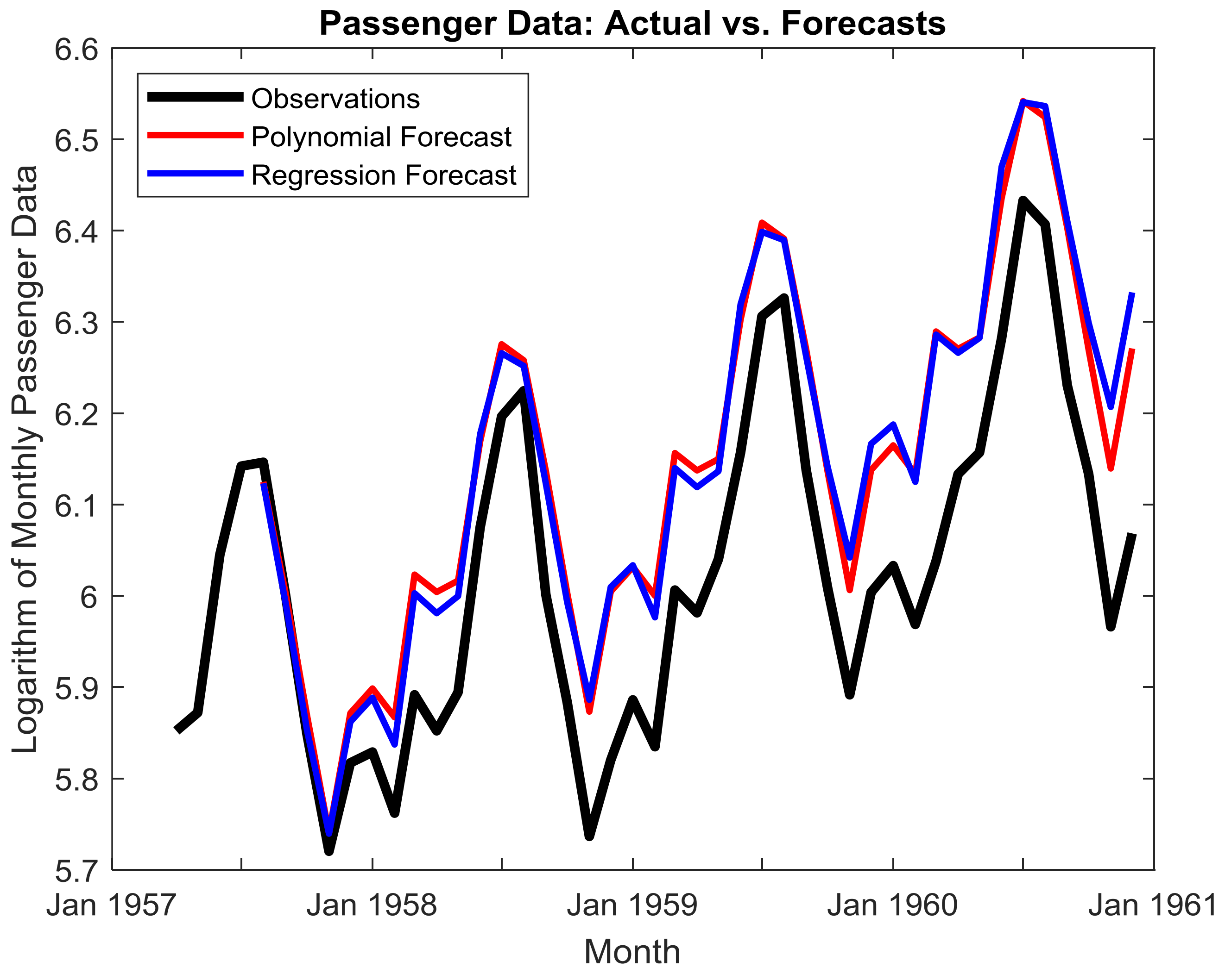
ARIMAX(0,1,1) Model Seasonally Integrated (Gaussian Distribution):

	Value	StandardError	TStatistic	PValue
Constant	0	0	NaN	NaN
MA{1}	-0.40711	0.084387	-4.8242	1.4053e-06
Beta(1)	-0.002577	0.025168	-0.10239	0.91845
Beta(2)	-0.0057769	0.031885	-0.18118	0.85623
Beta(3)	-0.0022034	0.030527	-0.072179	0.94246
Beta(4)	0.00094737	0.019867	0.047686	0.96197
Beta(5)	-0.0012146	0.017981	-0.067551	0.94614
Beta(6)	0.00487	0.018374	0.26505	0.79097
Beta(7)	-0.0087944	0.015285	-0.57535	0.56505
Beta(8)	0.0048346	0.012484	0.38728	0.69855
Beta(9)	0.001437	0.018245	0.078758	0.93722
Beta(10)	0.009274	0.014751	0.62869	0.52955
Beta(11)	0.0073665	0.0105	0.70158	0.48294
Beta(12)	0.00098841	0.014295	0.069146	0.94487
Variance	0.0017715	0.00024657	7.1848	6.7329e-13

Case 2 – SARIMA and ARMAX

- Both models produce similar results.

Case 2 – SARIMA and ARMAX



ARIMA with Long-range dependence

- The classical ARIMA can successfully model short-term dependences in the data, but the autocovariance diminish quickly when the time interval ***h*** increase

ARIMA with Long-range dependence

- More precisely the autocorrelation function is geometrically limited

$$|\rho(k)| \le Cr^k, \quad k = 1,2,\ldots,$$

$$C > 0, \quad 0 < r < 1$$

ARIMA with Long-range dependence

- A class of ARIMA models with slowly diminishing autocorrelation function is proposed by [Hosking (1981)] and [Granger and Joyeux (1980)]

ARIMA with Long-range dependence

- Their basic characteristic is introduction of fractional differencing
- Formally we use standard binomial series to define the fractional differencing

ARIMA with Long-range dependence

$$\nabla^d = (1-B)^d = \sum_{j=0}^{\infty} \binom{d}{j} (-B)^j = 1 - dB - \frac{1}{2} d(1-d) B^2 - \frac{1}{6} d(1-d)(2-d) B^3 - \ldots$$

$d - fraction$

$$\nabla^d = (1-B)^d = \sum_{j=0}^{\infty} \pi_j B^j,$$

$$\pi_j = \prod_{0<k\le j} \frac{k-1-d}{k} = \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)}, \quad j = 0,1,2,\ldots,$$

$where \quad \Gamma(\ ) is \ Gamma\ Function$

Gamma Function is an extension of factorial for complex numbers

ARIMA with Long-range dependence

- When $-1/2<d<1/2$ the process is stationary and invertible, and the coefficients of the infinite MA and AR representations diminish as a power function in contrast of standard ARIMA, where we have exponential decline

ARIMA with Long-range dependence

- For the processes with long-range dependence we have:
  - $\Sigma \; |\rho(k)| = \infty$
  - $\rho(k) \sim ck^{-\alpha}$, $\alpha \in (0, 1)$

ARIMA with Long-range dependence

- ARIMA processes with long-range dependence are usually called (Fractional ARIMA) or FARIMA