DATA 201: Time Series Analysis
Linear Regression Model
Lecture 13: Trends In Time Series & Out of Sample Forecasting

Lulu Wang

Data Analytics
Dickinson College

3/7/2025

The Auto-Regressive (AR) Model

Forecasting

Seasonality

Out-of-sample forecasts

Stationary

Trends in time series

Auto-Regressive (AR) Model

- In general, an **AR(p)** model is defined as:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + \epsilon_t$$

where:

- $\beta_0, \beta_1, \ldots, \beta_p$ are parameters to be estimated.
- ϵ_t is a white noise error term with mean zero and variance σ_ϵ^2.

Conditional Expectation in AR(p) Model

Property 1: Conditional Expectation

$$E(Y_t|Y_{t-1},\dots,Y_{t-p}) = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p}$$

It can be interpreted as a **forecast** since only past information is used to produce an expectation of the variable today.

Property 2: Unconditional Expectation

$$E(Y_t) = \frac{\beta_0}{1 - \sum_{j=1}^{p} \beta_j}$$

since:

$$E(Y_t) = \beta_0 + \beta_1 E(Y_{t-1}) + \cdots + \beta_p E(Y_{t-p}) + E(\epsilon_t)$$

Assuming that $E(Y_t) = E(Y_{t-k})$ for all values of k.
It represents the long-run mean of Y_t

Variance of AR(p) Model

Property 3: Variance

$$Var(Y_t) = \frac{\sigma_\epsilon^2}{1 - \sum_{j=1}^{p} \beta_j^2}$$

since:

$$Var(Y_t) = \beta_1^2 Var(Y_{t-1}) + \cdots + \beta_p^2 Var(Y_{t-p}) + Var(\epsilon_t)$$

$$= \left(\sum_{j=1}^{p} \beta_j^2 \right) Var(Y_t) + \sigma_\epsilon^2$$

This shows how past values contribute to the variance of the current value.

Persistence and Mean Reversion in AR(p)

Property 4: Persistence and Mean Reversion

$$\sum_{i=1}^{p} \beta_i$$

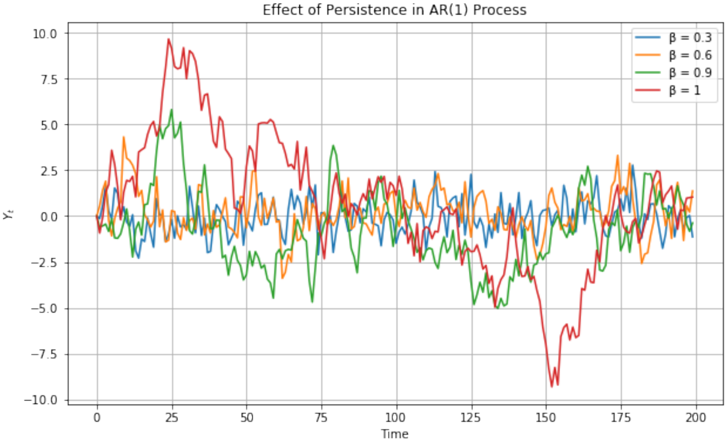
measures the **persistence** of the time series. Persistence can also be interpreted as **mean reversion**, which represents how quickly a time series reverts back to its mean.

- **Low persistence** $\Rightarrow$ quick mean reversion.
- **High persistence** $\Rightarrow$ slow mean reversion.

Key Insight: The closer the sum of β_i is to 1, the more persistent the process, meaning past values strongly influence future values.

Persistence and Mean Reversion in AR(1)

The plot below shows the **effect of persistence** in an **AR(1) process** with different values of β:



Forecasting with AR models

- The current period is time t, and we are interested in forecasting the value of the variable in future periods $t+1$, $t+2$, ...
- Statistically speaking, we want to calculate:

$$E(Y_{t+1}|Y_t), \quad E(Y_{t+2}|Y_t), \quad \ldots$$

- We simplify by assuming an **AR(1) model** ($p = 1$).

1. **Estimate the AR(1) model using OLS:**

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t$$

2. **Compute the forecast for $t+1$:**

$$E(Y_{t+1}|Y_t) = \hat{Y}_{t+1} = \hat{\beta}_0 + \hat{\beta}_1 Y_t$$

Forecasting with AR models (cont.)

3. Compute the forecast for $t + 2$:

$$\begin{aligned} E(Y_{t+2}|Y_t) &= \hat{Y}_{t+2} \\ &= \hat{\beta}_0 + \hat{\beta}_1 \hat{Y}_{t+1} \\ &= \hat{\beta}_0 + \hat{\beta}_1(\hat{\beta}_0 + \hat{\beta}_1 Y_t) \\ &= \hat{\beta}_0(1 + \hat{\beta}_1) + \hat{\beta}_1^2 Y_t \end{aligned}$$

4. Compute the forecast for $t + 3$:

$$\begin{aligned} E(Y_{t+3}|Y_t) &= \hat{Y}_{t+3} \\ &= \hat{\beta}_0 + \hat{\beta}_1 \hat{Y}_{t+2} \\ &= \hat{\beta}_0 + \hat{\beta}_1(\hat{\beta}_0(1 + \hat{\beta}_1) + \hat{\beta}_1^2 Y_t) \\ &= \hat{\beta}_0(1 + \hat{\beta}_1 + \hat{\beta}_1^2) + \hat{\beta}_1^3 Y_t \end{aligned}$$

Forecasting with AR models (cont.) : Forecasting daily returns

- Building a predictive time series model requires the following steps:
  1. *Model selection*: for AR models this means choosing p.
  2. *Model estimation*: estimating the parameters of the model.
  3. *Forecasting*: producing the forecasts based on the model and the estimated parameters.
  - The function `get_prediction()` takes the fitted object and returns the point forecasts and confidence intervals.

	Predicted Values	CI Lower	CI Upper
6905	0.000433	-0.021425	0.022290
6906	0.000408	-0.021487	0.022302
6907	0.000333	-0.021581	0.022248
6908	0.000339	-0.021576	0.022253
6909	0.000342	-0.021573	0.022257
6910	0.000341	-0.021574	0.022256
6911	0.000341	-0.021574	0.022256
6912	0.000341	-0.021574	0.022256

- The forecasts start from a different value but rapidly converge

Forecasting with AR models (cont.)

- The forecasting formula can be generalized to forecasting k steps ahead which is given by

$$\hat{E}(Y_{t+k}|Y_t) = \hat{\beta}_0(\sum_{j=1}^{k} \hat{\beta}_1^{\,k-1}) + \hat{\beta}_1^{\,k} \times Y_t$$

- Note, the The sum $\sum_{j=1}^{k} \hat{\beta}_1^{\,k-1} = 1 + \hat{\beta}_1 + \hat{\beta}_1^{\,2} + ... + \hat{\beta}_1^{\,k-1}$ is a geometric series with sum:

$$\sum_{j=1}^{k} \hat{\beta}_1^{\,k-1} = \frac{1 - \hat{\beta}_1^{\,k}}{1 - \hat{\beta}_1}, for|\beta_1| < 1$$

- Substituting this into our forecast equation:

$$\hat{E}(Y_{t+k}|Y_t) = \hat{\beta}_0(\frac{1 - \hat{\beta}_1^{\,k}}{1 - \hat{\beta}_1}) + \hat{\beta}_1^{\,k} \times Y_t$$

Forecasting with AR models (cont.)

- Substituting this into our forecast equation:

$$\hat{E}(Y_{t+k}|Y_t) = \hat{\beta}_0\left(\frac{1-\hat{\beta}_1^{\,k}}{1-\hat{\beta}_1}\right) + \hat{\beta}_1^{\,k} \times Y_t$$

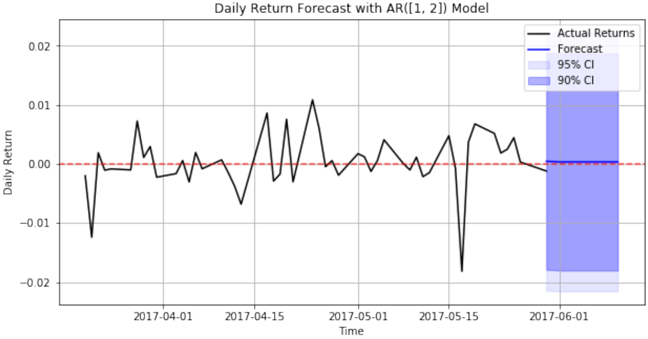
- Taking the limit as $k \to \infty$:

$$\lim_{k\to\infty} \hat{E}(Y_{t+k}|Y_t) = \frac{\hat{\beta}_0}{1-\hat{\beta}_1}.$$

- This is the long-run mean of the process.

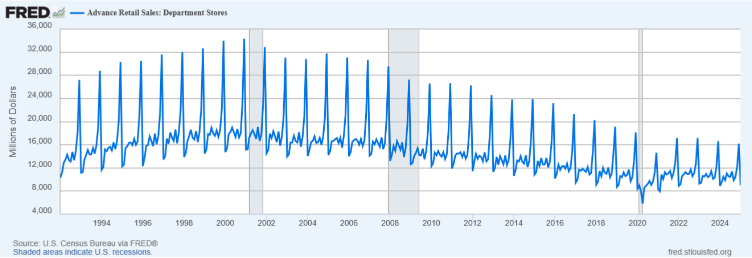
Visualizing the forecasts

- Visualizing the forecasts and the uncertainty around the forecast is a useful tool to understand the strength and the weakness of the forecasts.



Seasonality

- The term Seasonality refers to the characteristic of some time series to show a regular pattern related to the frequency of the variable (eg, daily, monthly, quarterly) that repeats over time.
- Below: Advance Retail Sales: Department Stores (FRED ticker RSDSELDN) at the monthly frequency



Seasonality : Department Stores Sales (Cont.)

- Let's assume that Y_t is the variable we are interested to model; seasonality can be accounted for as follows

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \gamma_2 FEB_t + \gamma_3 MAR_t + \gamma_4 APR_t + \gamma_5 MAY_t + \gamma_6 JUN_t + \gamma_7 JUL_t + \gamma_8 AUG_t + \gamma_9 SEP_t + \gamma_{10} OCT_t + \gamma_{11} NOV_t + \gamma_{12} DEC_t + \varepsilon_t$$

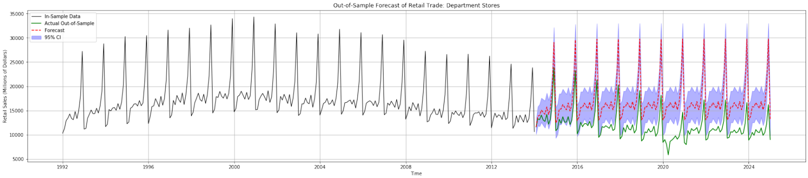
Seasonality : Department stores sales (cont.)

- The regression results are shown below:

```
==============================================================================
Dep. Variable:               RSDSELDN   No. Observations:                  397
Model:                   AutoReg-X(1)   Log Likelihood               -3421.039
Method:               Conditional MLE   S.D. of innovations           1366.555
Date:                Mon, 03 Mar 2025   AIC                           6870.077
Time:                        19:48:05   BIC                           6925.817
Sample:                    02-01-1992   HQIC                          6892.160
                         - 01-01-2025
================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const        -1.165e+04    630.140    -18.488      0.000   -1.29e+04   -1.04e+04
RSDSELDN.L1      0.8966      0.022     40.145      0.000       0.853       0.940
April          1.29e+04    428.435     30.108      0.000    1.21e+04    1.37e+04
August        1.431e+04    431.669     33.161      0.000    1.35e+04    1.52e+04
December      2.148e+04    380.459     56.467      0.000    2.07e+04    2.22e+04
February      1.349e+04    464.055     29.068      0.000    1.26e+04    1.44e+04
July          1.277e+04    426.582     29.935      0.000    1.19e+04    1.36e+04
June          1.261e+04    418.674     30.124      0.000    1.18e+04    1.34e+04
March         1.475e+04    454.657     32.443      0.000    1.39e+04    1.56e+04
May           1.404e+04    431.543     32.541      0.000    1.32e+04    1.49e+04
November      1.666e+04    422.558     39.427      0.000    1.58e+04    1.75e+04
October       1.411e+04    437.120     32.273      0.000    1.33e+04     1.5e+04
September     1.162e+04    415.192     27.994      0.000    1.08e+04    1.24e+04
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1153           +0.0000j            1.1153            0.0000
-----------------------------------------------------------------------------
```

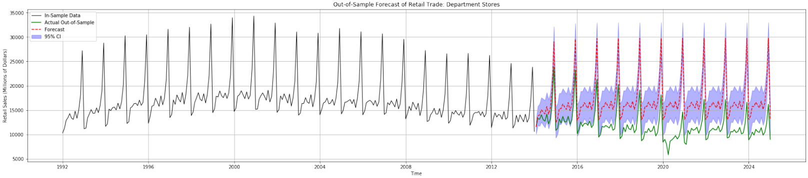
Out-of-sample forecasts

- An out-of-sample exercise consists of the following:
  1. Select and estimate your model up to a certain date (eg, **2014-01-01**)
  2. Forecast from that date forward (eg, 132 months ahead)
- In the graph below the red line is the 1 to 132 steps ahead forecast, while the shaded blue areas represent the 95% forecast interval
- The green line represents the actual realization of the variable.



Out-of-sample forecasts

- Notice:
  - We forecast a larger peak at the end of the year relative to what happened.
  - In the other quarters, the red line is typically above the green one, indicating that we are overpredicting sales.
  - Is our model *misspecified*, i.e., systematically wrong?



Stationary and Non-Stationary Time Series

- A time series is **stationary** if the *statistical properties of its distribution* (*e.g., mean, variance, autocorrelation*) are constant in the long-run.
- Example: if the returns are stationary, we expect that their distribution in 10 years will be the same as the distribution today; if they are non-stationary, their mean and/or variance will be different.

Stationary and Non-Stationary Time Series

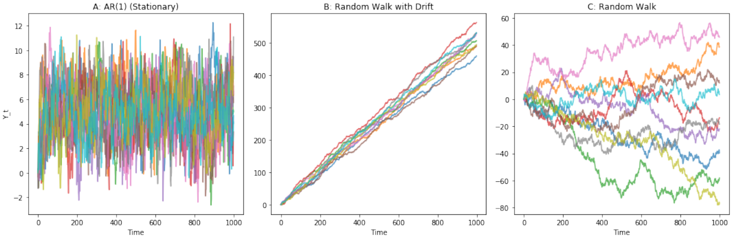
- This does not imply that the short-run mean and variance of the time series should be constant, but can be a function of past lags, macroeconomic variables, etc.
- Let's consider an AR(1) model:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t \quad (\text{with } |\beta_1| < 1)$$

  - (Short-run) **Conditional mean**: $E(Y_t|Y_{t-1}) = \beta_0 + \beta_1 \cdot Y_{t-1}$
  - (Long-run) **Unconditional mean**: $E(Y_t) = \frac{\beta_0}{1-\beta_1}$
- A time series is also stationary if the variance is constant in the long-run and, more generally, the complete distribution.

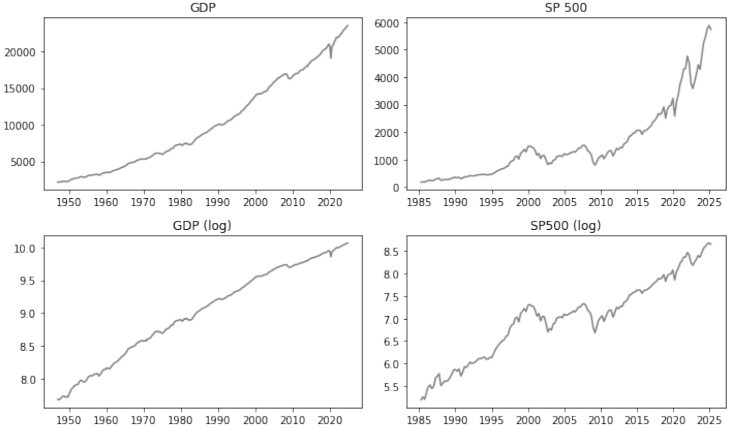
Simulated Time Series Models

- The Figure shows 10 time series simulated for 500 periods from the following models (with $\epsilon_t \sim N(0,1)$):
  1. $Y_t = 0.5 + 0.9Y_{t-1} + \epsilon_t$
  2. $Y_t = 0.5 + Y_{t-1} + \epsilon_t$
  3. $Y_t = Y_{t-1} + \epsilon_t$
- Who is who? Match models 1, 2, 3 to the plots A, B, C.
- Are these simulated series *stationary*? That is:
  - Does the mean seem approximately constant over time?
  - Does the variance seem approximately constant over time?



Economic and financial variables

- Variables that are expressed in $ are often transformed using the logarithm
- One reason for log-transforming a variable is to linearize its exponential behavior



Trend-Stationary Model

- A characteristic of many economic and financial variables is that they grow over time.
- One approach to account for this behavior is to assume a model in which the variable fluctuates in a stationary manner around a deterministic **trend**.
- The trend-stationary model assumes that a time series Y_t follows the process:

$$Y_t = \beta_0 + \beta_1 \cdot t + d_t$$

  where d_t is the deviation from the trend, which is assumed to be stationary.
- Two components:
  1. Permanent (*non-stationary*): $\beta_0 + \beta_1 t$
  2. Transitory (*stationary*): d_t (e.g., $d_t = \phi d_{t-1} + \epsilon_t$)

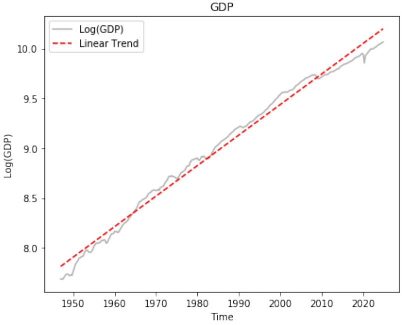
Estimation: Linear

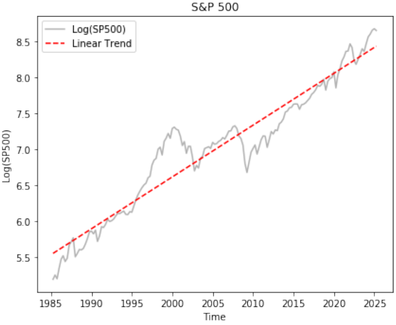
- The trend-stationary model $Y_t = \beta_0 + \beta_1 \cdot t + d_t$ can be estimated using OLS() or AutoReg()command.
- where the estimate of $\hat{\beta}_1$ is 0.0077 and indicates that real GDP is expected to grow 0.77% every quarter. for the S&P 500 $\hat{\beta}_1 = 0.0180$ and represents a quarterly growth of 1.8%.

	(Intercept)	trend
GDP	7.8059	0.0077
SP500	5.5394	0.0180

Linear trend

- These plots show the log of the variables considered above and the fitted linear trend



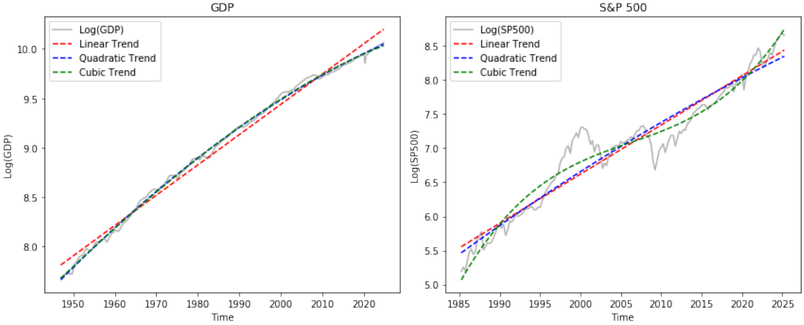


Nonlinear Trend

- The trend does not necessarily have to be linear, but could also be quadratic or cubic:

$$Y_t = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2 + \beta_3 \cdot t^3 + d_t$$

where t^2 and t^3 are the square and cube of the trend variable.



Residuals

- The plots below show the deviations of the variables from the linear/ nonlinear trend (d_t): stationary or non-stationary?