DATA 201: Time Series Analysis Linear Regression Model Lecture 12: The Auto-Regressive (AR) Model

Lulu Wang

Data Analytics
Dickinson College

3/4/2025

The Auto-Regressive (AR) Model

Forecasting

Seasonality

Interpretation of the Two Equations

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

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

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

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:

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

- 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:

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

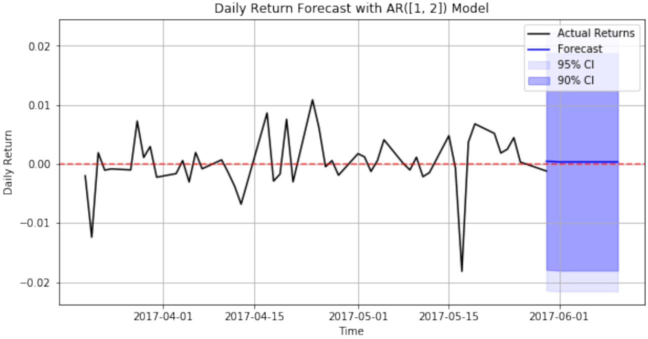
- Taking the limit as $k \to \infty$:

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

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

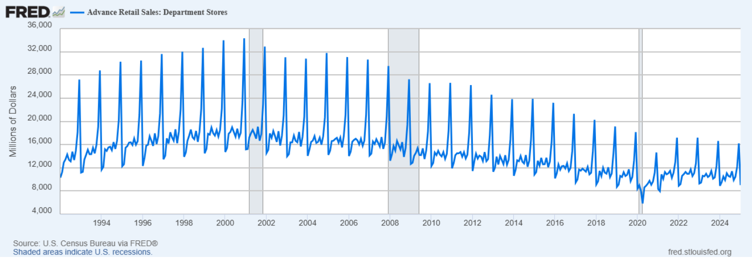
Visualizing the forecasts

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



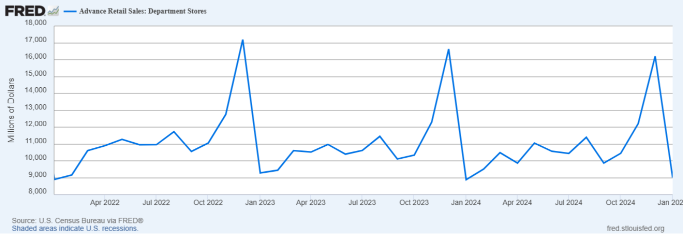
Seasonality

- 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

- Measures monthly sales at department stores in the U.S. This data is not seasonally adjusted.



- Strong Seasonality:
  - The data shows sharp peaks every December, indicating higher sales during the holiday season (Christmas shopping).
  - This suggests that department stores rely heavily on holiday spending.

Seasonality : Department stores sales (Cont.)

- There are many ways to model the seasonality in the data relatively simple approach is to use dummy variables to capture the seasonal pattern
- To model seasonality in time series, we create dummy variables that take value 1 in a certain period and are 0 otherwise
  - **Quarterly data**: **Q1, Q2, Q3, and Q4**
  - **Monthly data**: **JAN, FEB, MAR**, etc.
- With seasonal dummy variables, we need to be careful of the *dummy variable trap*

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 - 01-01-2025	HQIC	6892.160

	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

Key findings on retail sales at department stores

- **High Persistence**:
  - Retail sales exhibit strong persistence with an AR(1) coefficient of 0.8966.
- **Seasonal Coefficient Interpretation**:
  - January is the **baseline month** (left out dummy).
  - All other seasonal dummies represent **differences relative to January**.
- **Seasonal Pattern**:
  - Sales are higher than January for all months.
  - Gradual increase in sales from January, remaining stable during summer.
  - Significant spike in November and December.

Interpretation of the two equations

These two equations represent two alternative ways to model **seasonality** in a time series regression using **monthly dummy variables**.

- **Equation 1:** Excludes one dummy variable (baseline category).
- **Equation 2:** Includes all dummy variables but removes the intercept.

The choice between these two models impacts how we interpret the coefficients and the overall model structure.

Equation 1: Excluding one dummy variable (Baseline Model)

Regression Equation:

$$Y_t = \beta_0 + \gamma_2 FEB_t + \gamma_3 MAR_t + \gamma_4 APR_t + \gamma_5 MAY_t + \cdots + \gamma_{12} DEC_t + \varepsilon_t$$

Key Points:

- **Excludes January** to avoid the **dummy variable trap** (multicollinearity issue).
- β_0 represents the **expected value of Y_t in January**.
- Each γ_j measures the difference between a specific month and January.
- If $\gamma_2 > 0$, February has a **higher expected value** than January.

This is the **standard approach** in regression modeling.

Equation 2: Including all dummy variables (No Intercept Model)

Regression Equation:

$$Y_t = \gamma_1 JAN_t + \gamma_2 FEB_t + \gamma_3 MAR_t + \cdots + \gamma_{12} DEC_t + \varepsilon_t$$

Key Points:

- Includes **all 12 dummy variables** but **removes the intercept**.
- Each γ_j represents the **absolute expected value** of Y_t for that month.
- No baseline comparison; each month is treated separately.

Useful when absolute monthly effects are needed.

Comparison and when to use each approach

Comparison Table:

Feature	Equation 1 (Baseline Model)	Equation 2 (No Intercept Model)
Intercept	Represents January's mean (β_0)	Not included
Dummy Variables	11 (excluding January)	12 (one for each month)
Coefficient Meaning	Difference from January	Absolute mean for each month
Preferred For Regression?	Yes (standard)	No (special cases)

Equation 1: Preferred for interpretation & statistical modeling.
Equation 2: Useful in special cases when absolute effects are needed.

Sell in May and Go Away?

- The "Sell in May and Go Away" strategy suggests that stock returns are weaker from May to October and stronger from November to April. This is a seasonal anomaly in financial markets.
- **Regression Equation:**

$$Y_t = \gamma_1 JAN_t + \gamma_2 FEB_t + \gamma_3 MAR_t + \cdots + \gamma_{12} DEC_t + \varepsilon_t$$

Sell in May and Go Away?

- There is some support for the idea that winter months (November-April) perform better.
- There is no strong evidence that May-October underperforms significantly.

AutoReg Model Results

Dep. Variable:	Monthly_Return	No. Observations:	328
Model:	AutoReg-X(0)	Log Likelihood	589.517
Method:	Conditional MLE	S.D. of innovations	0.040
Date:	Mon, 03 Mar 2025	AIC	-1153.033
Time:	19:47:53	BIC	-1103.724
Sample:	02-01-1990 - 05-01-2017	HQIC	-1133.360

	coef	std err	z	P>|z|	[0.025	0.975]
April	0.0157	0.008	2.075	0.038	0.001	0.031
August	-0.0100	0.008	-1.292	0.196	-0.025	0.005
December	0.0170	0.008	2.202	0.028	0.002	0.032
February	0.0040	0.008	0.523	0.601	-0.011	0.019
January	0.0013	0.008	0.173	0.862	-0.014	0.016
July	0.0083	0.008	1.077	0.281	-0.007	0.023
June	-0.0050	0.008	-0.652	0.515	-0.020	0.010
March	0.0144	0.008	1.893	0.058	-0.001	0.029
May	0.0105	0.008	1.386	0.166	-0.004	0.025
November	0.0149	0.008	1.928	0.054	-0.000	0.030
October	0.0163	0.008	2.117	0.034	0.001	0.031
September	-0.0047	0.008	-0.604	0.546	-0.020	0.010