GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 15: TIME SERIES ANALYSIS

LAST TIME:

- DATA EXPLORATION
- EXPERIMENTAL DESIGN

QUESTIONS?

I. TIME SERIES DATA
II. DIFFERENCES IN TIME SERIES DATA
III. DIAGNOSTIC TOOLS
IV. AUTOREGRESSIVE MODELS

EXERCISE:
V. PREDICTIVE MODELING

I. TIME SERIES DATA

Time series vs. Cross sectional

Time series vs. Cross sectional

Time series data

Measurements of the same data taken over a period of (usually regular) intervals of time

Time series vs. Cross sectional

Time series data

Measurements of the same data taken over a period of (usually regular) intervals of time

Cross sectional data

A snapshot in time of a group of data

Time series vs. Cross sectional

Time series data

Measurements of the same data taken over a period of (usually regular) intervals of time

Stock returns, temperature

Cross sectional data

A snapshot in time of a group of data

Time series vs. Cross sectional

Time series data

Measurements of the same data taken over a period of (usually regular) intervals of time

Stock returns, temperature

Cross sectional data

A snapshot in time of a group of data

Beer, Mammals, Iris

Lab: Taking a look at time series data

II. DIFFERENCES IN TIME SERIES DATA

So what went wrong? Why didn't the linear regression model work?

Assumptions of linear regression models

Assumptions of linear regression models

1. Linear relationship between dependent and independent variables

Assumptions of linear regression models

1. *Linear relationship between dependent and independent variables*
2. *Independence of the errors (serial correlation)*

Assumptions of linear regression models

1. *Linear relationship between dependent and independent variables*
2. *Independence of the errors (serial correlation)*
3. *Homoscedasticity (constant variance of the errors)*
    a. *Over time*
    b. *Between the variables over time*

Assumptions of linear regression models

1. *Linear relationship between dependent and independent variables*
2. *Independence of the errors (serial correlation)*
3. *Homoscedasticity (constant variance of the errors)*
    a. *Over time*
    b. *Between the variables over time*
4. *Normality of the errors*

Assumptions of linear regression models

1. *Linear relationship between dependent and independent variables*
2. *Independence of the errors (serial correlation)*
3. *Homoscedasticity (constant variance of the errors)*
    a. *Over time*
    b. *Between the variables over time*
4. *Normality of the errors*

Assumptions of linear regression models

1. *Linear relationship between dependent and independent variables*
2. *Independence of the errors (serial correlation)*
3. *Homoscedasticity (constant variance of the errors)*
    a. *Over time*
    b. *Between the variables over time*
4. *Normality of the errors*

Assumptions of linear regression models

1. *Linear relationship between dependent and independent variables*
2. *Independence of the errors (serial correlation)*
3. *Homoscedasticity (constant variance of the errors)*
    a. *Over time*
    b. *Between the variables over time*
4. *Normality of the errors*

Assumptions of linear regression models

1. *Linear relationship between dependent and independent variables*
2. *Independence of the errors (serial correlation)*
3. *Homoscedasticity (constant variance of the errors)*
    a. *Over time*
    b. *Between the variables over time*
4. *Normality of the errors*

Assumptions of linear regression models

1. *Linear relationship between dependent and independent variables*
2. *Independence of the errors (serial correlation)*
3. *Homoscedasticity (constant variance of the errors)*
    a. *Over time*
    b. *Between the variables over time*
4. *Normality of the errors*

Another important difference

Unlike cross sectional models, the order of the data is VERY important

Things to look for in time series data

Trends

measurements tend to increase or decrease over time

Seasonality / Cyclicality

An observable cycle in the data (days, years, weeks, secs, etc)

III. DIAGNOSTIC TOOLS

Assumptions of linear regression models

1. *Linear relationship between dependent and independent variables*
2. *Independence of the errors (serial correlation)*
3. *Homoscedasticity (constant variance of the errors)*
    a. *Over time*
    b. *~~Between the variables over time~~*
4. *Normality of the errors*

Assumptions of linear regression models

-------1. Linear relationship between dependent and independent variables

2. Independence of the errors (serial correlation)

3. Homoscedasticity (constant variance of the errors)

a. Over time

~~b. Between the variables over time~~

4. Normality of the errors

Assumptions of linear regression models

1. Linear relationship between dependent and independent variables

-------2. Independence of the errors (serial correlation)

3. Homoscedasticity (constant variance of the errors)

a. Over time

~~b. Between the variables over time~~

4. Normality of the errors

Assumptions of linear regression models

1. Linear relationship between dependent and independent variables

2. Independence of the errors (serial correlation)

-------3. Homoscedasticity (constant variance of the errors)

a. Over time

~~b. Between the variables over time~~

4. Normality of the errors

Assumptions of linear regression models

1. Linear relationship between dependent and independent variables

2. Independence of the errors (serial correlation)

3. Homoscedasticity (constant variance of the errors)

a. Over time

b. Between the variables over time

-------4. Normality of the errors

4. Normality of the errors

Lab: Residuals plot

2. Independence of the errors (serial correlation)

2. Independence of the errors (serial correlation)

Indicates that there is room for improvement in our model – there is some association in the data that we are not taking into account.

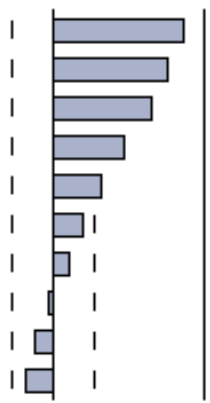
2. Independence of the errors (serial correlation)

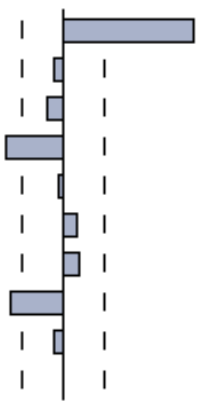
Indicates that there is room for improvement in our model – there is some association in the data that we are not taking into account.

Our tool for this is called an ***Autocorrelation*** *plot*

Autocorrelation and Partial Autocorrelation

Included observations: 50

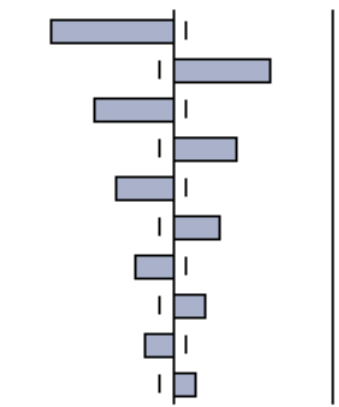




	AC	PAC	Q-Stat	Prob
1	0.894	0.894	42.382	0.000
2	0.787	-0.057	75.953	0.000
3	0.673	-0.099	101.00	0.000
4	0.496	-0.382	114.91	0.000
5	0.339	-0.027	121.56	0.000
6	0.217	0.098	124.35	0.000
7	0.114	0.113	125.14	0.000
8	-0.019	-0.351	125.17	0.000
9	-0.114	-0.052	125.98	0.000
10	-0.185	0.005	128.22	0.000

Autocorrelation and Partial Autocorrelation

Included observations: 500



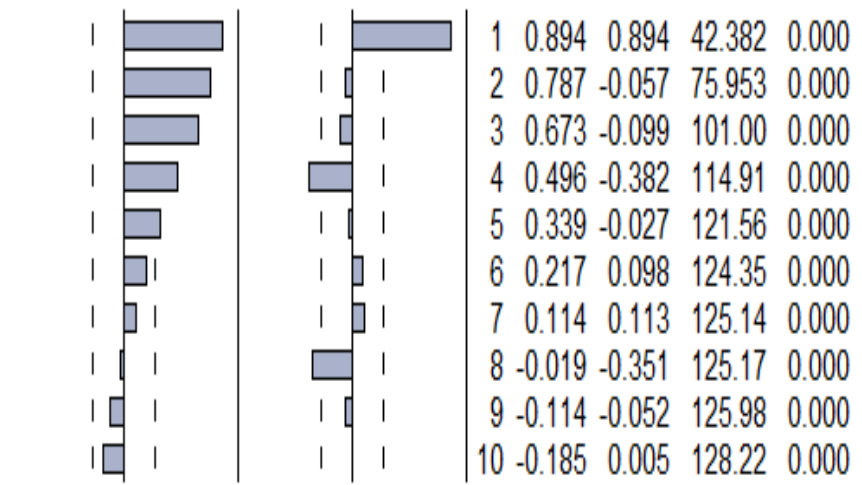
Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.796	-0.796	318.73	0.000
		2	0.641	0.019	525.64	0.000
		3	-0.502	0.036	653.11	0.000
		4	0.415	0.057	740.26	0.000
		5	-0.362	-0.056	806.54	0.000
		6	0.297	-0.047	851.22	0.000
		7	-0.240	0.011	880.44	0.000
		8	0.208	0.050	902.60	0.000
		9	-0.177	0.012	918.59	0.000
		10	0.143	-0.028	929.08	0.000

Autocorrelation and Partial Autocorrelation

Autocorrelation

– A variable's relationship with itself in a previous period

Included observations: 50



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.894	0.894	42.382	0.000
		2	0.787	-0.057	75.953	0.000
		3	0.673	-0.099	101.00	0.000
		4	0.496	-0.382	114.91	0.000
		5	0.339	-0.027	121.56	0.000
		6	0.217	0.098	124.35	0.000
		7	0.114	0.113	125.14	0.000
		8	-0.019	-0.351	125.17	0.000
		9	-0.114	-0.052	125.98	0.000
		10	-0.185	0.005	128.22	0.000

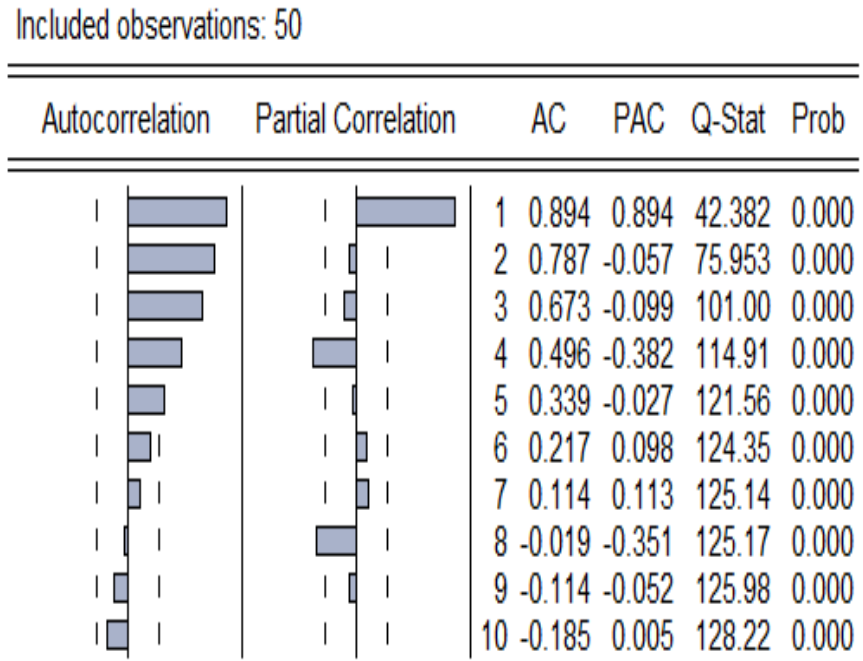
Autocorrelation and Partial Autocorrelation

Autocorrelation

– A variable's relationship with itself in a previous period

Partial Autocorrelation

– Marginal autocorrelation



Included observations: 50

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.894	0.894	42.382	0.000
		2	0.787	-0.057	75.953	0.000
		3	0.673	-0.099	101.00	0.000
		4	0.496	-0.382	114.91	0.000
		5	0.339	-0.027	121.56	0.000
		6	0.217	0.098	124.35	0.000
		7	0.114	0.113	125.14	0.000
		8	-0.019	-0.351	125.17	0.000
		9	-0.114	-0.052	125.98	0.000
		10	-0.185	0.005	128.22	0.000

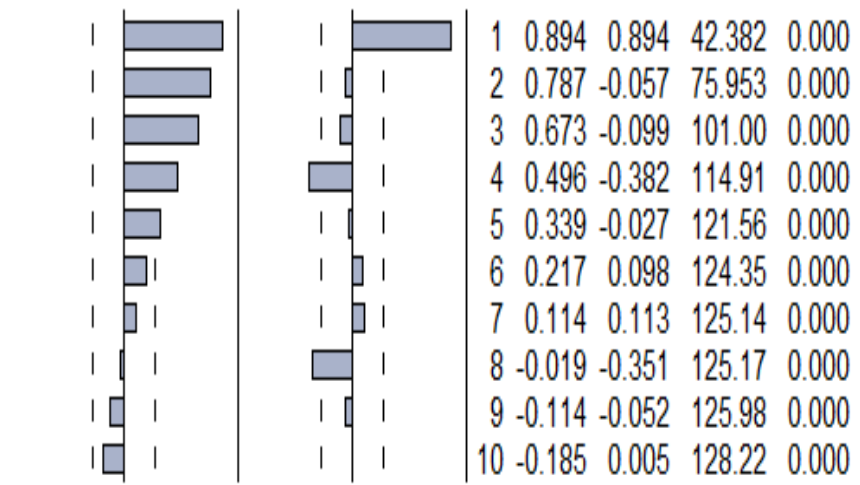
Autocorrelation and Partial Autocorrelation

Interpretation

| | = confidence intervals (95%)

Most of the lags falling between the confidence intervals indicates that our model appropriately reflects the autoregressive nature of our data

Included observations: 50



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.894	0.894	42.382	0.000
		2	0.787	-0.057	75.953	0.000
		3	0.673	-0.099	101.00	0.000
		4	0.496	-0.382	114.91	0.000
		5	0.339	-0.027	121.56	0.000
		6	0.217	0.098	124.35	0.000
		7	0.114	0.113	125.14	0.000
		8	-0.019	-0.351	125.17	0.000
		9	-0.114	-0.052	125.98	0.000
		10	-0.185	0.005	128.22	0.000

Autocorrelation and Partial Autocorrelation

Lab: Sunspots data ACF and PACF (correlograms)

IV. AUTOREGRESSIVE MODELS

Autoregressive models

Autoregressive models

What do they do?

Autoregressive models

What do they do?

They use the value at time -1 to predict the value at time 0

Basic linear regression

$$Y = \alpha + \beta x + \varepsilon$$

Basic linear regression

$$Y = \alpha + \beta x + \varepsilon$$

Multivariate regression

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Autoregressive regression AR(1)

$$Y_i = \Phi_0 + \Phi_1 Y_{i-1} + \varepsilon_i$$

Y_i = *target variable*
Φ_0 = *intercept (~ a, constant)*
Φ_1 = *coefficient (~ β)*
Y_{i-1} = *lagged variable (~ x)*
ε_i = *error*

Autoregressive regression AR(1)

$$Y_i = \Phi_0 + \Phi_1 Y_{i-1} + \varepsilon_i$$

Autoregressive regression AR(3)

$$Y_i = \Phi_0 + \Phi_1 Y_{i-1} + \Phi_2 Y_{i-2} + \Phi_3 Y_{i-3} + \varepsilon_i$$

Further reading:

Stationarity

Unit root processes

MA (Moving average) regressions

ARMA (Autoregressive moving average) regressions

ARCH (Autoregressive conditional heteroscedasticity)

EGARCH (Exponential generalized ARCH)