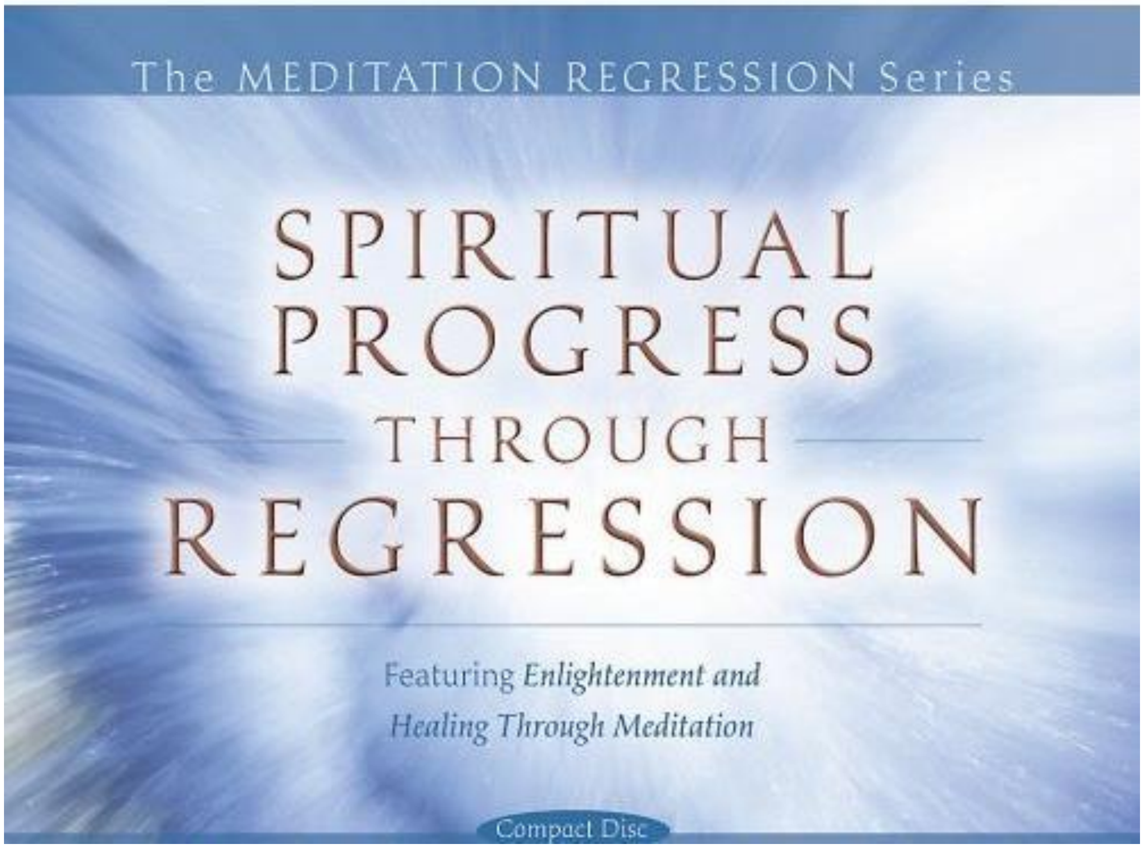
Modeling Autocorrelation: AR models, two-layer models, and evaluating predictability



Forecasting Analytics

Prof. Galit Shmuéli

Autocorrelation

Autocorrelation measures how strong the values of a time series are related to their own past values

Technically: compute the **correlation** between the series and the lagged series (approximately)

Lag(1) autocorrelation = correlation between (y_1, $y_{2,}$ …, y_{t-1}) and (y_2,y_3,…, y_t)
Lag(k) autocorrelation = correlation between (y_1, $y_{2,}$ …, y_{t-k}) and (y_{k+1},y_{k+2},…,y_t)

Note: autocorrelation measures **linear** relationship

Uses of autocorrelation

- Check forecast errors for independence
- Model remaining information
- Evaluate predictability

Example Recap: Coca Cola Sales

Coca Cola.xls contains quarterly sales of Coca Cola (in millions of $) from Q1-86 to Q2-96

Possible Goals:

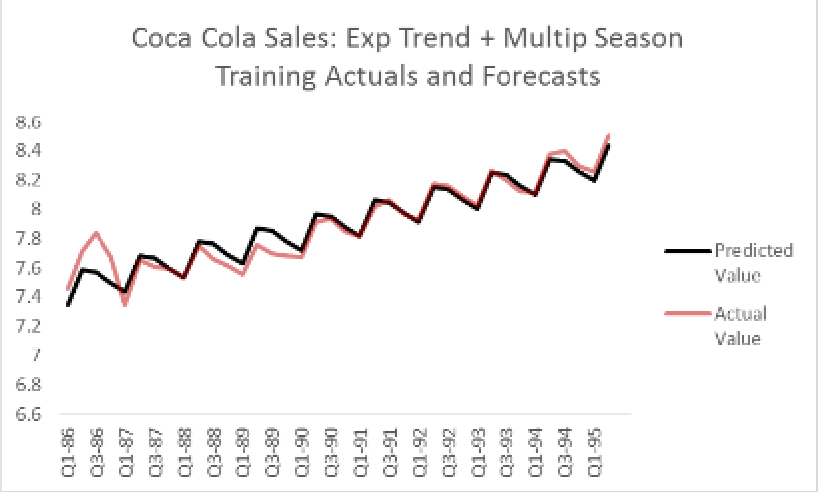
- Time series forecasting: *create forecasts* for the next 4 quarters
- Time series analysis: *quantify* the components (patterns and noise)

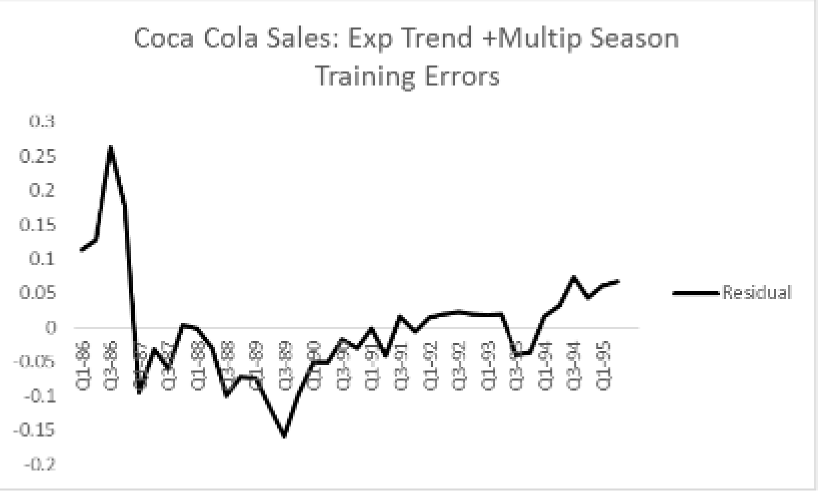
Multiplicative model for Coca Cola sales

$$\log(y_t) = a + bt + b_2 D_2 + b_3 D_3 + b_4 D_4$$

Regression Model

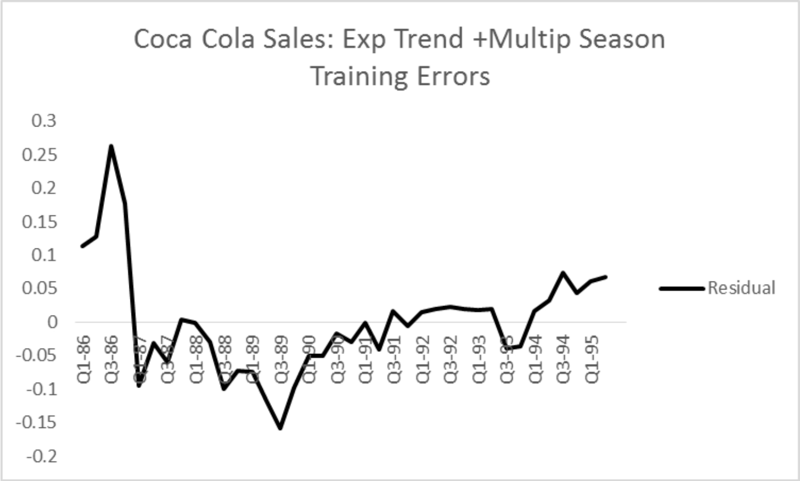
Input Variables	Coefficient	Std. Error	t-Statistic	P-Value
Intercept	7.32222	0.03637944	201.2735679	1.46E-52
t	0.023603	0.001272756	18.5450182	5.08E-19
Quarter index_2	0.218458	0.03845744	5.680501478	2.47E-06
Quarter index_3	0.18125	0.03948962	4.589817795	6.14E-05
Quarter index_4	0.082226	0.039510125	2.081141391	0.045257



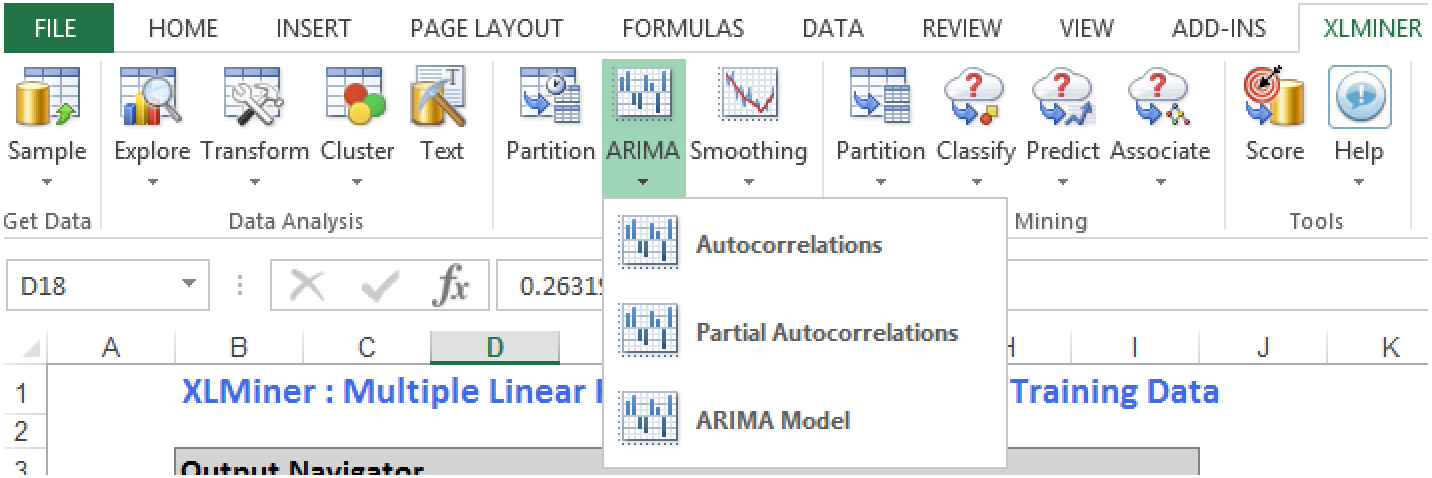


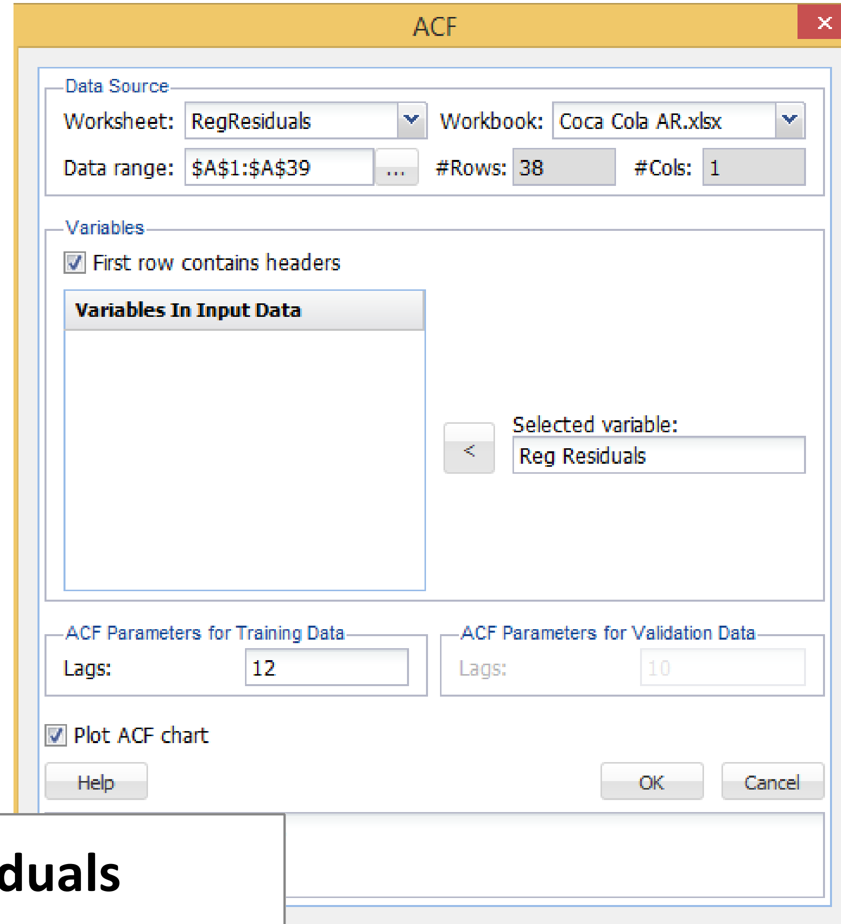
Measuring autocorrelation

Forecast errors from multiplicative seasonal model (Coca Cola sales)



XLMiner: ARIMA > Autocorrelations

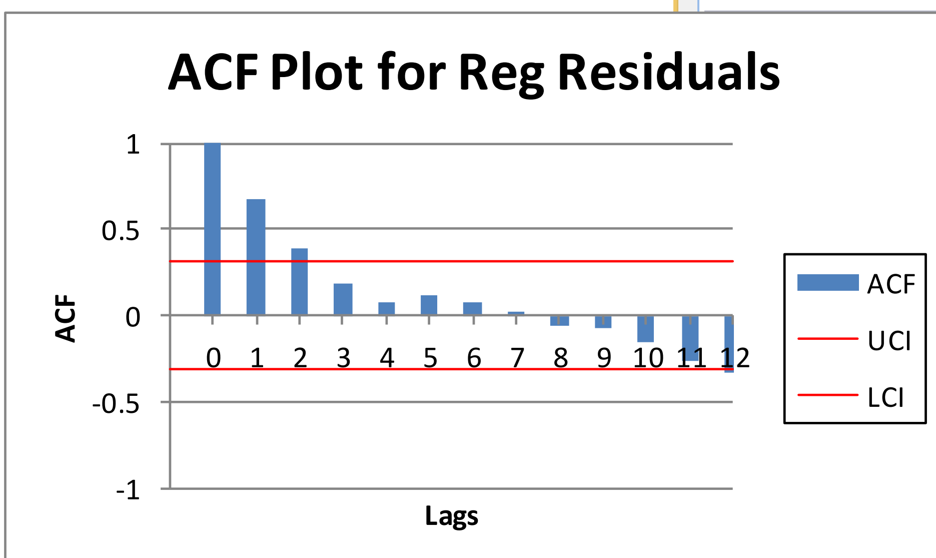




How to model the remaining structure?

ACF Values

Lags	ACF
0	1
1	0.673291
2	0.387009
3	0.179345
4	0.077985
5	0.111203
6	0.081314
7	0.021778
8	-0.05407
9	-0.07948
10	-0.15333
11	-0.26495
12	-0.33681



Autocorrelation – cont.

Positive lag-1 autocorrelation (“**stickiness**”):

high values usually immediately follow ______ values, and low values usually immediately follow ______ values

Negative lag-1 autocorrelation (“**swings**”):

high values usually immediately follow ______ values, and low values usually immediately follow ________ values

High positive autocorrelation at multiples of a certain lag (e.g. lags 4, 8, 12…) indicates ___________.

What to do?

Option 1: multi-layer model

- Model the forecast errors, by treating them as a time series
- Then examine autocorrelation of “errors of forecast errors”
  - If random, stop there, and forecasts using the sum of the sub-forecasts
  - If autocorrelated, continue modeling the level-2 errors (not practical)

Option 2: model the dependence directly (AR and ARIMA models)

Autoregressive (AR) models for modeling forecast errors

1. Use any method to generate forecasts (regression, smoothing)
2. Examine forecast errors for autocorrelation (time plot of forecast errors ; ACF plot)

If autocorrelation exists, fit an AR model to the forecast errors series

Note: AR model can also be fit to original data (more complicated)

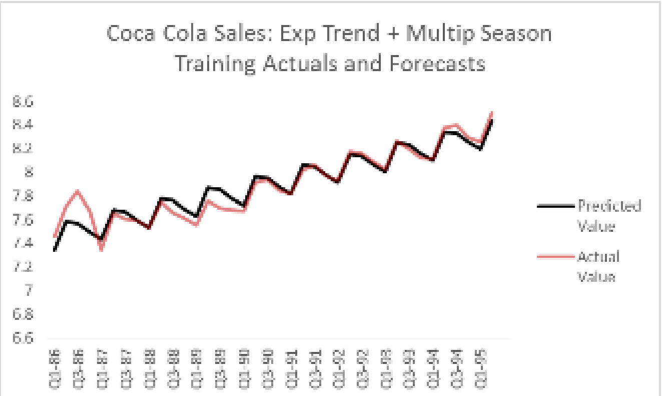
MLR_Output3

Coca Cola Sales: Multiplicative regression-based model (exp trend)

$$\log(y_t) = \beta_0 + \beta_1 t + \beta_2 D_2 + \beta_3 D_3 + \beta_4 D_4 + \varepsilon$$

Regression Model

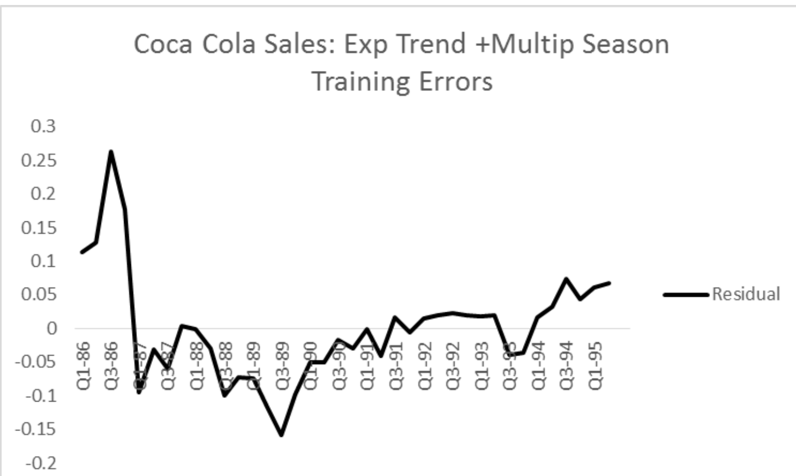
Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	7.32222	0.03637944	201.2735679	1.46E-52	7.248205	7.396234	2373.006
t	0.023603	0.001272756	18.5450182	5.08E-19	0.021014	0.026193	2.576354
Quarter index_2	0.218458	0.03845744	5.680501478	2.47E-06	0.140215	0.2967	0.13176
Quarter index_3	0.18125	0.03948962	4.589817795	6.14E-05	0.100908	0.261592	0.1237
Quarter index_4	0.082226	0.039510125	2.081141391	0.045257	0.001842	0.16261	0.031993



ACF_Output

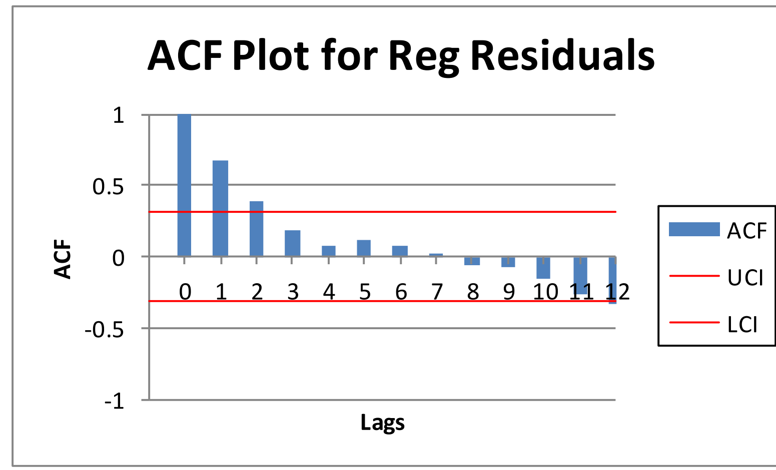
Regression captures seasonality and trend

But forecast errors exhibit autocorrelation



ACF Values

Lags	ACF
0	1
1	0.673291
2	0.387009
3	0.179345
4	0.077985
5	0.111203
6	0.081314
7	0.021778
8	-0.05407
9	-0.07948
10	-0.15333
11	-0.26495
12	-0.33681



Positive autocorrelation at lag 1 (stickiness) and lag2

Autoregressive (AR) Models

The idea: Model the autocorrelation directly in regression model, using past observations as predictors

Useful in modeling time series of **forecast errors** (2-level model)

Example: Suppose series exhibits autocorrelation at lags 1 and 2

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \varepsilon_t$$

Called an AR model of order 2, or AR(2)

Fitting & Using AR models

AR model estimation is roughly similar to linear regression, with the lagged series as predictors

Some software will fit an AR directly (XLMiner)

Use the output

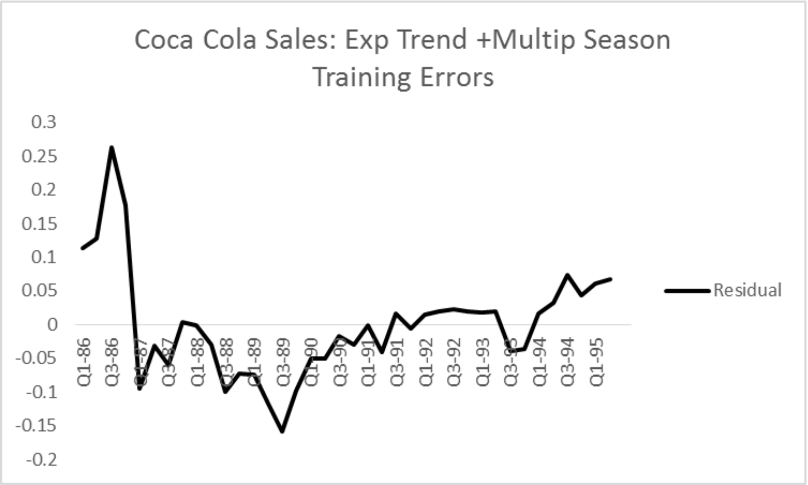
- To estimate coefficients, std error of estimate, etc.
- To forecast. Example for an AR(2) :

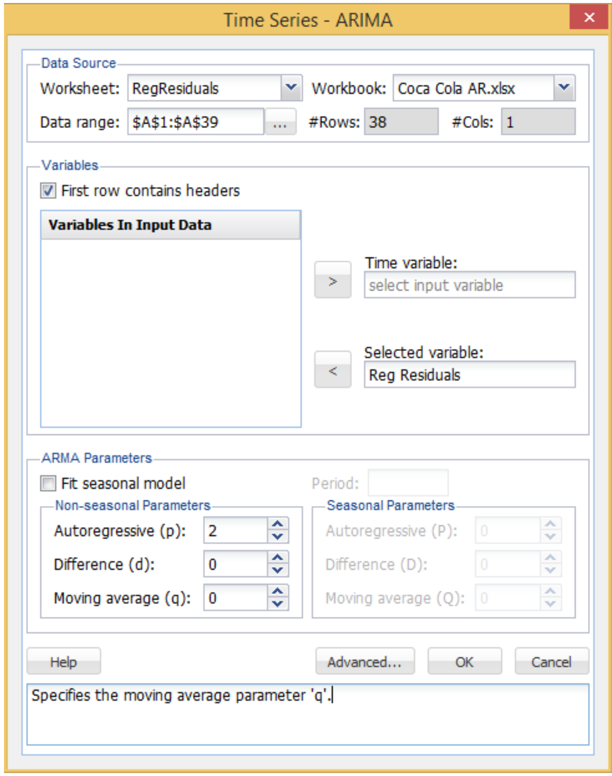
1-step ahead forecast: $F_{t+1} = a + b_1 Y_t + b_2 Y_{t-1}$
2 steps ahead $F_{t+2} = a + b_1 F_{t+1} + b_2 Y_t$
3 steps ahead $F_{t+3} = a + b_1 F_{t+2} + b_2 F_{t+1}$

IMPROVING PREDICTIVE ACCURACY: MODELING FORECAST ERRORS USING AR MODELS

Fitting an AR(2) Model in XLMiner

Example: residuals from model 4 ($R_1 \ldots R_{38}$)





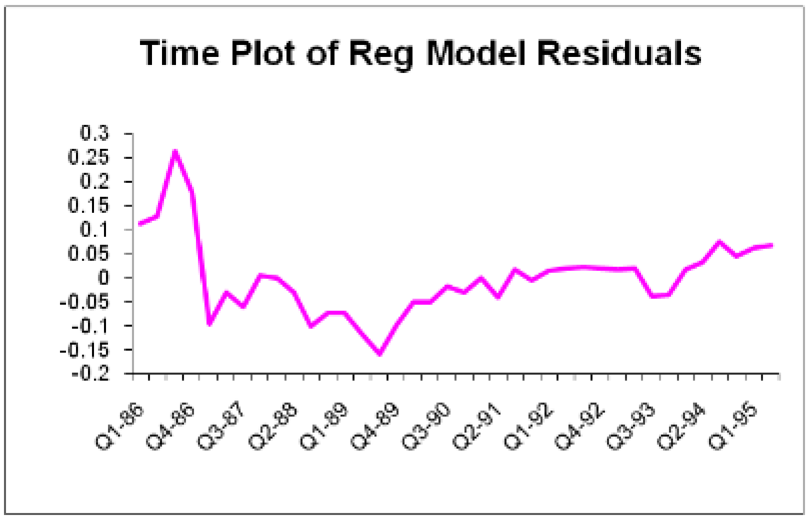
XLMiner> *ARIMA > ARIMA Model*

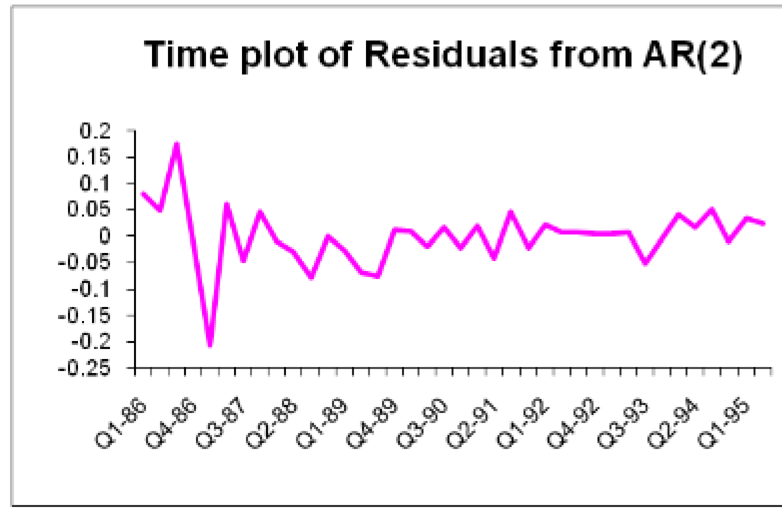
ARIMA = AutoRegressive-Integrated-Moving-Average

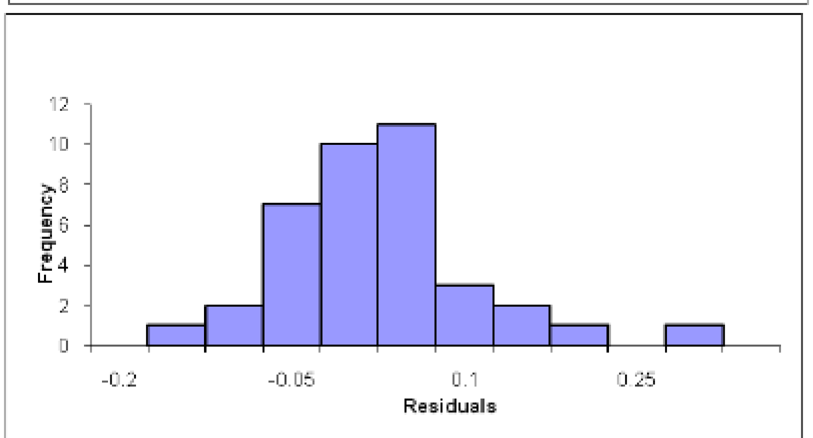
Set *autoregressive parameter* = 2

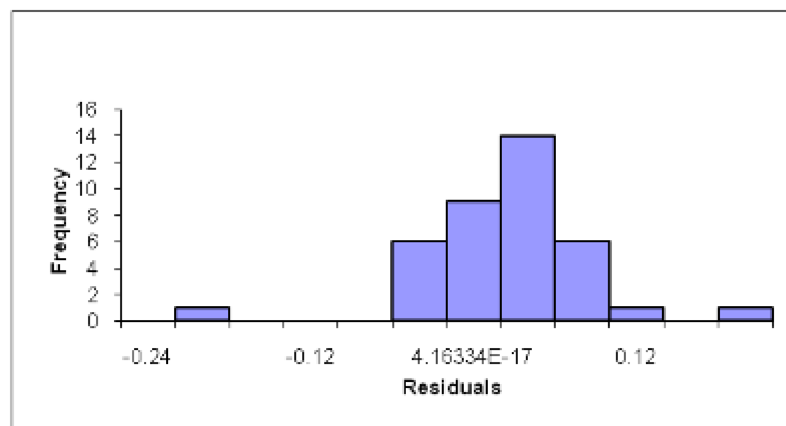
Forecast errors before AR(2)

Forecast errors after AR(2) (errors of forecast errors)









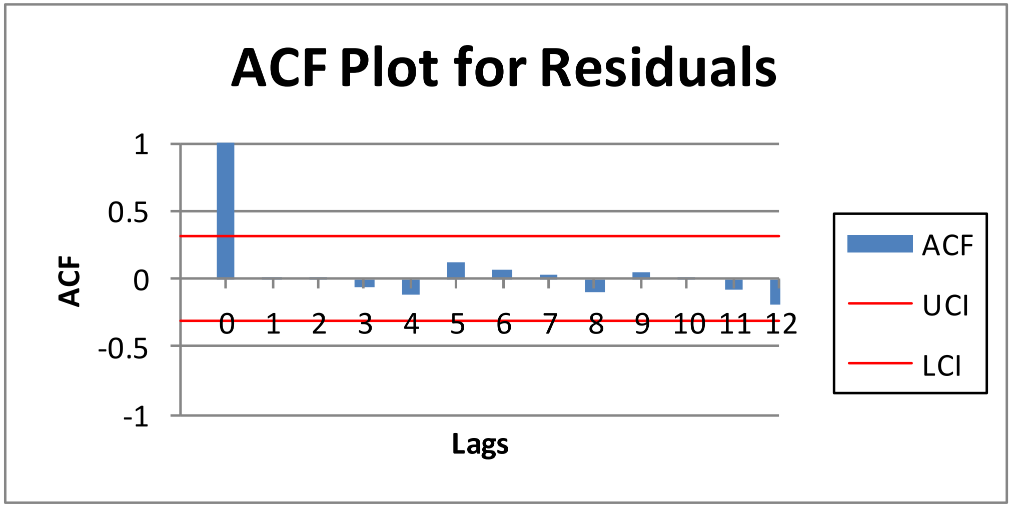
ARIMA_Residuals

ACF_Output1

Any autocorrelation left?
Compute autocorrelation of errors-of-errors

ACF Values

Lags	ACF
0	1
1	0.012059
2	0.019801
3	-0.06917
4	-0.11583
5	0.120415
6	0.056936
7	0.025882
8	-0.10121
9	0.048068
10	0.017034
11	-0.08724
12	-0.20016



ARIMA_Output

AR(2) Output

Inputs

Data	
Workbook	Coca Cola AR.xlsx
Worksheet	RegResiduals
Range	$A$1:$A$39
Selected Variable	Reg Residuals
# Records in Input Data	38

Parameters/Options	
AR	2
MA	0
Ordinary Difference	0
Show Var/Covar Output	No
Show Forecasting Output	No
#Forecasts	N.A.
Confidence Level	N.A.
Show Residual Output	Yes

ARIMA Model

ARIMA	Coeff	StErr	p-value
Const. term	1.03381E-15	0.004257	1
AR1	0.767113049	0.298655	0.010212
AR2	-0.09474876	0.298655	0.751053

Mean	3.15537E-15
-2LogL	-109.059763
Res. StdDev	0.059307746
#Iterations	109

Ljung-Box Test Results on Residuals

Lag	12	24	36
p-Value	0.999395277	0.999898	1
ChiSq	5.111254761	7.967949	19.11468
df	10	22	34

Forecasting with 2-level models

Level 1: Model /method applied to raw data, produces forecasts + forecast errors

Level 2: AR applied to forecast errors, produces errors-of-forecast-errors

Piecing it together: getting improved forecasts

Use level 1 to forecast next value of series F_{t+1}

Use AR to forecast next **forecast error (**residual) $\hat{E}_{t+1}$

Combine the two to get an improved forecast F^*_{t+1} :

$$F^*_{t+1} = F_{t+1} + \hat{E}_{t+1}$$

Example: Forecasting Q3-95

Level 1: Forecast Q3-95 sales from Model 4:

$$\log(F_{39}) = 7.322 + (0.0236)(39) + (0.181)(1) = 8.4240$$
$$F_{39} = e^{8.4240} = \$4555.077 \text{ million}$$

Level 2: Forecast of Q3-95 **forecast error** (assuming that the above model 4 is used):

$$\hat{E}_{39} = -0.000 + (0.767)(R_{38}) - (0.095)(R_{37}) =$$
$$-0.000 + (0.767)(0.0667) - (0.095)(0.0613) = 0.04536$$

Combine: Improved forecast based on both:

$$\log(F_{39}^{*}) = 8.4240 + 0.0454 = 8.4694$$
$$F_{39}^{*} = \$4766.466 \text{ million}$$

When does a 2^{nd} AR layer make sense?

Depends on the forecast horizon!

EVALUATING PREDICTABILITY

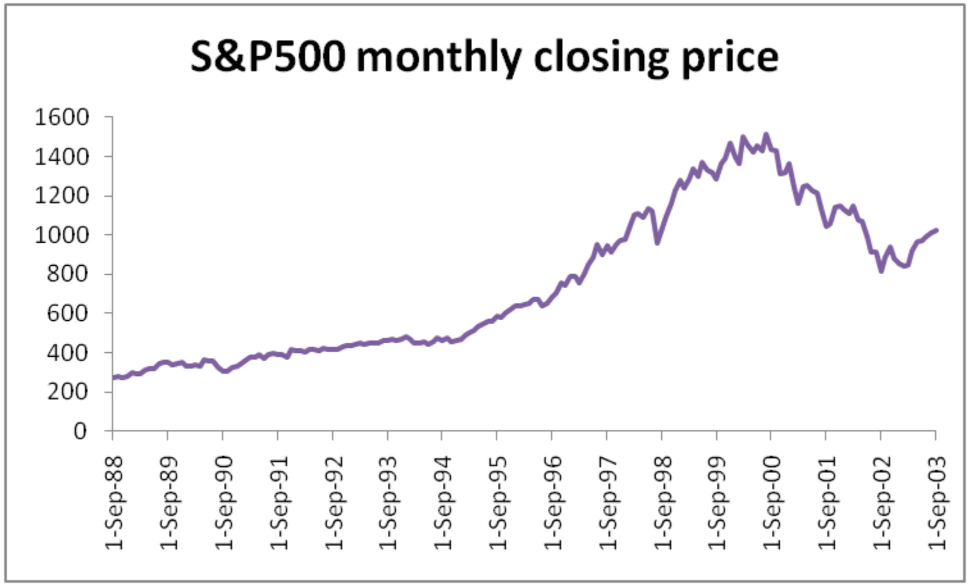
Figuring out whether the forecasting effort is useful:

should we go beyond naïve forecasts?

Example: S&P500

The S&P500 index in the last 15 years (finance.yahoo.com) – see S&P500.xls

Monthly closing values from 9/1988-8/2003



Goal: Forecast Sept 2003

Random Walk: Special Case of AR(1)

Random walk = AR(1) model with *b*=1

$$Y_t = a + Y_{t\text{-}1} + \varepsilon_t$$

a = **drift parameter**

σ (std of ε) = **volatility**.

In a random walk the *changes* from one period to the next are *random*

Does the series behave like a random Walk? (Predictability test)

Option 1: test the hypothesis β=1.

Option 2:

Series: $Y_1, Y_2, \ldots, Y_T$

Differenced Series: $Y_2-Y_1, Y_3-Y_2, \ldots, Y_T - Y_{T-1}$

If original series is a random walk, then differenced series behaves like a random series (but its mean can be non-zero)

Equation for differenced series: $Y_t - Y_{t-1} = a + \varepsilon_t$

Moral: see if the differenced series is random! (check autocorrelations)

ARIMA_last100days, Differenced, ACF_Diff

Example: S&P500 – with XLM

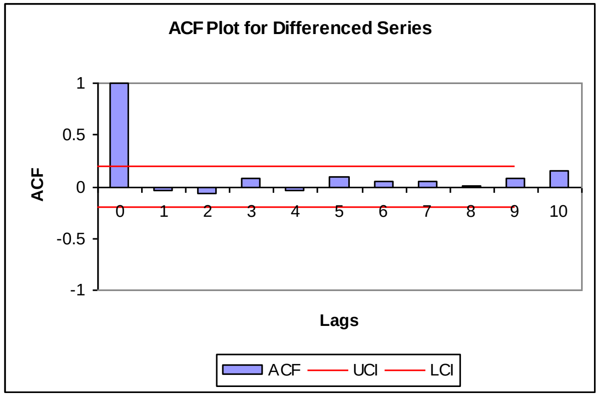
- Look at last 100 months
- *ARIMA>ARIMA Model* with AR coefficient = 1
- Compute differenced series, and then *Time Series > ACF*

ARIMA Model

ARIMA	Coeff	StErr	p-value
Const. term	15.62566853	3.68750787	0.00002261
AR1	0.98479182	0.01436355	0

Current XLMiner rounds StErr to 0

Close	Differenced values
271.91	7.06
278.97	-5.28
273.69	4.03
277.72	19.75
297.47	-8.61
288.86	6.01
294.87	14.77
309.64	10.88



Estimating the drift and volatility

Using the series:

- How can we estimate the drift (a)?

- How can we estimate the volatility (σ)?

Compute these for the S&P500 series.

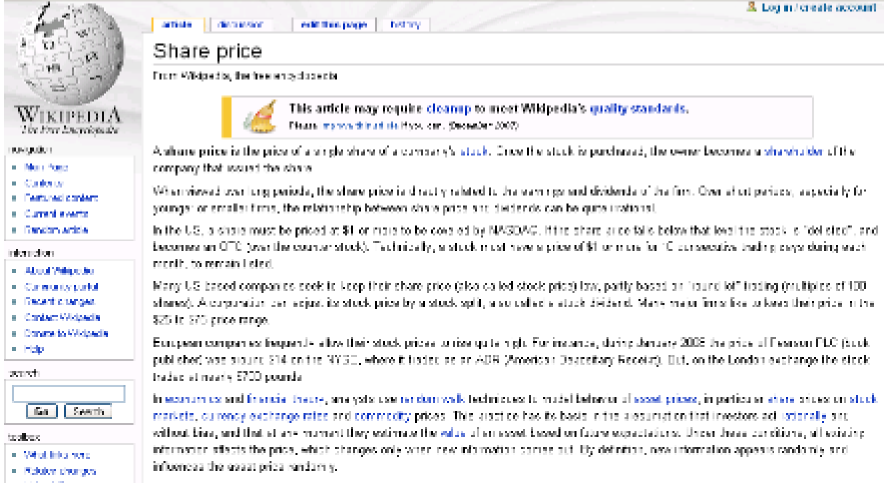
Forecasting a Random Walk

Estimated by average of differenced series

- One-step-ahead forecast: $F_{t+1} = a + Y_t$
- Two-step-ahead forecast: $F_{t+2} = a + Y_{t+1} = 2a + Y_t$
- *k*-step-ahead forecast : $F_{t+k} = ka + Y_t$

- If the drift parameter is 0, then the *k*-step-ahead forecast is F_{t+k} = _____ for all *k.*
- Economic implications of a random walk: **The Efficient Market Hypothesis**
  - At any given time, security prices fully reflect all available information; buying and selling securities in an attempt to outperform the market will effectively be a game of chance rather than skill

Stocks and random walks..



Wikepedia > Share price
(http://en.wikipedia.org/wiki/Share_price)

In economics and financial theory, analysts use random walk techniques to model behavior of asset prices, in particular share prices on stock markets, currency exchange rates and commodity prices. This practice has its basis in the presumption that investors act rationally and without bias, and that at any moment they estimate the value of an asset based on future expectations. Under these conditions, all existing information affects the price, which changes only when new information comes out. By definition, new information appears randomly and influences the asset price randomly.

Empirical studies have demonstrated that prices do not completely follow random walk. Low serial correlations (around 0.05) exist in the short term; and slightly stronger correlations over the longer term. Their sign and the strength depend on a variety of factors, but transaction costs and bid-ask spreads generally make it impossible to earn excess returns.

FROM AR MODELS TO ARIMA MODELS

From AR to ARMA

AR(p) model:

$$Y_t = \alpha + \beta_1 Y_{t\text{-}1} + \dots + \beta_\text{p} Y_{t\text{-p}} + \varepsilon_t$$

Autoregressive Moving Average (ARMA(p,q)) model:

$$Y_t = \alpha + \beta_1 Y_{t\text{-}1} + \dots + \beta_\text{p} Y_{t\text{-p}} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q}$$

The idea: add lags of the series and/or lags of the forecast errors to capture all forms of autocorrelation

… and to ARIMA

Autoregressive Moving Average (ARMA(p,q)) model:

$$Y_t = \alpha + \beta_1 Y_{t\text{-}1} + \ldots + \beta_{\mathrm{p}} Y_{t\text{-}\mathrm{p}} + \varepsilon_t - \theta_1 \varepsilon_{t\text{-}1} - \ldots - \theta_q \varepsilon_{t\text{-}q}$$

ARIMA (p,d,q) = Autoregressive Integrated Moving Average

same as ARMA applied to the *d*-times differenced series (difference the series at lag 1 over and over again, *d* times)

The idea: add lags of the differenced series and/or lags of the forecast errors

Using ARIMA models

- Require expertise (two-step identification/estimation process)
- More volatile
- Require stationary series = data with no patterns (trend, seasonality)
- Some software do “automated ARIMA” – highly sensitive to optimization setup (R forecast package has *auto_arima*)
- Less popular with management

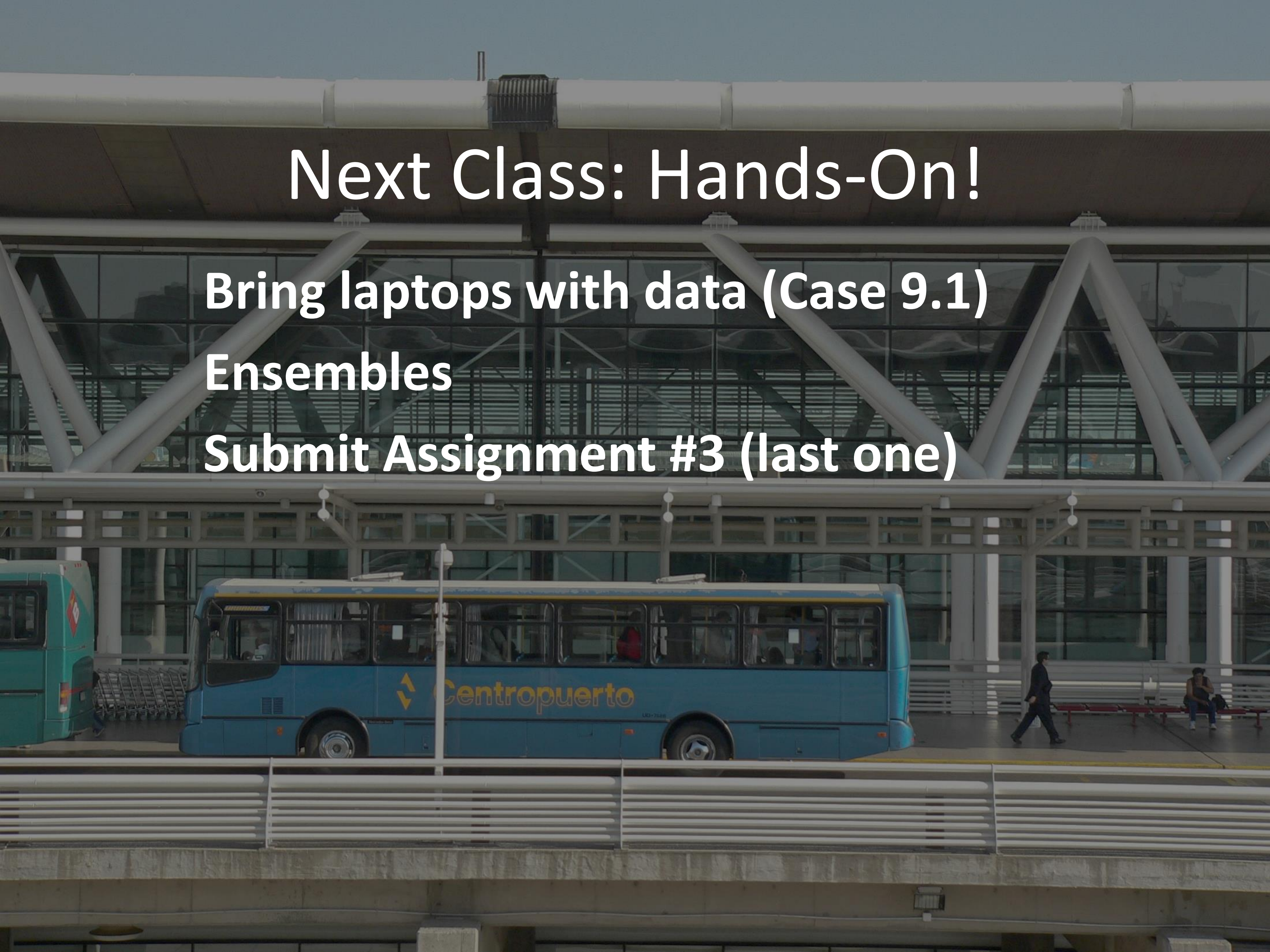
Summary

Autoregressive models are useful as second-level models of forecast errors (residuals), *in some applications*

They capture/model autocorrelation directly

AR(1) models help evaluate predictability of series beyond naïve forecasts

ARIMA models are more complicated; less popular with management (blackbox); careful choice of software



Next Class: Hands-On!

Bring laptops with data (Case 9.1)

Ensembles

Submit Assignment #3 (last one)