Chapter 16 – Regression Based Forecasting
基于回归的预测

Instructor: Zach Zhizhong ZHOU,
Shanghai Jiao Tong University
主讲教师：周志中，上海交通大学

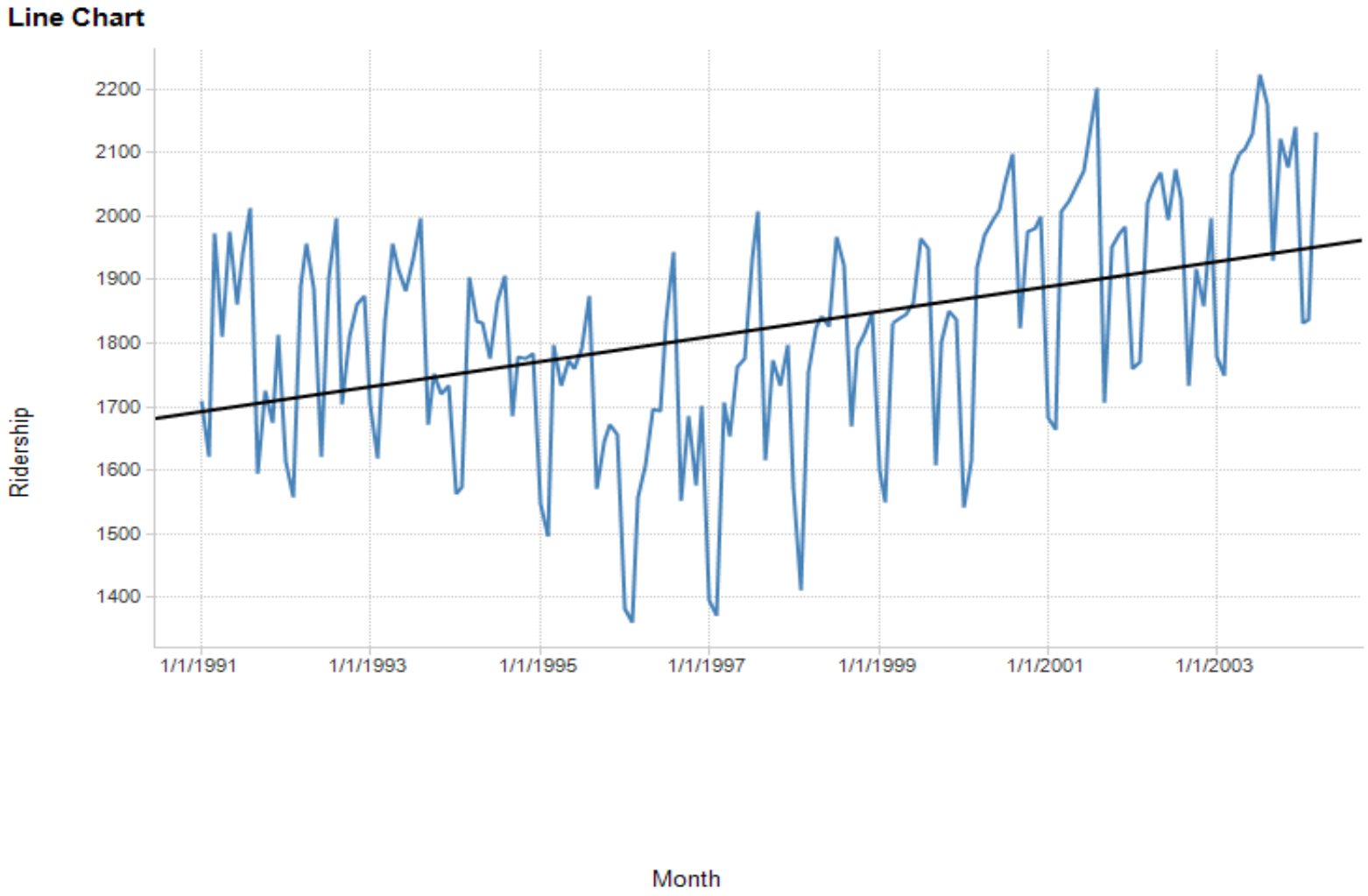
Data Mining for Business Intelligence

Shmueli, Patel & Bruce

Main ideas 主要思想

- Fit linear trend, time as predictor 拟合线性趋势，使用时间作为预测因子。
- Modify & use also for non-linear trends 模型做更改之后也可以用于对非线性趋势的预测
  - Exponential 指数趋势
  - Polynomial 多项式趋势
- Can also capture seasonality 也可以用来捕捉季节性

Linear fit to Amtrak ridership data
(Doesn’t fit too well – more later)



The regression model 回归模型

Ridership Y is a function of time (t) and noise (error = e)

$$Y_i = B_0 + B_1 * t + e$$

Thus we model 3 of the 4 components:
我们模型4个要素中的3个

- Level (B_0) 水平
- Trend* (B_1) 趋势
- Noise (e) 噪音

*Our trend model is linear, which we can see from the graph is not a good fit (more later) 我们的趋势模型是线性的，从图上看拟合效果不好。

XLMiner Output
(see chapter on multiple linear regression for details on RMS error, etc.)

The Regression Model

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	1713.028809	27.08552361	0	477456500
t	1.2053107	0.31751993	0.00021544	384546.3125

Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
3869551.676	162.2451256	-3.84852E-05

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
529326.616	210.0251207	168.8524156

Exponential Trend 指数趋势

Appropriate model when increase/decrease in series over time is multiplicative 合适的模型是增长率/减少率在大段时间内的一小段时间是常数。

e.g., t_1 is x% more than t_0, t_2 is x% more than t_1…

Replace Y with log(Y), then fit linear regression 将Y替换为log(Y)，然后用线性模型进行拟合

$$log(Y_i) = B_0 + B_1 t + e$$

Forecast errors 预测错误

Note that performance measures in standard linear regression software are not in original units 注意到线性模型中衡量预测表现的指标并未针对原来的被预测变量。

Model forecasts will be in the form log(*Y*) 模型的预测值是log(Y)的形式

Return to original units by taking exponent of model forecasts 将模型的预测值取指数得到原来被预测变量的预测值。

Calculate standard deviation of these forecast errors to get RMSE 计算与被预测变量实际值之间的差（预测错误）的标准差以得到RMSE。

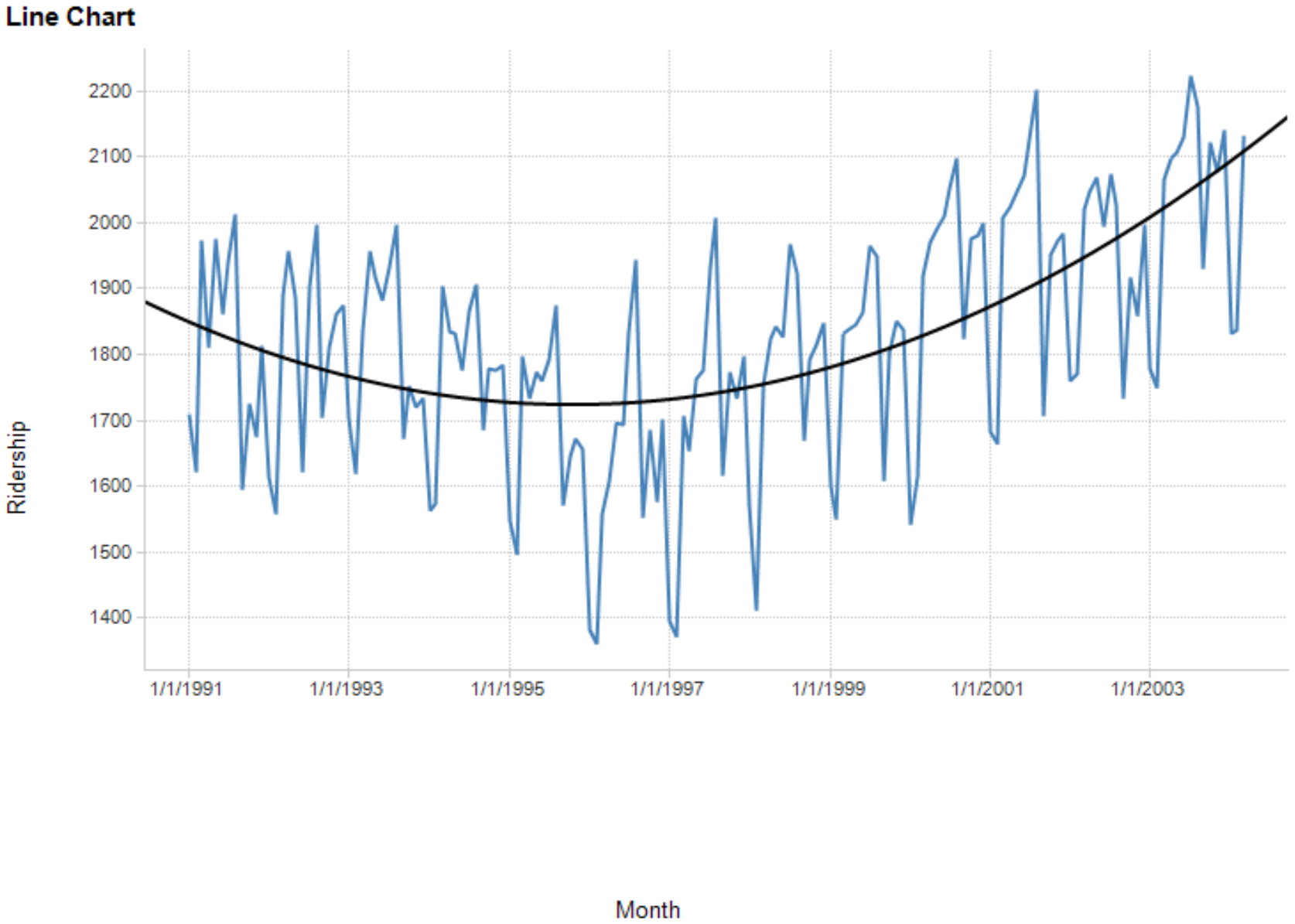
Polynomial Trend 多项式趋势

Add additional predictors as appropriate 增加合适的新的预测变量。

For example, for quadratic relationship add a t^2 predictor 例如，对二项式关系加上一个t^2作为预测因子。

Fit linear regression using both t and t^2 对于线性回归同时使用t和 t^2

Quadratic fit to Amtrak data



Quadratic fit to Amtrak Data

Now appears to capture trend 看起来捕捉住了趋势

Seasonality remains 但季节性未被捕捉到

Handling Seasonality 处理季节性

- Seasonality is any recurring cyclical pattern of consistently higher or lower values (daily, weekly, monthly, quarterly, etc.) 季节性是反复出现的持续的数值高低相间的周期模式（可以天、周、月、季为周期）。

- Handle in regression by adding categorical variable for season, e.g., 加入类别型变量来代表季节，如：

Month	Ridership	Season
Jan-91	1709	Jan
Feb-91	1621	Feb
Mar-91	1973	March
Apr-91	1812	April

Creating Binary dummies 创建二元虚拟变量

Logistic regression software usually requires transforming categorical variables into dummies

To avoid multicollinearity problems, use *m-1* dummies for *m* categories

XLMiner output coefficients for each season

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	1855.235962	33.95079803	0	477456500
season_Aug	139.3903351	48.01367569	0.00431675	483721.3125
season_Dec	-19.82307816	48.01367569	0.68036187	33314.77734
season_Feb	-288.9631348	47.08128357	0	665331.9375
season_Jan	-251.2854462	47.08128357	0.00000034	598841.0625
season_Jul	94.34428406	48.01367569	0.05147372	187691.7656
season_Jun	-10.11090946	48.01367569	0.83352947	11869.09277
season_Mar	11.57308865	47.08128357	0.80620199	48930.94922
season_May	31.24033737	48.01367569	0.51637506	114420.9141
season_Nov	-63.96651077	48.01367569	0.18502063	3121.062012
season_Oct	-54.12883377	48.01367569	0.26158884	14579.31641
season_Sep	-193.6371613	48.01367569	0.00009163	224972.1094

Seasonality types 季节性的类型

Additive – described above (model shows amounts by which seasonal values exceed or fall below those in the reference season) 增量型：如上述（模型显示季节性取值超过或低于被参照季节的数量）。

Multiplicative - (model shows percentages by which seasonal values exceed or fall below those in the reference season) 乘积型：模型显示季节性取值高于或低于被参照季节的百分比。

Proceed as above, but use log(Y) as output 可以使用log(Y)作为被预测变量。

Final model, Amtrak data

Incorporates trend and seasonality

13 predictors

- 11 monthly dummies
- t
- t^2

XLMiner output - coefficients

The Regression Model

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	1932.998779	27.85863113	0	477456500
season_Aug	135.1726227	30.52143288	0.00001955	483721.3125
season_Dec	-29.65872955	30.53801155	0.33320817	33314.77734
season_Feb	-306.3078308	29.94875526	0	665331.9375
season_Jan	-267.444458	29.94642067	0	598841.0625
season_Jul	91.31225586	30.5189991	0.00330446	187691.7656
season_Jun	-12.04474545	30.51724434	0.69370645	11869.09277
season_Mar	-7.04482555	29.95207596	0.81441271	48930.94922
season_May	30.31717491	30.51618195	0.32228076	114420.9141
season_Nov	-72.26641083	30.53282547	0.01938256	3121.062012
season_Oct	-60.98049164	30.52834129	0.04781064	14579.31641
season_Sep	-199.1280975	30.52454758	0	224972.1094
t	-5.246521	0.58674908	0	398979.7188
t^2	0.0437566	0.00384071	0	725213.9375

Model Performance
(superior performance on validation data is unusual)

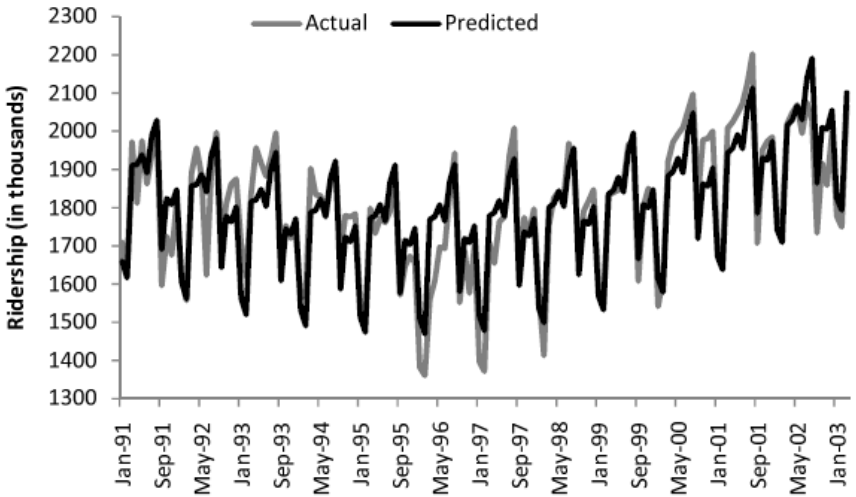
Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
743110.0191	71.0997201	-6.05149E-05

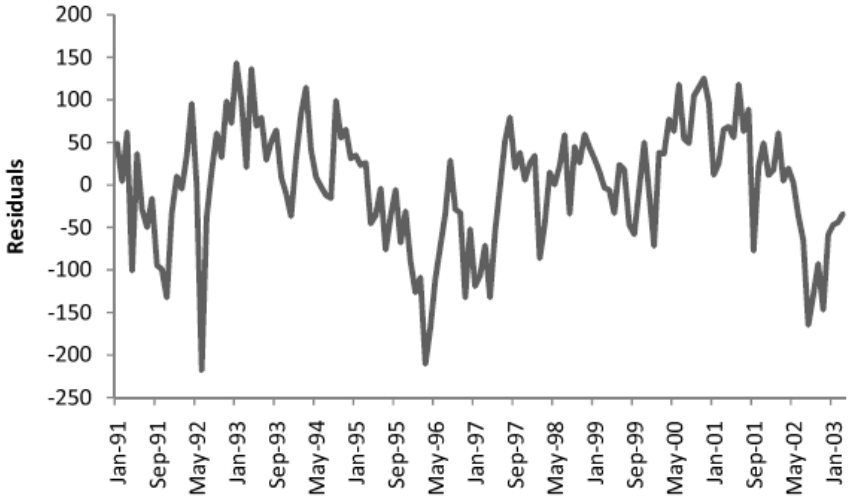
Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
30722.61731	50.59859789	-34.11397564

Actual vs. Predicted



Residuals



Autocorrelation and ARIMA

自相关和ARIMA

Autocorrelation 自相关

Unlike cross-sectional data, time-series values are typically correlated with nearby values ("autocorrelation") 与横截面数据不同，时间序列数据特别会与附近的值有相关性（自相关）

Ordinary regression does not account for this 通常的回归没有将此纳入考虑。

Computing autocorrelation 计算自相关

Create “lagged” series 创建“滞后”时间序列

Copy of the original series, offset by one or more time periods 复制原来的时间序列数据，将其偏置一个或者多个时间段。

Compute correlation between original series and lagged series (lag-1, lag-2, etc.) 计算原时间序列与被偏置之后的时间序列之间的相关性。

Amtrak – original series and Lag-1, Lag-2

TABLE 16.1 FIRST 24 MONTHS OF AMTRAK RIDERSHIP SERIES

Month	Ridership	Lag-1 Series	Lag-2 Series
Jan-91	1709		
Feb-91	1621	1709	
Mar-91	1973	1621	1709
Apr-91	1812	1973	1621
May-91	1975	1812	1973
Jun-91	1862	1975	1812
Jul-91	1940	1862	1975
Aug-91	2013	1940	1862
Sep-91	1596	2013	1940
Oct-91	1725	1596	2013
Nov-91	1676	1725	1596
Dec-91	1814	1676	1725
Jan-92	1615	1814	1676

Autocorrelation 自相关

Positive autocorrelation at lag-1 = stickiness 一期滞后项

Strong autocorrelation (positive or negative) at a lag > 1 indicates seasonal (cyclical) pattern 大于1期滞后项的强自相关（正或者负）表明有强的季节性周期模式。

Autocorrelation in residuals indicates the model has not fully captured the seasonality in the data 残差项如出现自相关则代表模型没有完全捕捉住数据中的季节性因素。

Amtrak residuals – autocorrelation function (ACF) plot

ACF Values

Lags	ACF
0	1
1	0.64821321
2	0.51890093
3	0.40798336
4	0.31966141
5	0.26237851
6	0.21345751
7	0.22334783
8	0.22640951
9	0.19724335
10	0.14933859
11	0.17307311
12	0.12726976

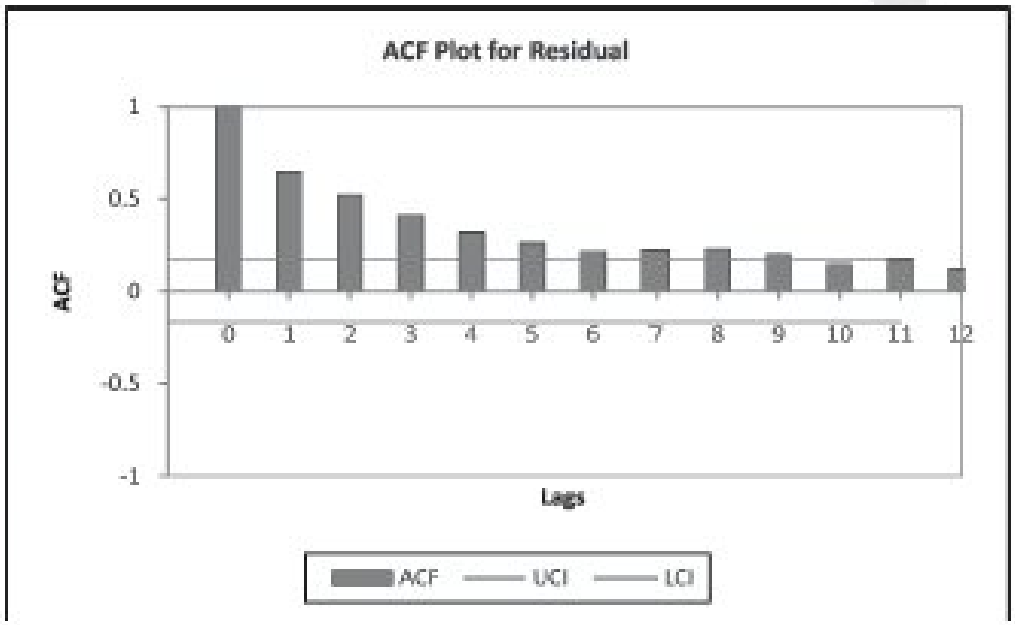


FIGURE 16.11 XLMINER OUTPUT SHOWING AUTOCORRELATION OF RESIDUAL SERIES FROM FIGURE 16.9

Incorporating autocorrelation into models 第二种方法：将自相关引入模型中

Use a forecasting method to forecast *k*-steps ahead 使用一个预测模型预测k-期后的项

Fit AR (autoregressive) model to residuals 用AR模型拟合残差。

Incorporate residual forecasts 考虑对残差的预测

Improved $F_{t+k} = F_{t+k} + E_{t+k}$

Choose order of the AR model
选择AR模型的阶次

If autocorrelation exists at lag-1, a lag-1 model should be sufficient to capture lags at other periods as well 如果原数据中存在滞后1期或者滞后更多期的自回归，那么残差项的滞后1期自回归模型应该足以捕捉到原数据中的滞后1期或者多期的自回归。

$$E_t = B_0 + B_1 E_{t-1} + e$$

Where E_t is residual (forecast error) at time t

In XLMiner

- Fit AR Model, use ARIMA
- In *nonseasonal parameters*, set *Autoregressive (p)* to the required order
- Set *Moving Average (q)* to 0
- Request forecasts, fitted values, residuals via *Advanced* menu

Amtrak residuals – AR(1) model

ARIMA Model

ARIMA	Coeff	StErr	p-value
Const. term	-0.00002147	4.86492062	0.99999648
AR1	0.64688748	0.06221873	0

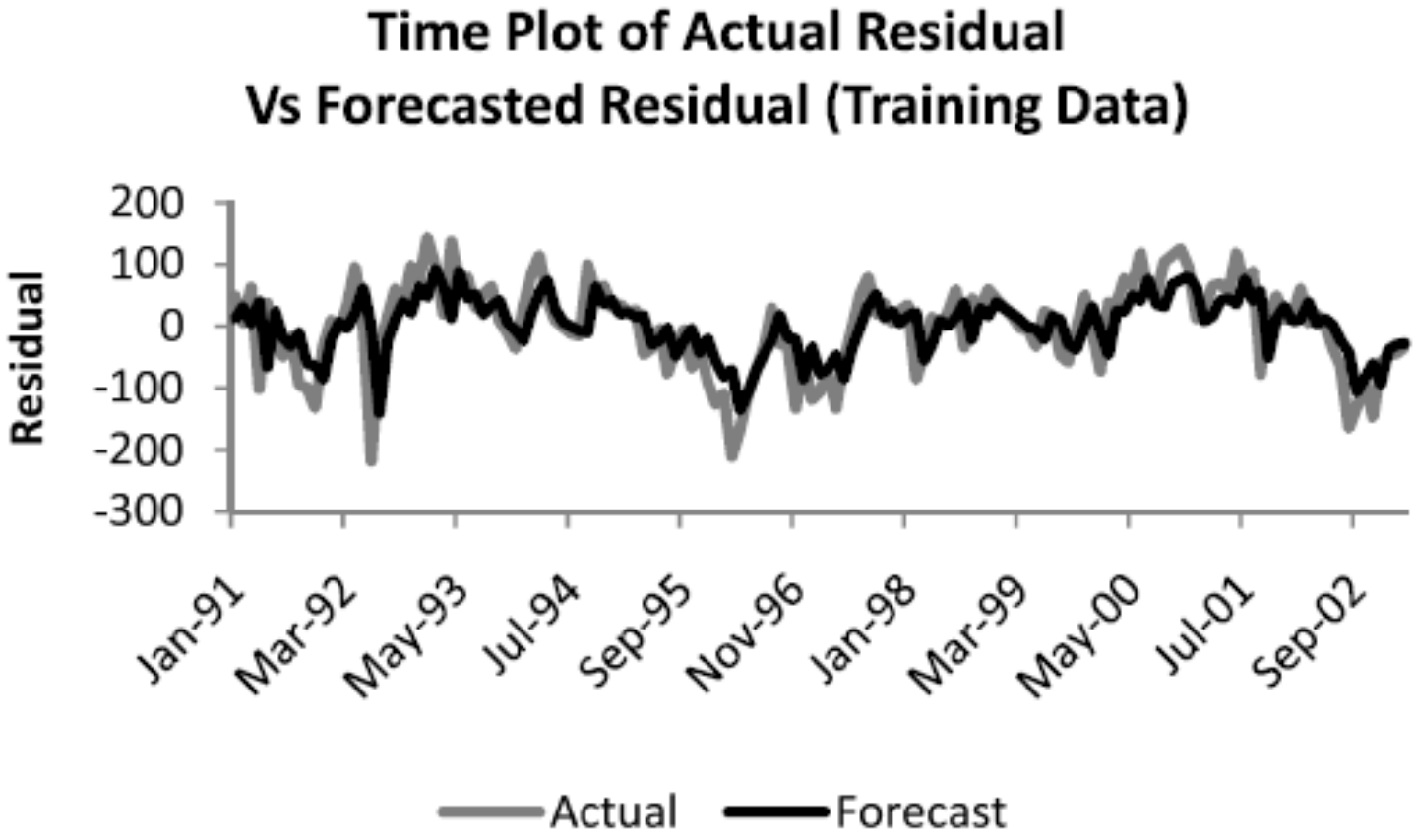
Mean	-0.000061
-2LogL	1590.729736
S	54.047494
#Iterations	8

Lag	12	24	36	48
p-Value	0.577618	0.48778015	0.60872775	0.52462631
ChiSq	9.48063087	22.54175186	32.10258102	45.74367905
df	11	23	35	47

Forecast

Month	Forecast
Apr-03	-21.85559082

AR (1) actual vs. forecast residual



Random walks

- Before forecasting, consider "is the time series predictable?"
- Or is it a random walk?
- Test the hypothesis that *slope = 1* in an AR(1) model (i.e., that the forecast for a period is the most recently-observed value)
- If hypothesis cannot be rejected, series is statistically equivalent to a random walk (i.e., we have not shown that it is predictable).

Summary – Regression Based Forecasting

- Can use linear regression for exponential models (use logs) and polynomials (exponentiation)

- For seasonality, use categorical variable (create dummies)

- Incorporate autocorrelation by modeling it, then using those error forecasts in the main model