Regression Therapy:

To go back in time to the origin or root cause of a problem and then release or heal it. This might entail going back to past lives or earlier parts of your current life.

Regression-Based Forecasting Methods
Part I: Linear Regression

Forecasting Analytics

Prof. Galit Shmuéli

Recall: Time series components

Systematic part

- Level
- Trend
- Seasonal patterns

Non-systematic part

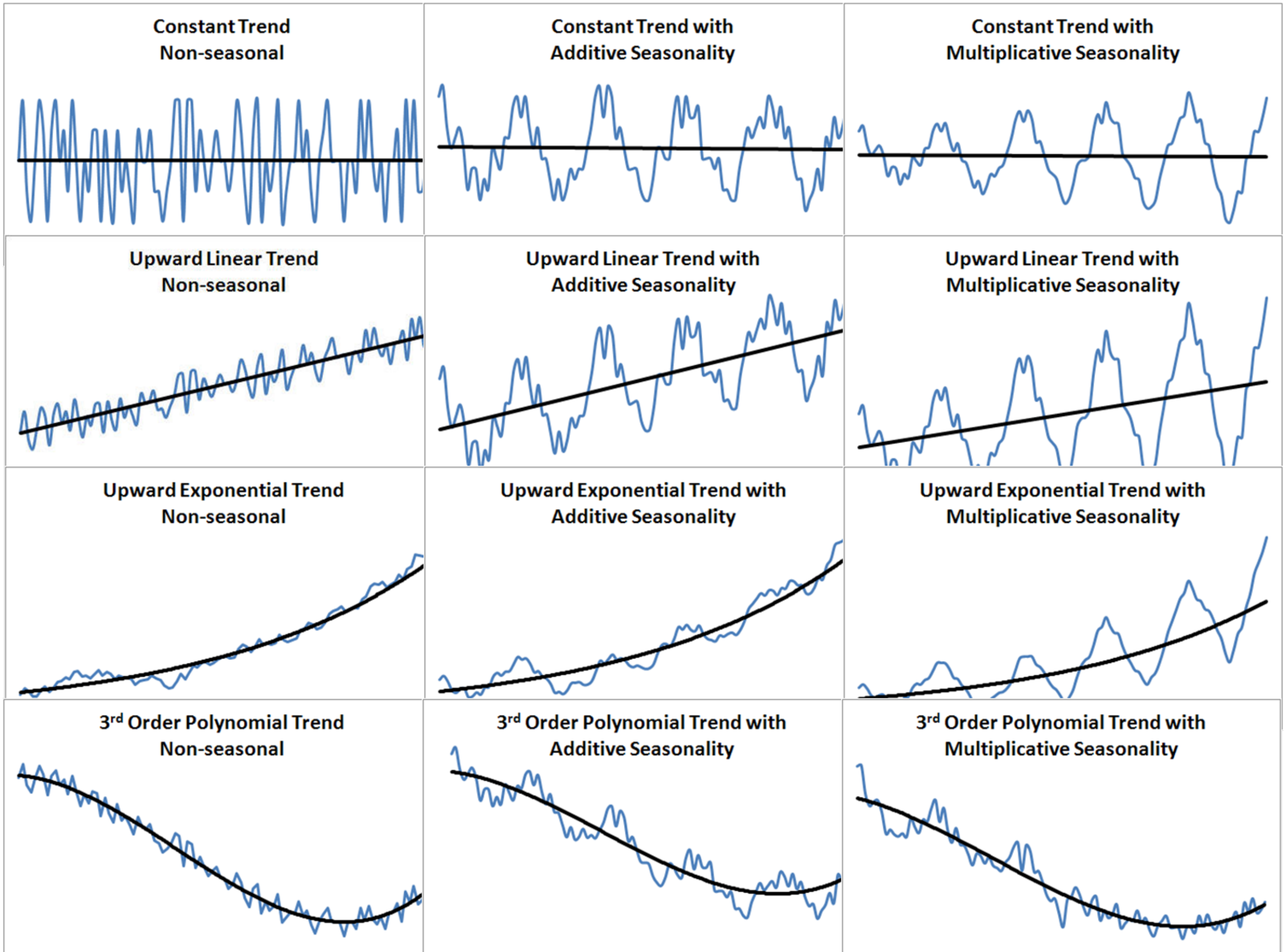
- “Noise”

Additive:

Y_t = Level + Trend + Seasonality + Noise

Multiplicative:

Y_t = Level x Trend x Seasonality x Noise



Constant Trend
Non-seasonal
Constant Trend with
Additive Seasonality
Constant Trend with
Multiplicative Seasonality
Upward Linear Trend
Non-seasonal
Upward Linear Trend with
Additive Seasonality
Upward Linear Trend with
Multiplicative Seasonality
Upward Exponential Trend
Non-seasonal
Upward Exponential Trend with
Additive Seasonality
Upward Exponential Trend with
Multiplicative Seasonality
3rd Order Polynomial Trend
Non-seasonal
3rd Order Polynomial Trend with
Additive Seasonality
3rd Order Polynomial Trend with
Multiplicative Seasonality

Overview

Assumptions

Fitting a trend (linear, exponential, other)

Fitting a seasonal component

Fitting other patterns

Autocorrelation

Example: Coca Cola Sales

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

Possible Goals:

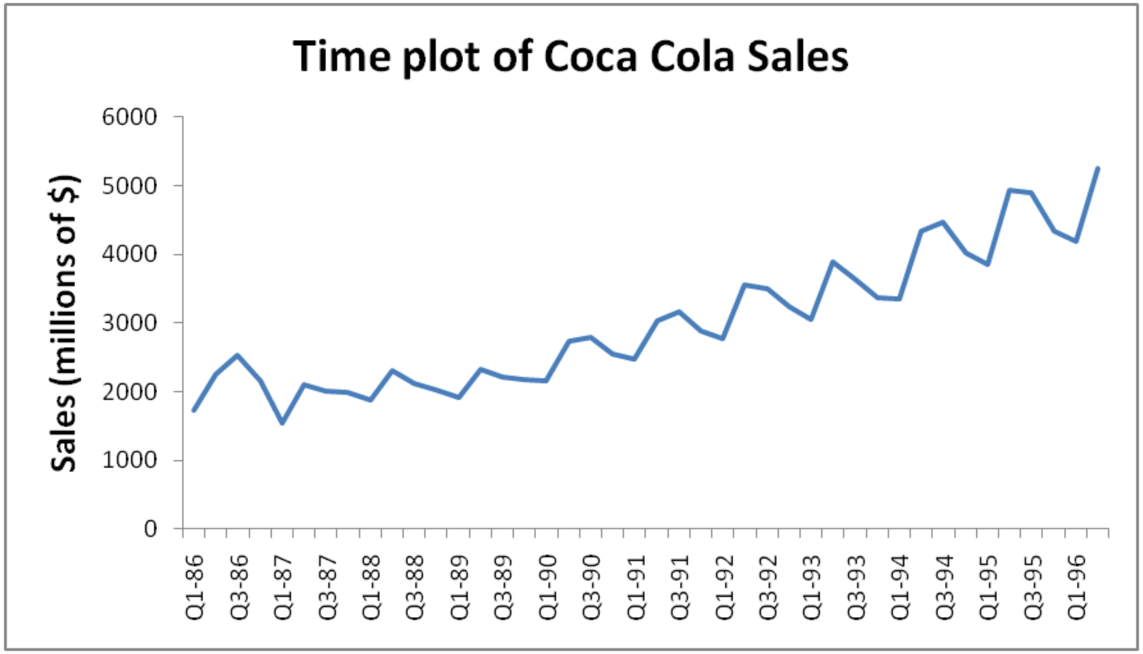
Time series forecasting:

create forecasts for the next 4 quarters

Time series analysis:

quantify the components (patterns and noise)

Step I: Time Plot



What do you see?

Data Partitioning

Goal: forecast future quarterly sales

Data partitioning:

- Training: first 38 quarters
- Validation: last 4 quarters

XLMiner: Time Series> Partition Data

- Allows specifying **last rows** as the validation set

Major Assumption: *Stationarity*

The future is “similar” to the past (in a probabilistic sense)

- What to do if assumption violated?
- Is more data always better?

Notation

T = Number of periods (#observations)
t = period/observation number: t =1,2,… T
We assume equally spaced intervals

$\boldsymbol{y_t}$ **= Observed value** at time t

$\boldsymbol{F_{t-k,\, t}}$ **= Forecast** for time t, based on data collected up to time t-k

$\boldsymbol{e_{t-k,\, t}} = \boldsymbol{y_t} - \boldsymbol{F_{t-k,\, t}}$ **= Forecast error** (**Residual**)

If clear from context, drop first subscript

Principles for modeling time series

Time series analysis

- Interpretation and parsimony
- Goodness of fit (residual analysis)

Time series forecasting

- Forecast accuracy (or some cost function)
- Parsimony

Toolkit of Regression-based Models

Fitting a trend (linear, exponential, etc.)

Fitting a seasonal component (additive, multiplicative)

Fitting trend + seasonal component

Capturing special events (e.g., holidays)

Capturing period-to-period correlation

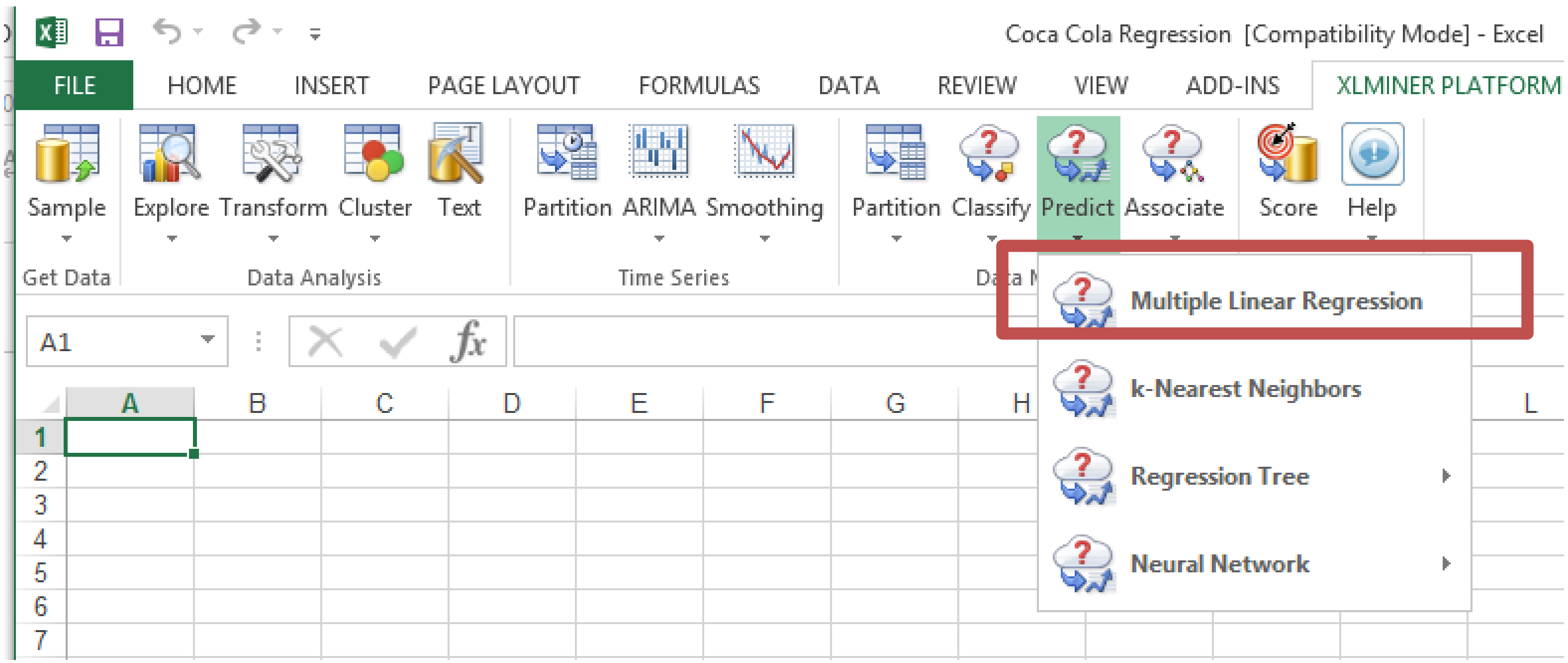
Model 1: Linear Trend Model

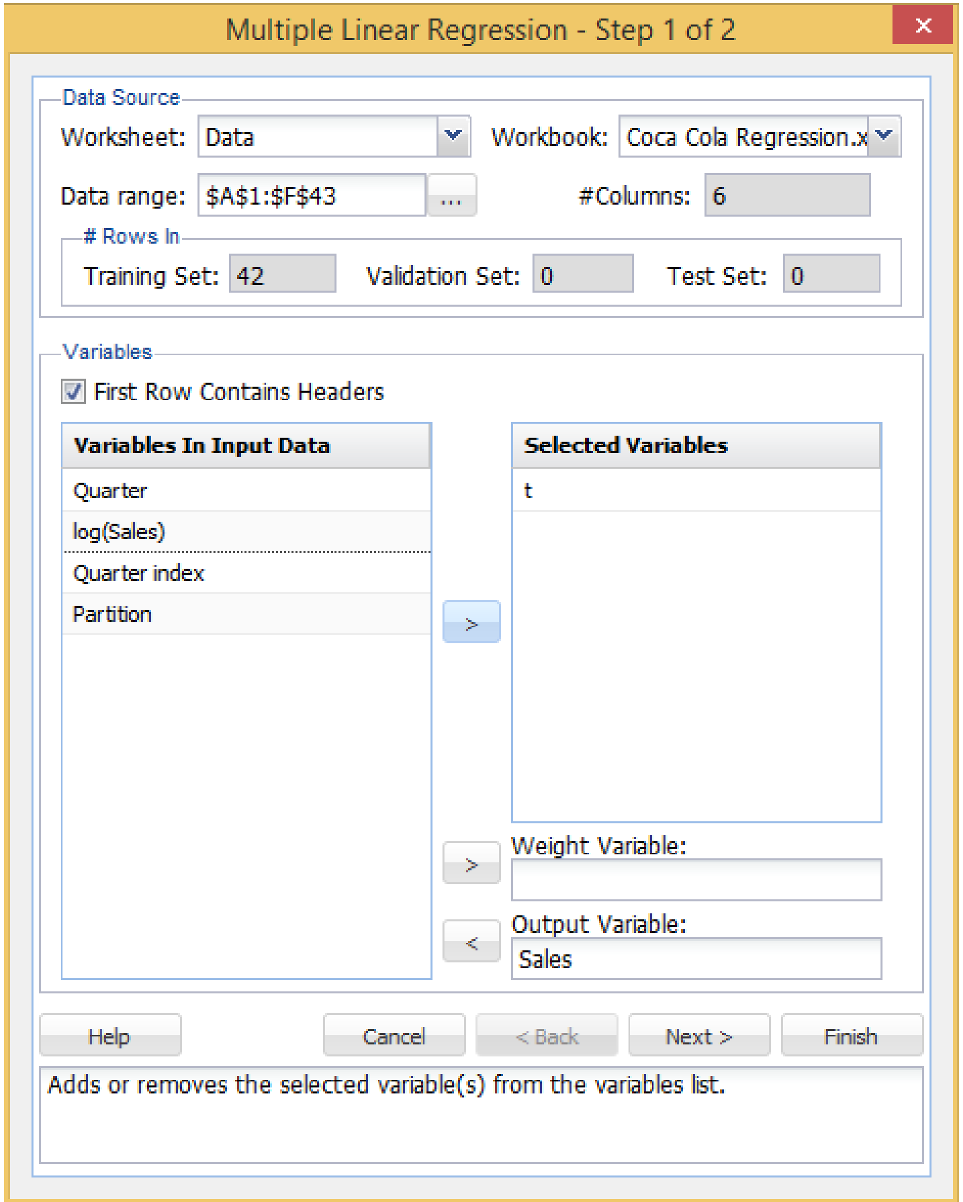
Trend = $ increase in quarterly growth

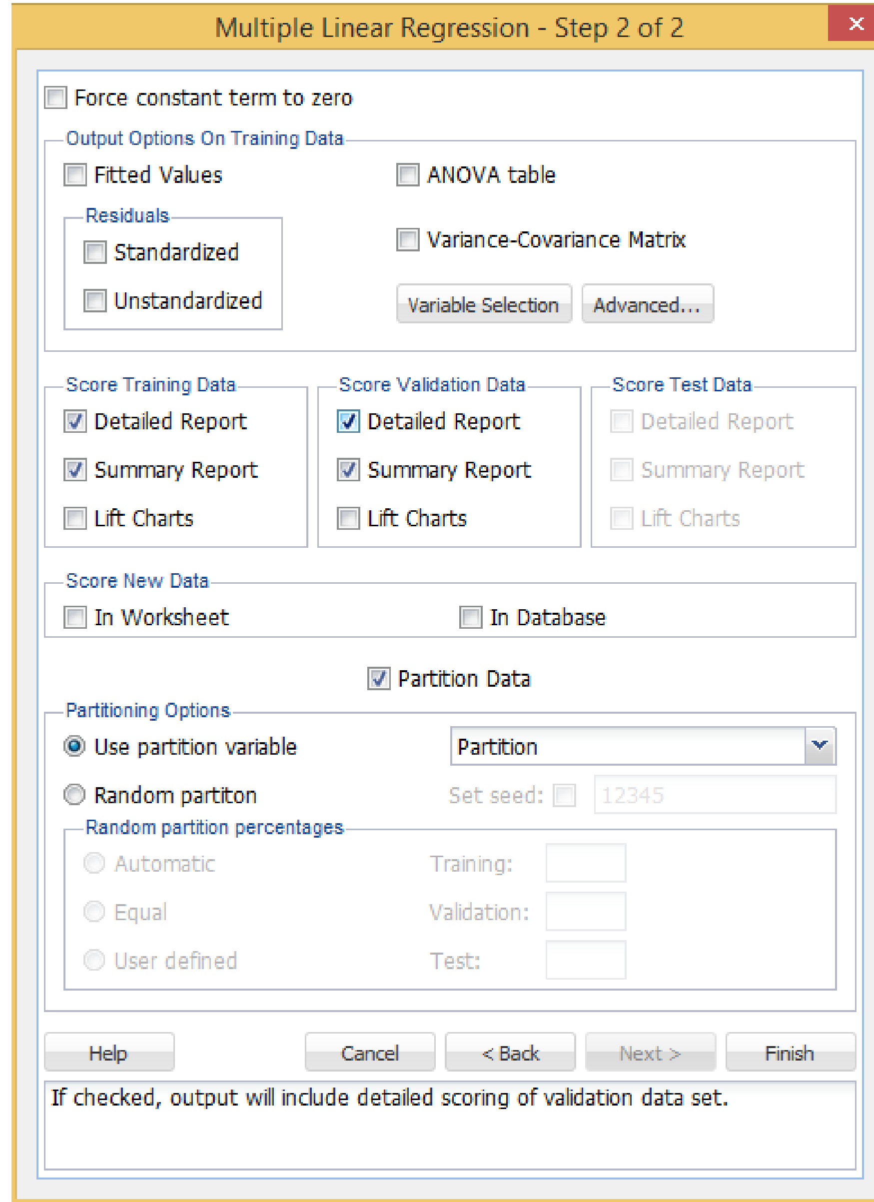
- Create running index variable *t=1,2,3…*
- Run regression model:

$$y_t = \beta_0 + \beta_1 t + \varepsilon$$

Linear Regression in XLMiner




Multiple Linear Regression - Step 1 of 2
Data Source
Worksheet: Data
Workbook: Coca Cola Regression.x
Data range: $A$1:$F$43
#Columns: 6
Rows In
Training Set: 42
Validation Set: 0
Test Set: 0
Variables
First Row Contains Headers
Variables In Input Data
Quarter
log(Sales)
Quarter index
Partition
Selected Variables
t
Weight Variable:
Output Variable:
Sales
Help
Cancel
< Back
Next >
Finish
Adds or removes the selected variable(s) from the variables list.


Multiple Linear Regression - Step 2 of 2
Force constant term to zero
Output Options On Training Data
Fitted Values
ANOVA table
Residuals
Standardized
Unstandardized
Variance-Covariance Matrix
Variable Selection
Advanced...
Score Training Data
Detailed Report
Summary Report
Lift Charts
Score Validation Data
Detailed Report
Summary Report
Lift Charts
Score Test Data
Detailed Report
Summary Report
Lift Charts
Score New Data
In Worksheet
In Database
Partition Data
Partitioning Options
Use partition variable
Partition
Random partiton
Set seed:
12345
Random partition percentages
Automatic
Training:
Equal
Validation:
User defined
Test:
Help
Cancel
< Back
Next >
Finish
If checked, output will include detailed scoring of validation data set.

Output for Fitting Model 1
(MLR_Output, MLR_TrainScore)

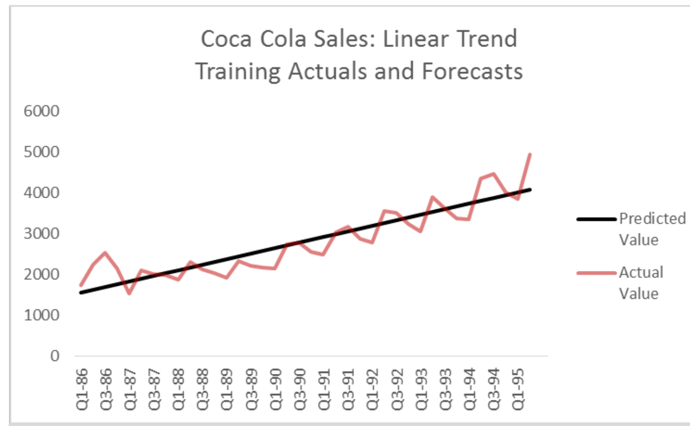
Regression Model

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	1490.736	122.247469	12.19441075	2.40996E-14	1242.806	1738.665	301784348.6
t	68.07001	5.464341527	12.4571293	1.29104E-14	56.98781	79.15221	21172897.6

Residual DF	36
R^2	0.811696
Adjusted R^2	0.806465
Std. Error Estim	369.379
RSS	4911870

Training Data Scoring - Summary Report

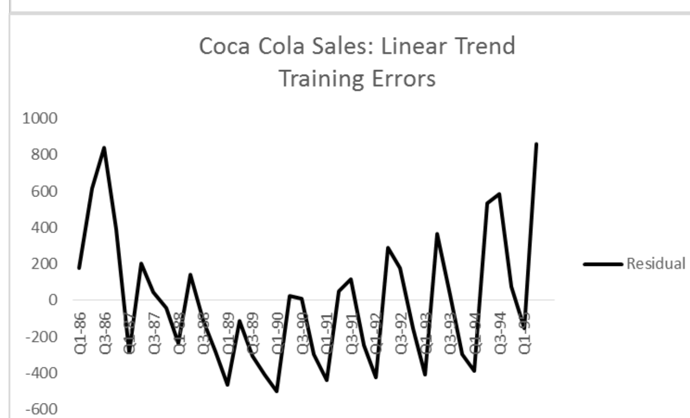
Total sum of squared errors	RMS Error	Average Error
4911870	359.5271	4.48764E-13



Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
1399741	591.5533	421.1787876

Is this random? →



Using the linear trend model

Forecast sales in Q3-95 (*validation set report*):

$$F_{39} = 1490.7358 + 68.07(39) = \$4{,}145.466\ \textbf{million}$$

Interpret b_1 = 68.07 :

Sales increase by average of $68.07 million/quarter

But is the interpretation valid, given the model fit?

Model 2: Exponential Trend Model

Trend = ***percentage*** quarterly growth

$$y_t = \alpha \, e^{\beta t} \, \varepsilon$$

$$\log(y_t) = \beta_0 + \beta_1 t + \varepsilon$$

Fit linear regression with $\log(Y_t)$ as output and *t* as predictor

Output for Fitting Model 2
(MLR_Output1, MLR_TrainScore1)

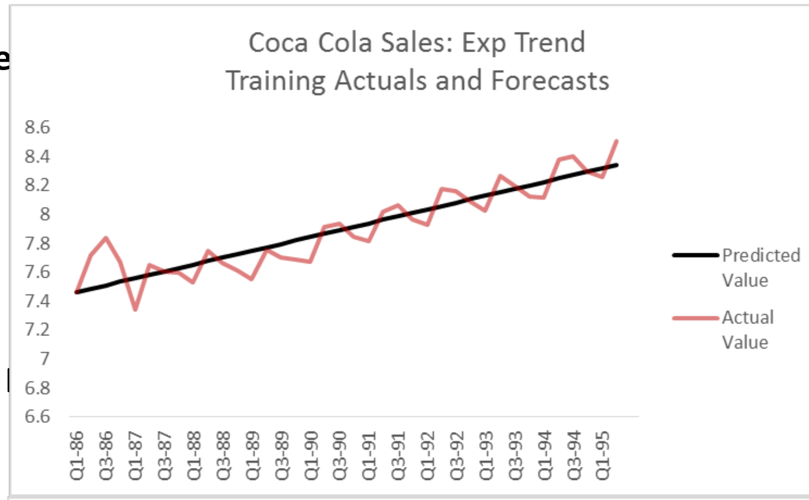
Regression Model

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	7.439351	0.04020242	185.0473408	3.19181E-55	7.357817	7.520885	2373.006
t	0.023745	0.00179701	13.21351162	2.24111E-15	0.0201	0.027389	2.576354

Residual DF	36
R^2	0.829057
Adjusted R^2	0.824309
Std. Error Estim	0.121474
RSS	0.531216

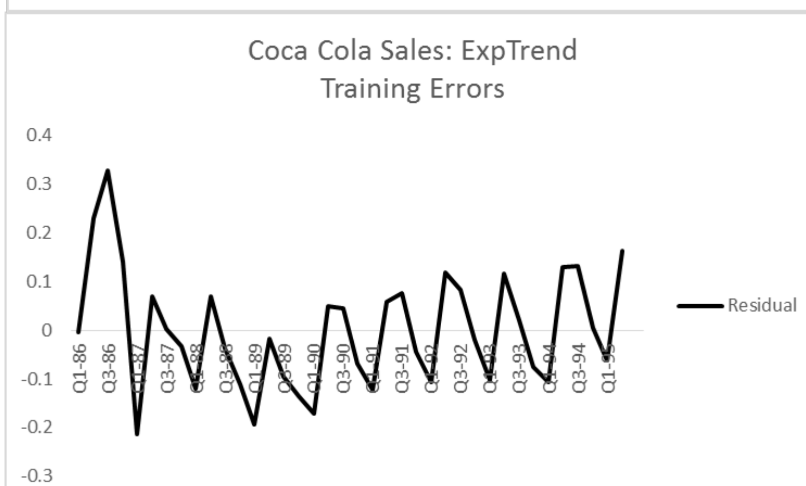
Training Data Scoring - Summary Re

Total sum of squared errors	RMS Error	Average Error
0.531216	0.118234	6.7782E-16



Validation Data Scoring - Summary I

Total sum of squared errors	RMS Error	Average Error
0.039266	0.099079	0.04347205



Using the exponential trend model

Forecast sales in Q3-95:

$$\log(F_{39}) = 7.4394 + 0.0237(39) = 8.3654$$
$$F_{39} = e^{8.3654} = \$4295.822 \text{ million}$$

Interpret $b_1 = 0.0237$:

Sales increase by average of 2.37% per quarter

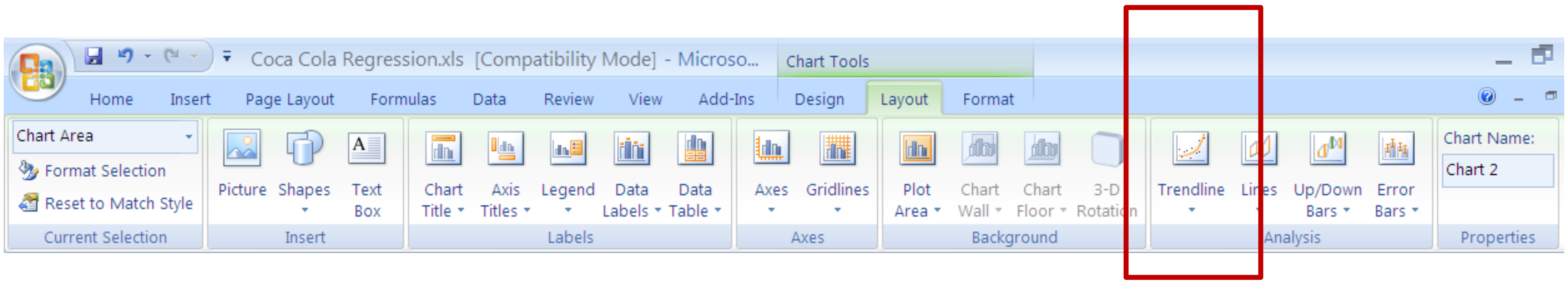
But is the interpretation valid, given the model fit?

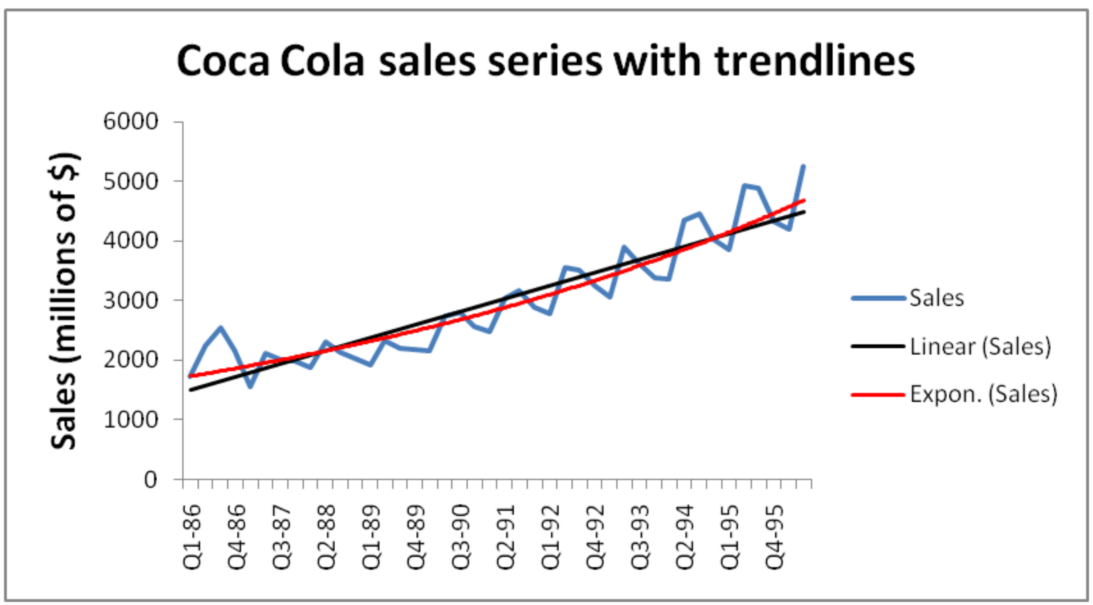
Comparing forecast accuracy

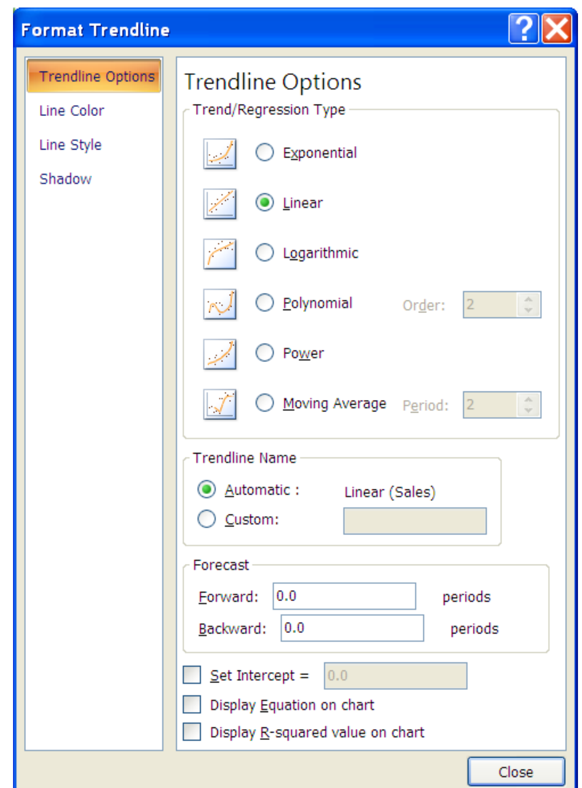
Validation Data Scoring - Summary Report

	Total sum of squared errors	RMS Error	Average Error
	0.039266	0.099079	0.04347205
Transformed to $	869548.7	466.248	215.596591

Trend lines in Excel







Model 3: Additive Seasonal Model

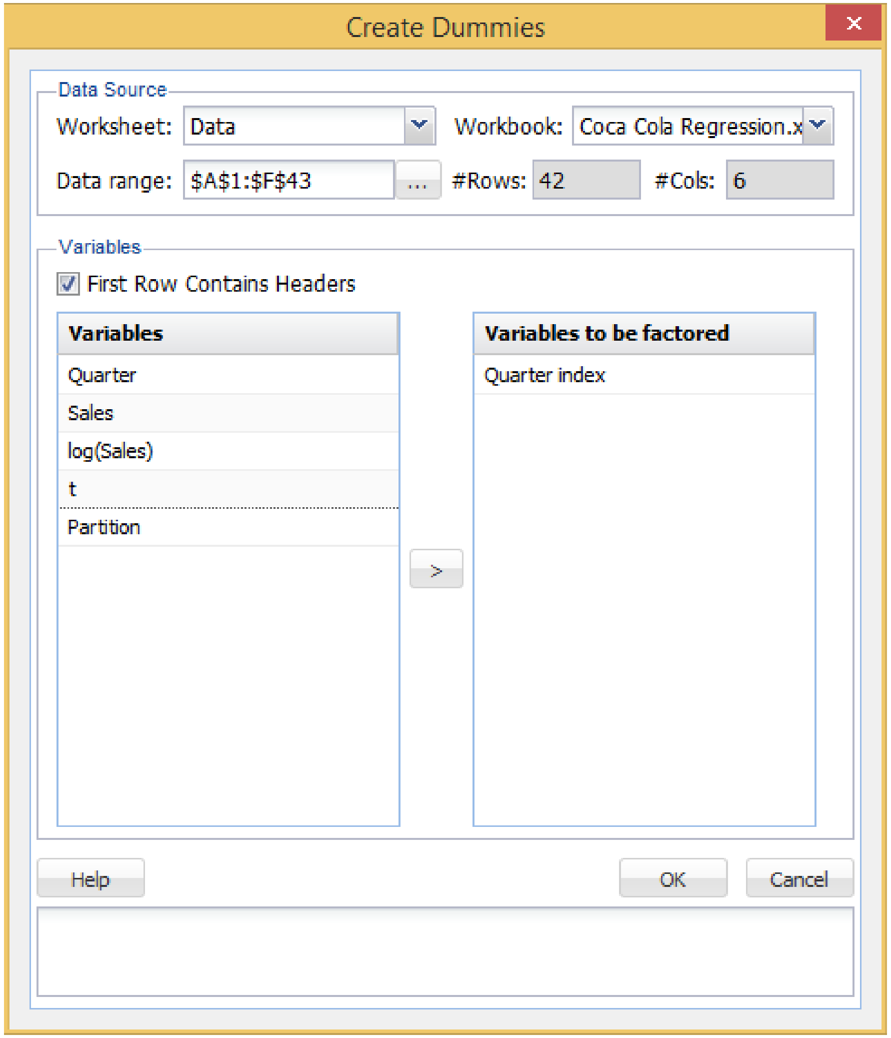
Additive = $ change from season to season

Coca Cola sales: how many seasons?

Augment linear trend model (Model 1) by including additional seasonal predictor(s)

How many dummy variables?

XLMiner: Transform > Transform Categorical Data > Create Dummies



Output for fitting Model 3

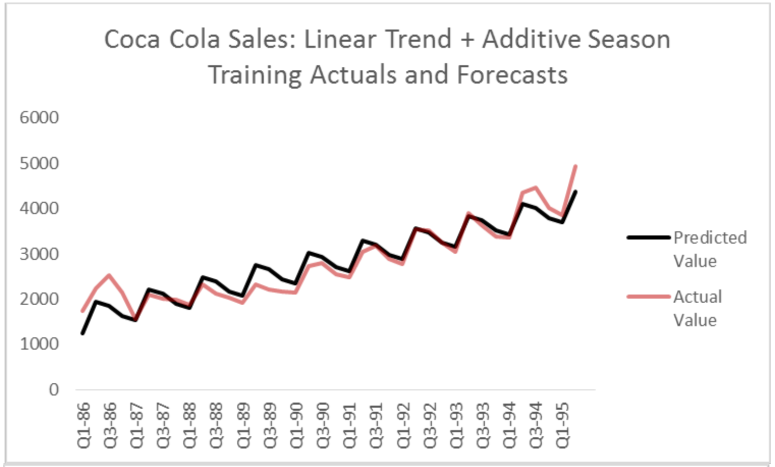
Regression Model

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	1186.416	120.2812966	9.863676838	2.28E-11	941.7017	1431.13	301784348.6
t	67.69155	4.208109892	16.08597375	3.52E-17	59.13008	76.25301	21172897.6
Quarter ind	609.7534	127.151785	4.795476157	3.37E-05	351.0611	868.4456	1204992.613
Quarter ind	465.879	130.5644789	3.568191174	0.001125	200.2436	731.5145	901043.8372
Quarter ind	172.6841	130.6322752	1.321910236	0.195289	-93.08922	438.4575	141105.1677

Residual DF	33
R^2	0.897844
Adjusted R^2	0.885461
Std. Error Estim.	284.1643
RSS	2664728

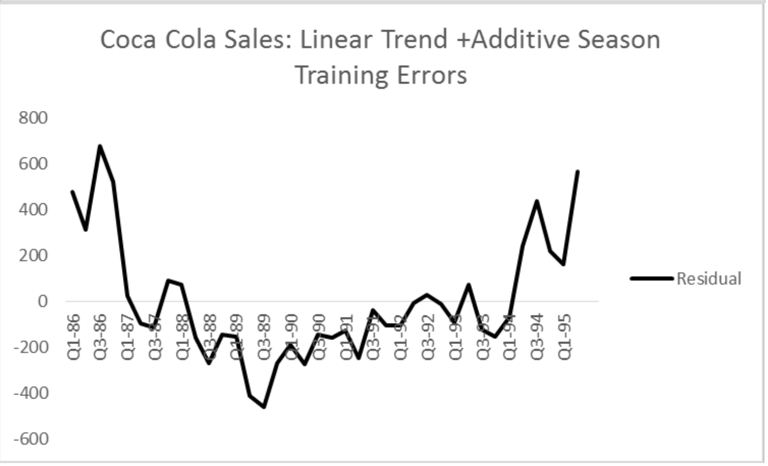
Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
2664728	264.8102	3.53028E-13



Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
864836.4	464.9829	428.7474485



Model 3: fitting the model

Dummy variables

$$D_2 = \begin{cases} 1, & \text{if } Q_2 \\ 0, & \textit{otherwise} \end{cases}$$

$$D_3 = \begin{cases} 1, & \text{if } Q_3 \\ 0, & \textit{otherwise} \end{cases}$$

$$D_4 = \begin{cases} 1, & \text{if } Q_4 \\ 0, & \textit{otherwise} \end{cases}$$

Model:

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

Forecasting with the additive seasonality model

For **Q3-95:**

t = _39_, D2 = _0_, D3 = _1_, D4 = _0_

Forecasted sales:

$$F_{39} = 1186.4 + 67.6915(39) + 609.8(0) + $$
$$+ 465.88(1) + 172.7(0) = \$4292.2651 \text{ million}$$

For **Q1-98** (into the future):

t = _49_, D2 = _0_, D3 = _0_, D4 = _0_

Forecasted sales:

$$F_{49} = 1186.4 + 67.6915(49) = \$4848.67 \text{ million}$$

Interpreting the Coefficients

Interpret $b_1 = 67.6915$:

Seasonally adjusted sales increase by an average of $67.6915 million per quarter

Interpret $b_2 = 609.7$:

After adjusting for trend, sales in Q2 are higher than sales in Q1 by an average of $609.7 million

But is the interpretation valid, given the model fit?

Model 4: Multiplicative Seasonal Model

Multiplicative = ***percentage*** change from season to season

$$y_t = c \cdot e^{bt} \cdot e^{b_2 D_2} \cdot e^{b_3 D_3} \cdot e^{b_4 D_4} \cdot \varepsilon$$

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

Exponential trend

Output for fitting Model 4

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

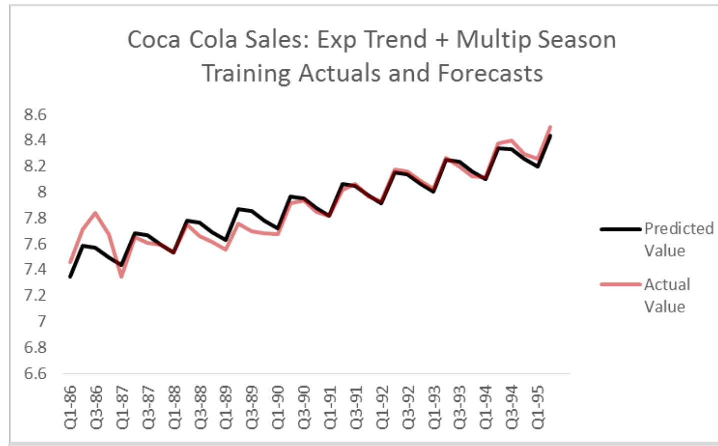
Residual DF	33
R^2	0.921558
Adjusted R^2	0.91205
Std. Error Estim	0.085946
RSS	0.243764

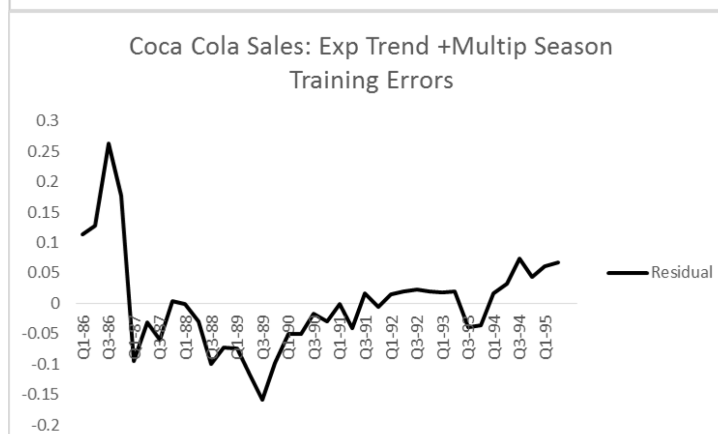
Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
0.243763542	0.080093	3.15537E-15

Validation Data Scoring - Summary Report

	Total sum of squared errors	RMS Error	Average Error
	0.009667825	0.049163	0.04585164
Transformed to $	203445.0028	225.5244	209.3619406





Interpreting the coefficients

Interpret $b_1 = 0.0236$:

Seasonally-adjusted sales increase by an average of 2.36% per quarter

Interpret $b_2 = 0.218$:

After adjusting for trend, Q2 sales are higher than Q1 by an average of 21.8%

But is the interpretation valid, given the model fit?

Forecasting with a multiplicative model (model 4)

For Q3-95: t =39, D_2 =0, D_3 =1 ,D_4 =0

$$\log(F_{39}) = 7.3222 + 0.0236(39) + 0.1812(1) = 8.4240$$

$$F_{39} = e^{8.4240} = \$4555.077 \text{ million}$$

What if we want different trends for different seasons?

Same (linear) trend for all seasons:

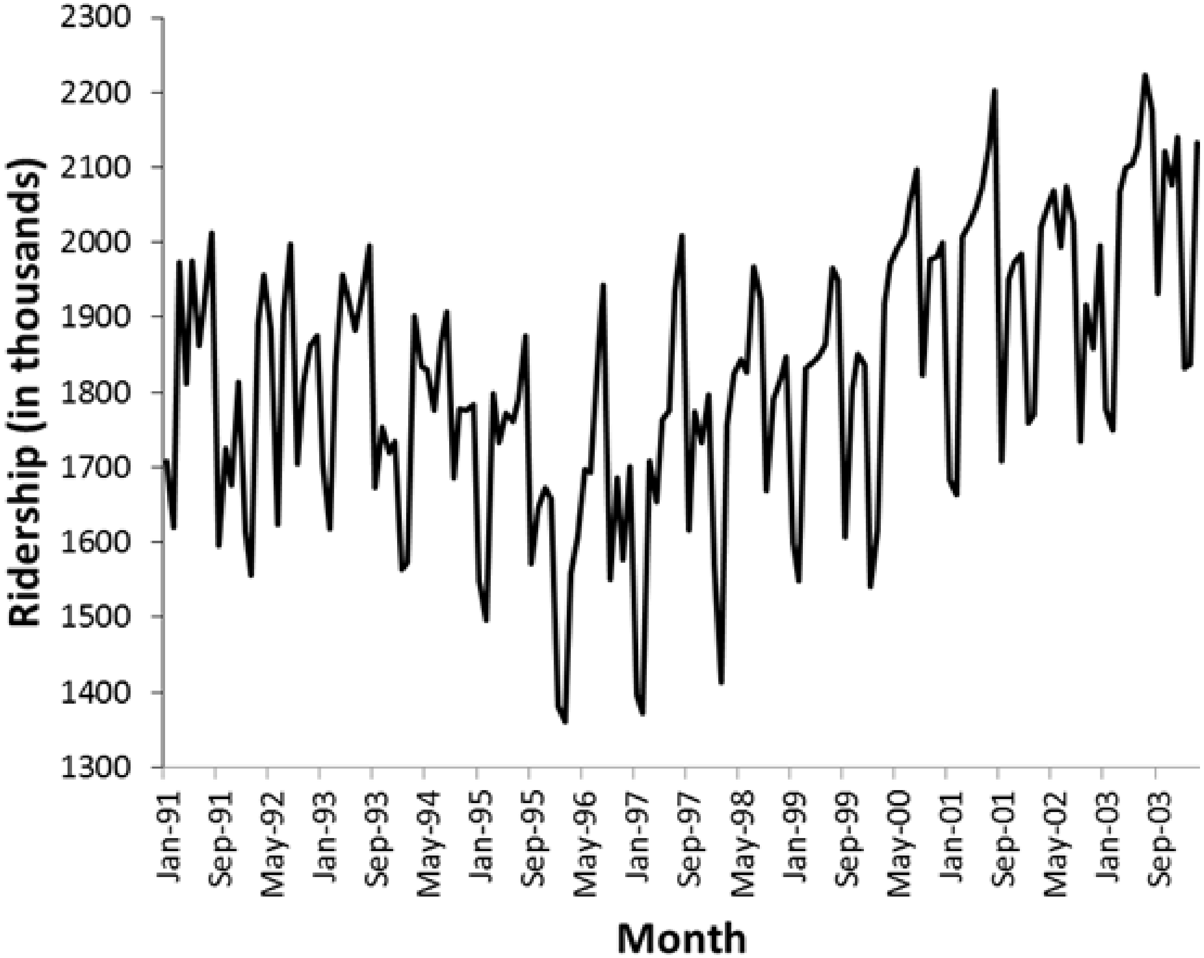
$$y_t = \beta_0 + \beta_1 t + \beta_2 D_2 + \beta_3 D_3 + \beta_4 D_4 + \varepsilon$$

Different slopes = use interaction terms

$$y_t = \beta_0 + \beta_1 t + \beta_2 D_2 + \beta_3 D_3 + \beta_4 D_4 +$$

$$+ \beta_5 t^*D_2 + \beta_6 t^*D_3 + \beta_7 t^*D_4 + \varepsilon$$

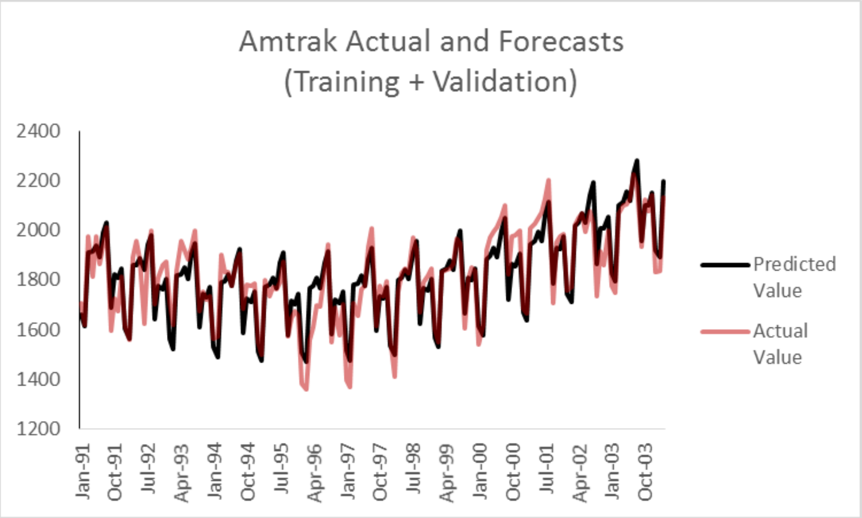
Another example: Amtrak Ridership



Regression Model

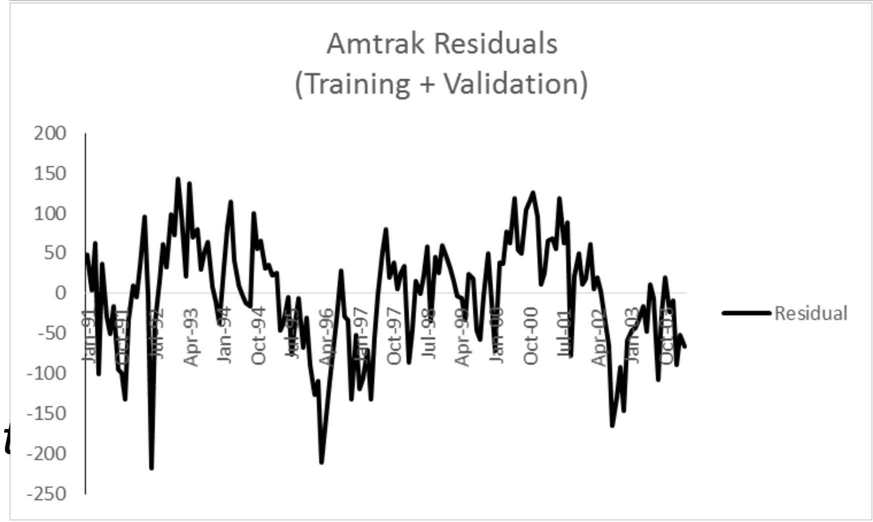
Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	1932.999	27.85863142	69.38599004	2.5E-106	1877.895	1988.102	477456482
t	-5.246521	0.586749089	-8.94167729	2.83E-15	-6.407088	-4.085954	384546.28
t^2	0.043757	0.003840708	11.39284872	2.09E-21	0.03616	0.051353	602582.64
Season_Aug	135.1726	30.52143549	4.428776488	1.96E-05	74.8024	195.5428	500917.22
Season_Dec	-29.65873	30.53801409	-0.97120688	0.333208	-90.06174	30.74428	31069.144
Season_Feb	-306.3078	29.94875664	-10.2277316	1.8E-18	-365.5453	-247.0704	720827.09
Season_Jan	-267.4445	29.94642287	-8.9307646	3.01E-15	-326.6773	-208.2116	653637.09
Season_Jul	91.31223	30.51900051	2.991979756	0.003305	30.9468	[illegible]	[illegible]
Season_Jun	-12.04475	30.5172469	-0.39468662	0.693706	-72.4066	[illegible]	[illegible]
Season_Mar	-7.044826	29.95207719	-0.23520327	0.814413	-66.2888	[illegible]	[illegible]
Season_May	30.31717	30.51618347	0.993478552	0.322281	-30.0426	[illegible]	[illegible]
Season_Nov	-72.26639	30.53282675	-2.36684257	0.019383	-132.659	[illegible]	[illegible]
Season_Oct	-60.98049	30.52834304	-1.99750405	0.047811	-121.364	[illegible]	[illegible]
Season_Sep	-199.1281	30.52454877	-6.5235394	1.32E-09	-259.504	[illegible]	[illegible]

Residual DF	133
R^2	0.825319
Adjusted R^2	0.808245
Std. Error Estim.	74.7482
RSS	743110



Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
743110.0191	71.09972	1.77877E-13



Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
30722.56431	50.59855	-34.1139112

Amtrak-Regression.xlsx, worksheet *MLR_Output*

XLMiner: generating forecasts for new periods ("scoring new data")

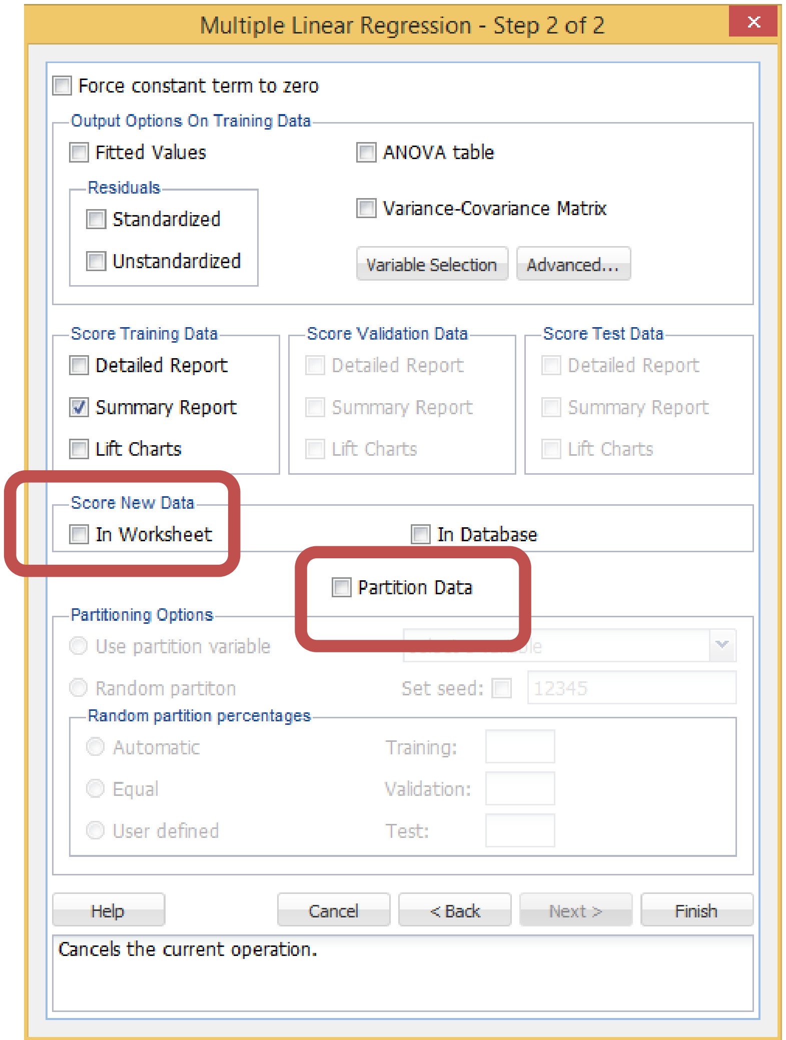
Prepare sheet with periods to forecast

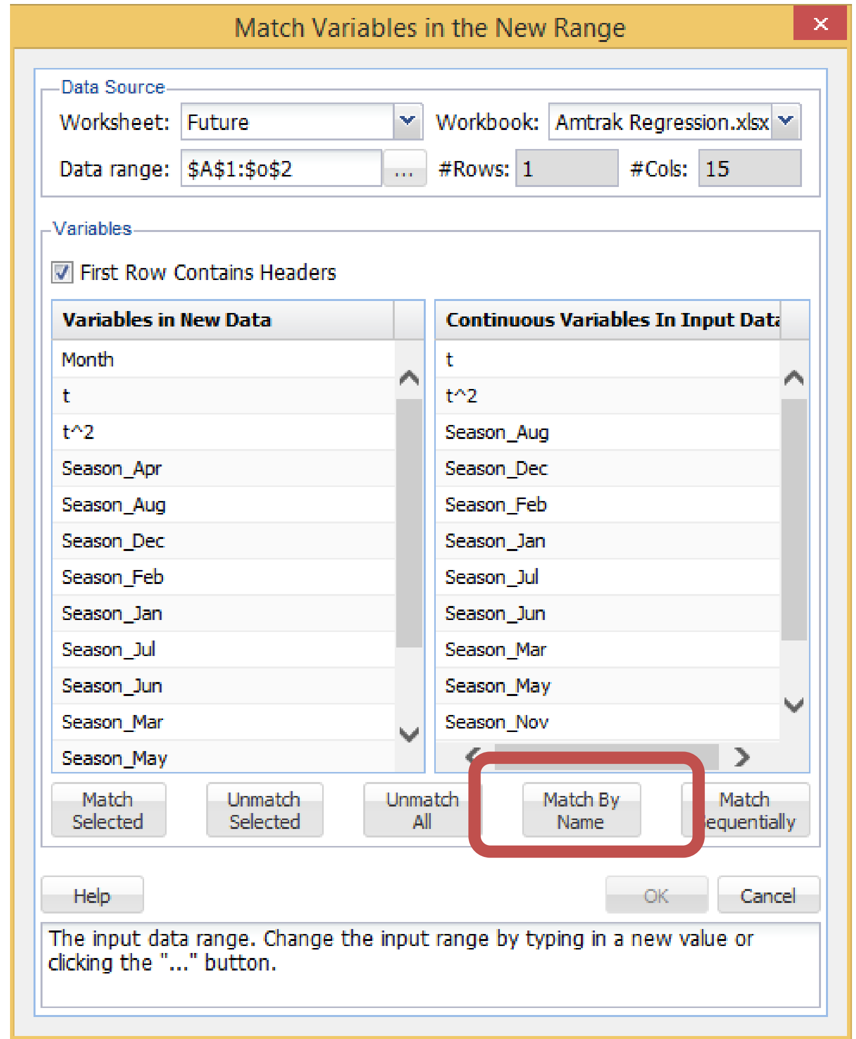
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Month	t	t^2	Season_Apr	Season_Aug	Season_Dec	Season_Feb	Season_Jan	Season_Jul	Season_Jun	Season_Mar	Season_May	Season_Nov	Season_Oct	Season_Sep
2	Apr-04	160	25600	1	0	0	0	0	0	0	0	0	0	0	0
3															
4															

Amtrak-Regression.xlsx, worksheet *Future*

XLMiner: generating forecasts ("scoring new data")

Run regression on un-partitioned data ; Choose "score new data"





XLMiner: result

New worksheet with forecasted period

XLMiner : Multiple Linear Regression - Prediction of New Data

Output Navigator			
New Data Detail Rpt.	Inputs	Predictors	Regress. Model

Workbook	Amtrak Regression.xlsx
Worksheet	Future
Range	$A$1:o2

Predicted Value	Confidence Intervals		Prediction Intervals	
	Lower	Upper	Lower	Upper
2193.895	2141.723	2246.068	2041.127	2346.664

t	t^2
160	25600

Amtrak-Regression.xlsx, worksheet *MLR_NewScore1*

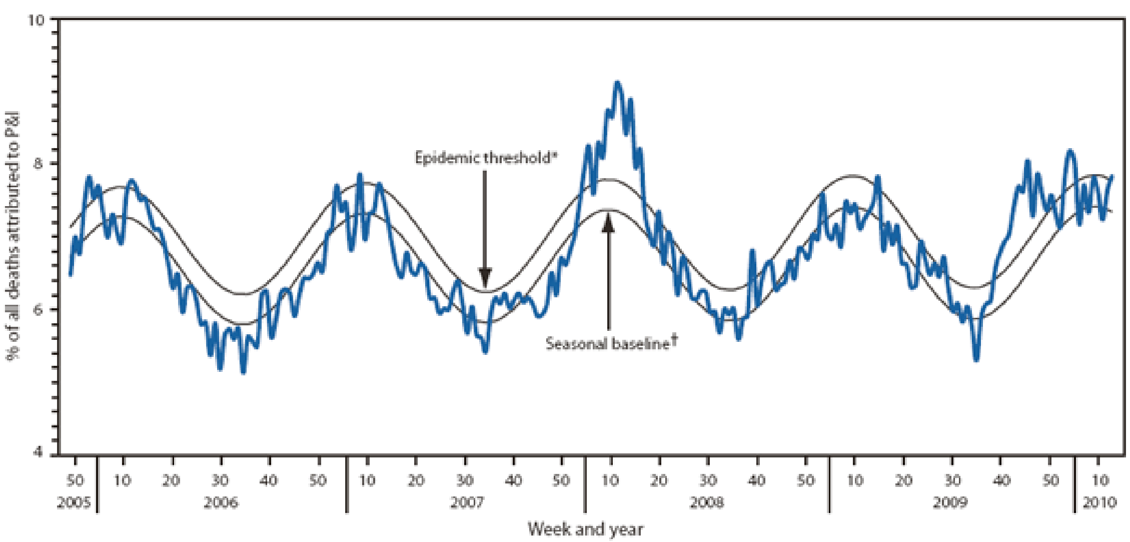
More trend/seasonality shapes

Polynomial trend (good for interpretation?):

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \ldots + \varepsilon$$

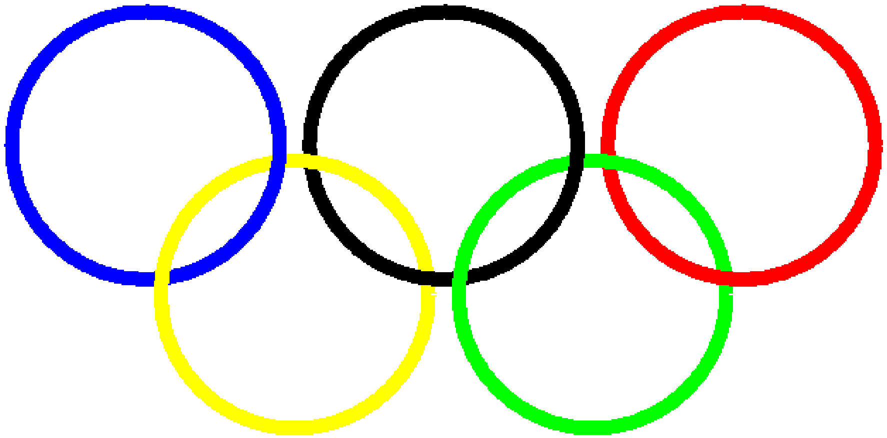
Smoothly-transitioning annual seasonality:

$$y_t = \beta_0 + \beta_1 \sin(2\pi t/52.18) + \beta_2 \cos(2\pi t/52.18) + \varepsilon$$



Irregular patterns

Explained/unexplained



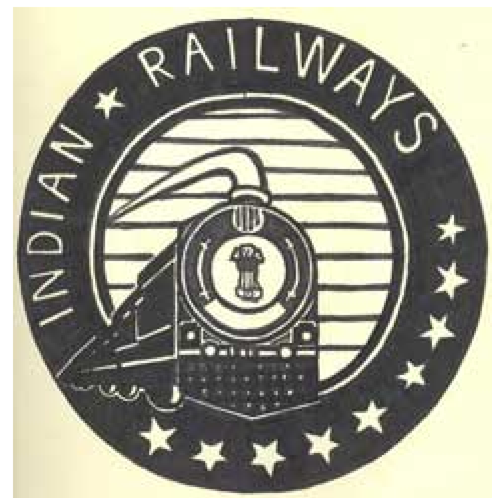
- Outliers
- Special events (holidays, sporting events)
- Interventions (promotion, policy change)

Solutions (depends on goal and data):

- Remove unusual periods from the model
- Model separately
- Keep in the model, using dummy variable

What if the pattern changes?

- Global / local pattern
- Do we know when it changes?



Separate models
External predictor

Including External Information

- Easy – include as predictors
- Make sure predictors are available at time of prediction

"In the NBC Internet example, we've found that TV ads have an impact for about six months, and a simple but good model would be
**Sales(t) = *g*{ *f*(sales(t-1, t-2, ... , t-6), a_1*SQRT[AdSpend(t-1)] + ...
+ a_6*SQRT[AdSpend(t-6)] }**

where the time unit is one month … and both *g* and *f* are functions that need to be identified via cross-validation and model fitting techniques"

-blog post by Vincent Granville, Dec 2011, www.analyticbridge.com
"Sales forecasts: how to improve accuracy while simplifying models?"

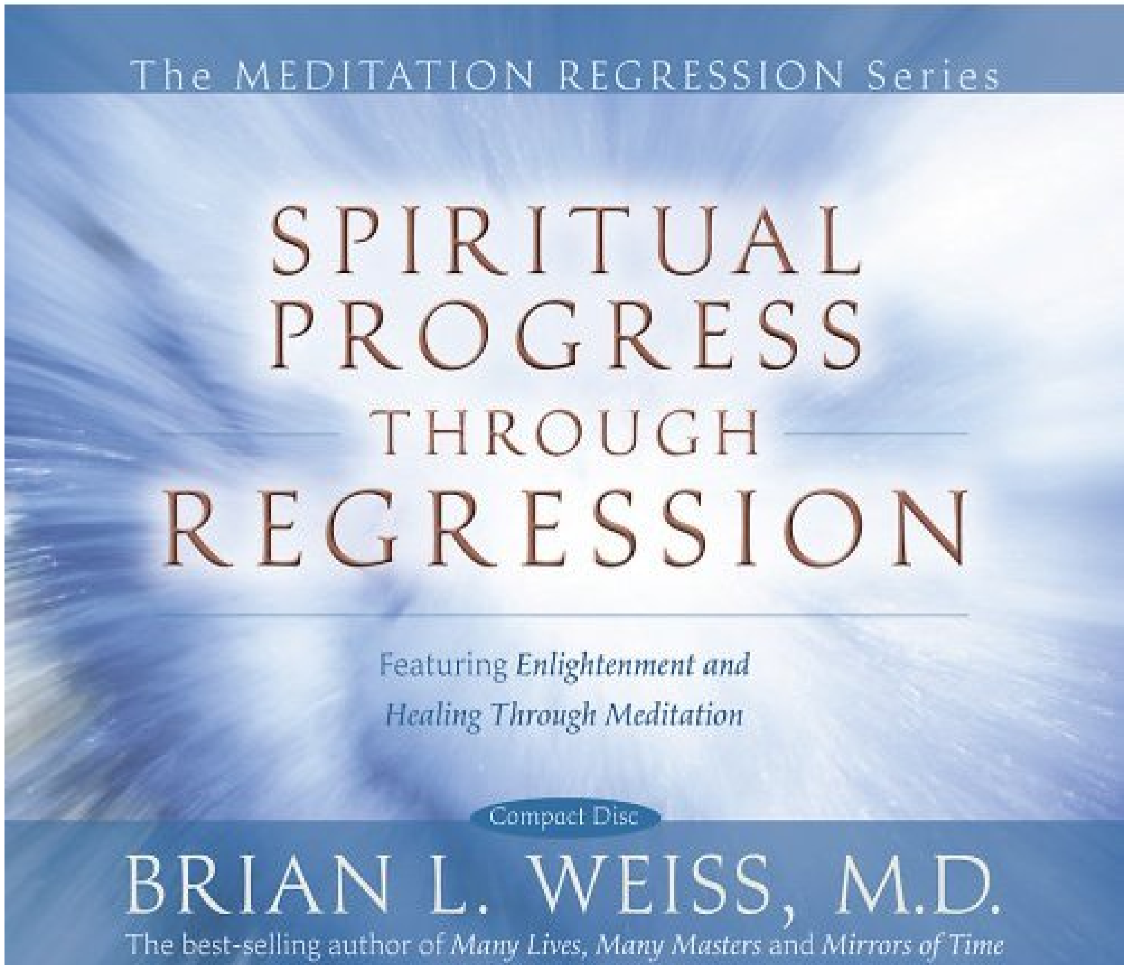
What about missing values?

- In the training period?
- In the validation period?

Summary: Linear regression for forecasting time series

- Useful for time series analysis and forecasting
- Global trend
  - linear trend (constant growth) - use time index as predictor.
  - exponential trend (% growth) - use log(Y) as response and time index as predictor
- Additive or multiplicative seasonality (or other shape)
  - Use dummy variables for seasons
  - Y for additive model or log(y) for multiplicative model
- Irregular patterns
- External predictors

Next class: Regression-based models for capturing period-to-period correlation



Next class

In class: Present proposed project (with chart of series) – 2 slides

Upload slides by 7am
Upload proposal (1 page) to LMS

What are the guidelines for the project proposal?

The proposal should be a single-page description of the following:

- Description of the business problem or goal
- Description of the forecasting problem (what are you forecasting? Forecast horizon? Etc.
- The particular data to be used (which series?)
- A relevant chart of the data
- Steps you have taken thus far (data cleaning, exploration, partitioning, modeling, etc.)

In class, use two slides only to present

- the business problem and forecasting problem, and
- display a relevant chart of the data.