Regression Models for Time Trends: A Second Example

INSR 260, Spring 2009

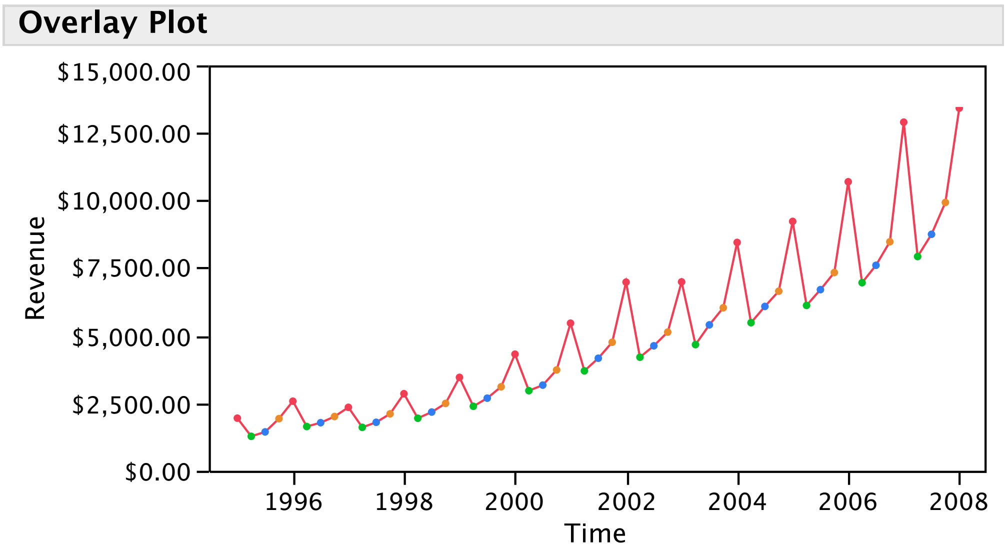
Bob Stine

Overview

- Resembles prior textbook occupancy example
  - Time series of revenue, costs and sales at Best Buy, in millions of dollars
  - Quarterly from 1995-2008
- Similar features
  - Log transformation
  - Seasonal patterns via dummy variables
  - Testing for autocorrelation: Durbin-Watson, lag residuals
  - Prediction with autocorrelation adjustments
- Novel features
  - Use of segmented model to capture change of regime
  - Decision to set aside some data to get consistent form

Forecasting Problem

- Predict revenue at Best Buy for next year
  - Q1, 1995 through Q1, 2008
  - 53 quarters
  - Forecast revenue for the rest of 2008
  - Estimate forecast accuracy
- Evident patterns
  - Growth
  - Seasonal
  - Variation



- Forecast of profit needs an estimate of cost of goods sold and amount of sales: then difference.

Initial Modeling

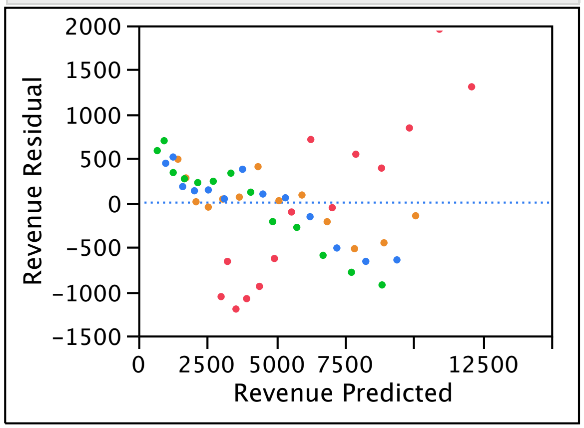
- Quadratic trend + quarterly seasonal pattern
- Overall fit is highly statistically significant

Summary of Fit

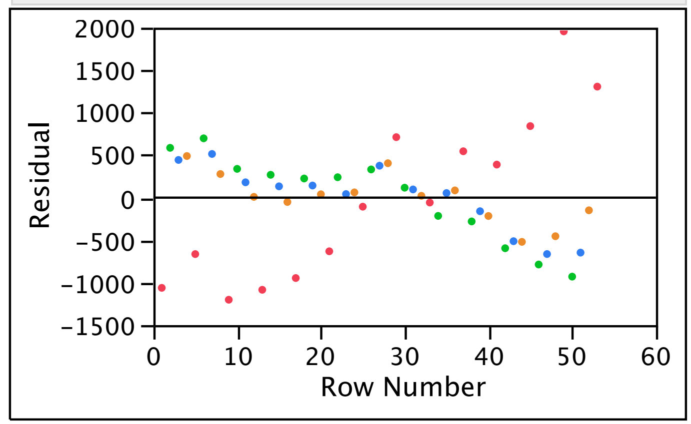
RSquare	0.959712
RSquare Adj	0.955426
Root Mean Square Error	632.221
Mean of Response	4952.975
Observations (or Sum Wgts)	53

- Nonetheless model shows problems in residuals

Residual by Predicted Plot



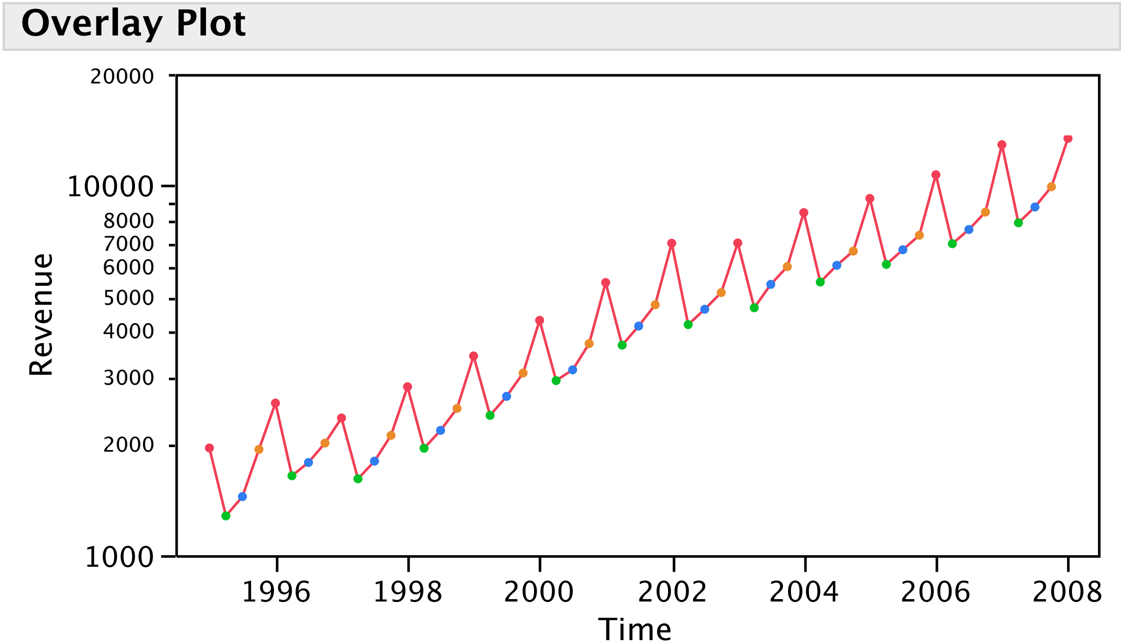
Residual by Row Plot



Trend in the first quarter of each year (red) appears different from those in other quarters... interaction.

Two Ways to Fix

- Two approaches
  - Add interactions that allow slopes to differ by quarter
    - Do you want to predict quadratic growth?
  - Log transformation
- Use log
  - Curvature remains, but variance seems stable with consistent patterns in the quarters



Model on Log Scale

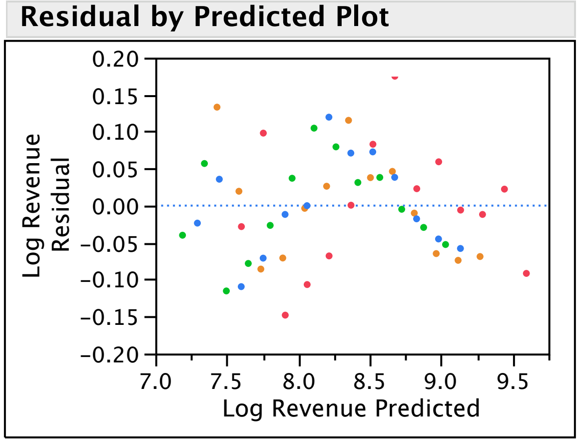
- Model of logs on time and quarter is highly statistically significant,

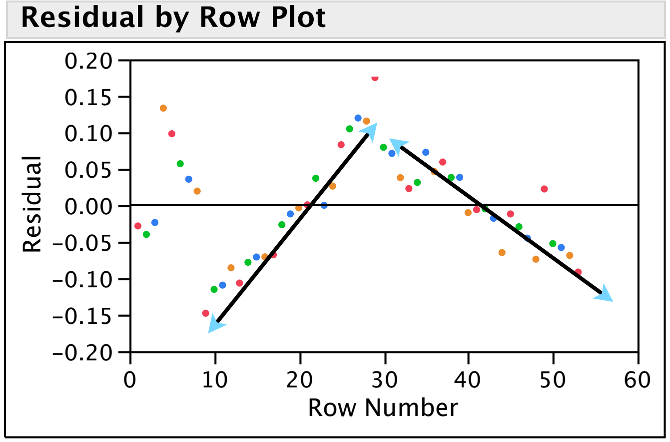
Summary of Fit	
RSquare	0.987077
RSquare Adj	0.986
Root Mean Square Error	0.073872
Mean of Response	8.324368
Observations (or Sum Wgts)	53

Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	-298.6066	5.316919	48.00	-56.16	<.0001*
Time	0.1533451	0.002656	48.00	57.73	<.0001*
Quarter[1]	0.2856838	0.02846	48.00	10.04	<.0001*
Quarter[2]	-0.164648	0.029005	48.00	-5.68	<.0001*
Quarter[3]	-0.09888	0.028982	48.00	-3.41	0.0013*

- But residuals show lack of fit and dependence

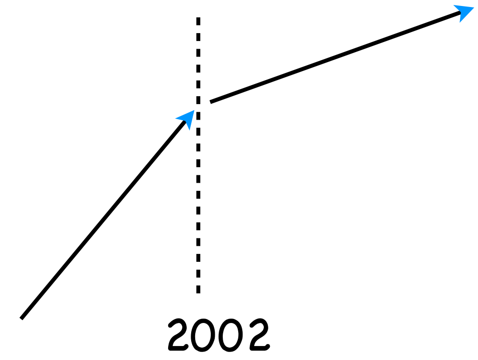




- Why does slope (% growth rate) seem to change?

Modified Trend

- Introduce "period" dummy variable
  - Exclude first two years of data (8 quarters)
  - Add Pre-Post Dot Com indicator
  - Allows slope to shift at start of 2002
  - Another shift is possible!



- Better model?
  - Summary statistics

Summary of Fit

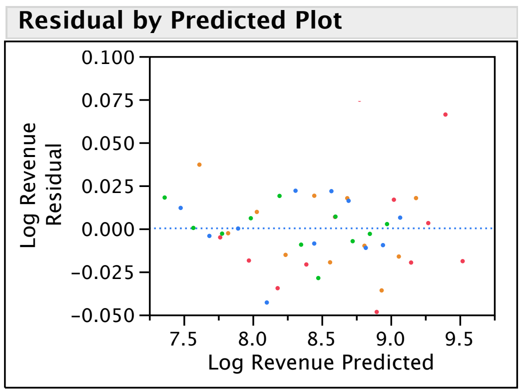
RSquare	0.998093
RSquare Adj	0.997792
Root Mean Square Error	0.025882
Mean of Response	8.473075
Observations (or Sum Wgts)	45

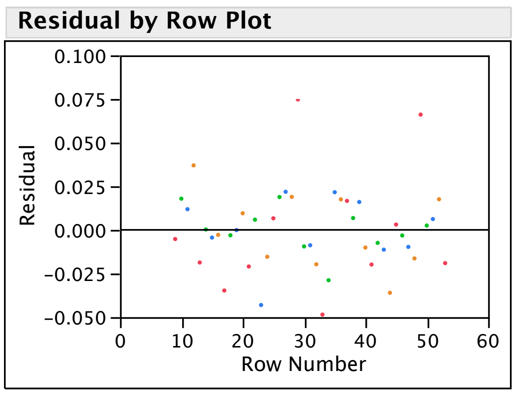
Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	-408.1624	8.094352	38.00	-50.43	<.0001*
Time	0.2081232	0.004048	38.00	51.41	<.0001*
Quarter[1]	0.306712	0.010896	38.00	28.15	<.0001*
Quarter[2]	-0.147721	0.011102	38.00	-13.31	<.0001*
Quarter[3]	-0.083811	0.011053	38.00	-7.58	<.0001*
Pre/Post Dot Com[post]	167.27411	9.912849	38.00	16.87	<.0001*
Time*Pre/Post Dot Com[post]	-0.083569	0.004953	38.00	-16.87	<.0001*

Huge shift in rate of growth

  - Residual plots



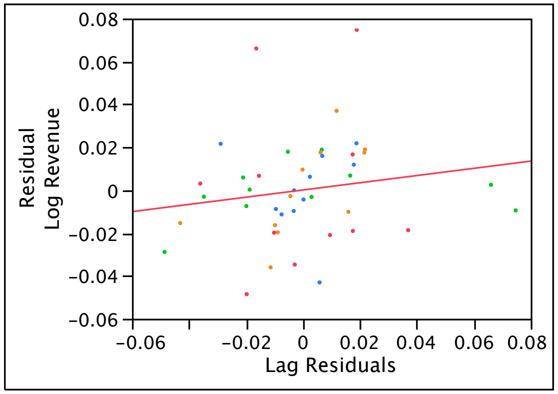


Autocorrelation?

- Dependence absent from sequence plot
  - Confirmed by Durbin-Watson, residual scatterplot

Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
1.6527607	45	0.1660	0.0718



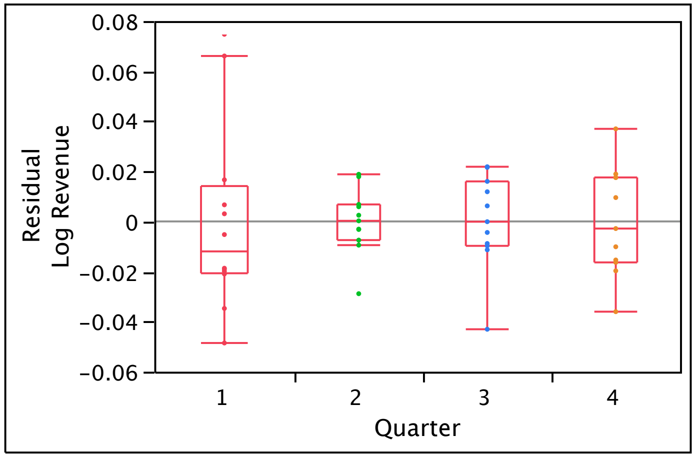
- No need to add lagged residual as explanatory variable; all captured by trend + seasonal

Indicator Function Parameterization

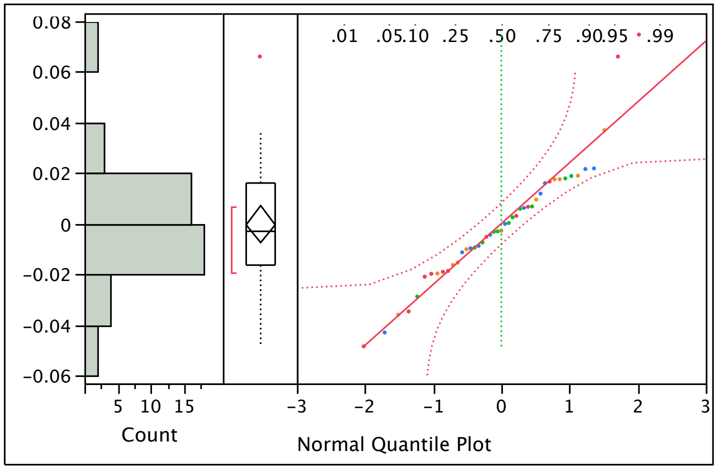
Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	-407.8512	8.821915	36.00	-46.23	<.0001*
Time	0.2079678	0.004412	36.00	47.14	<.0001*
Quarter[1]	0.3072369	0.011212	36.00	27.40	<.0001*
Quarter[2]	-0.148054	0.011246	36.00	-13.16	<.0001*
Quarter[3]	-0.083831	0.011189	36.00	-7.49	<.0001*
Pre/Post Dot Com[post]	166.99646	10.55057	36.00	15.83	<.0001*
Time*Pre/Post Dot Com[post]	-0.08343	0.005272	36.00	-15.82	<.0001*
Lag Residuals	0.1691184	0.165917	36.00	1.02	0.3149

More Diagnostics

- Residual plots show little remaining structure
  - Similar variances in quarters?



  - Normality seems reasonable (albeit outliers in Q1)



Forecasting

- Forecast log revenue for rest of 2008
  - $\hat{y}_{n+j} = (-408.162 + 167.274 + Q_j) +$ seasonal
    $(0.20812-0.08357)$ time time trend
  - Overall intercept plus adjustment for pre/post

- Examples for Q2, Q3, Q4 of 2008
  - $\hat{y}_{53+1} = (-408.162 + 167.274 - 0.148)$ $Q_2 = -0.148$
    $+ 0.12455\ (2008.25)$
    $\approx \underline{9.092}$
  - $\hat{y}_{53+2} = (-408.162 + 167.274 - 0.084)$ $Q_3 = -0.100$
    $+ 0.1245\ (2008.50)$
    $\approx \underline{9.187}$
  - $\hat{y}_{53+3} = (-408.162 + 167.274)$ $Q_4 = 0$
    $+ 0.1245\ (2008.75)$
    $\approx \underline{9.302}$

Forecast Accuracy

- Since model does not have autocorrelation and data meet assumptions of MRM, we can use the JMP prediction intervals

- One period out
  - $\hat{y}_{53+1} \pm t_{.025}$ SE(indiv pred) = 9.0415 to 9.1587

- Two periods out
  - $\hat{y}_{53+2} \pm t_{.025}$ SE(indiv pred) = 9.1363 9.2540

- Three periods out
  - $\hat{y}_{53+3} \pm t_{.025}$ SE(indiv pred) = 9.2510 9.3692

Prediction Intervals

- Obtain predictions of revenue, not the log of revenue

- Conversion
  - Form interval as we have done on transformed scale
  - Exponentiate

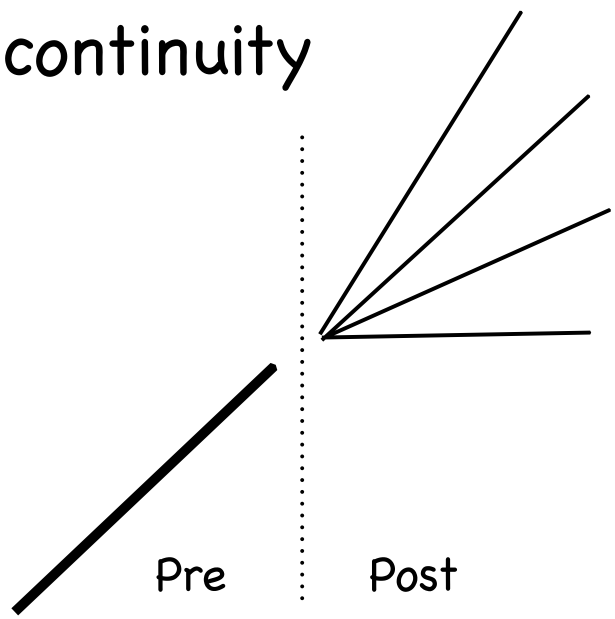
$$9.0415 \text{ to } 9.1587 \quad \Rightarrow \quad e^{9.0415} \text{ to } e^{9.1587}$$

$8446 to $9497 (million)

- As in prior example, the prediction interval is much wider than you may have expected from the R^2 and RMSE of the model on the log scale.
  - Small differences on log scale are magnified on $ scale

Alternative Segments

- Prior approach adds two variables to segment
  - Dummy variable for period allows new intercept
  - Interaction allows slope to change
- Models fit in the two periods are "disconnected"
  - Not constrained to be continuous or intersect where the second period begins
- Alternative approach forces continuity
  - Add one parameter for change in the slope
  - No dummy variable needed.
  - Intercept defined by the location of the prior fit.



Building the Variables

- Model comparison
  - Break in structure (kink) at time T
  - Before ($t \le T$) : $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$
  - After ($t > T$) : $Y_t = \alpha_0 + (\beta_1 + \delta)X_t + \varepsilon_t$
  - Choose α_0 so that means match at time T

    $$\beta_0 + \beta_1 X_T = \alpha_0 + (\beta_1 + \delta)X_T \quad \Rightarrow \quad \alpha_0 = \beta_0 - \delta X_T$$

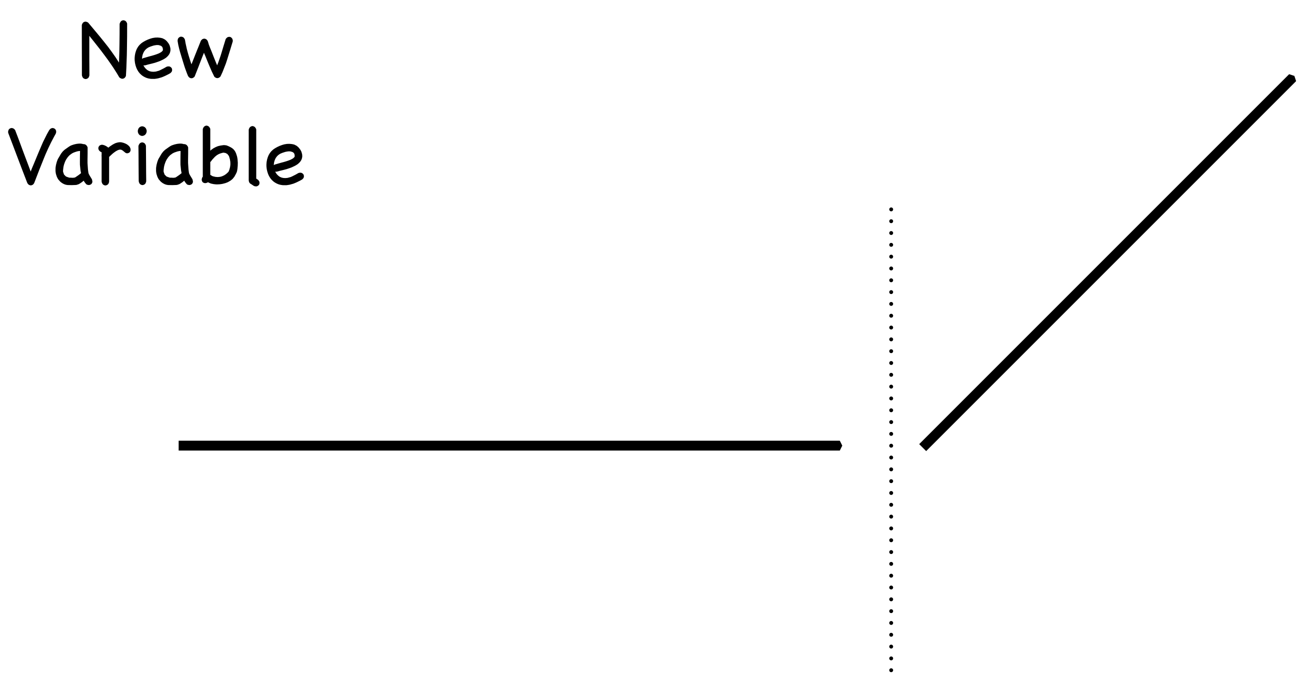
  - Hence, only need to estimate one parameter, δ

- To fit with regression, add the variable Z_t
  - $Z_t = 0$ for $t \le T$, $Z_t = X_t - X_T$ for $t > T$
  - Before T: no effect on the fit since 0
  - After T: $\beta_0 + \beta_1 X_t + \delta Z_t = \beta_0 + \beta_1 X_t + \delta (X_t - X_T)$

    $$= (\beta_0 - \delta X_T) + (\beta_1 + \delta) X_t$$

Changing the Slope

- Added variable is very simple
  - Prior to the change point, it’s 0
  - After the change point, its (x - time of change)
- Picture shows “dog-leg” shape of new variable with kink at the change point



Example

- Fit with distinct segments

Summary of Fit	
RSquare	0.998093
RSquare Adj	0.997792
Root Mean Square Error	0.025882
Mean of Response	8.473075
Observations (or Sum Wgts)	45

Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	−408.1624	8.094352	38.00	−50.43	<.0001*
Time	0.2081232	0.004048	38.00	51.41	<.0001*
Quarter[1]	0.306712	0.010896	38.00	28.15	<.0001*
Quarter[2]	−0.147721	0.011102	38.00	−13.31	<.0001*
Quarter[3]	−0.083811	0.011053	38.00	−7.58	<.0001*
Pre/Post Dot Com[post]	167.27411	9.912849	38.00	16.87	<.0001*
Time*Pre/Post Dot Com[post]	−0.083569	0.004953	38.00	−16.87	<.0001*

- Fit with continuous joint
  - Almost as large R^2, with one less estimated parameter
  - Similar shift in slope in two models.

Summary of Fit	
RSquare	0.997901
RSquare Adj	0.997632
Root Mean Square Error	0.026804
Mean of Response	8.473075
Observations (or Sum Wgts)	45

Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	−397.4332	6.166522	39.00	−64.45	<.0001*
Time	0.2027556	0.003083	39.00	65.76	<.0001*
Time Post	−0.081303	0.004988	39.00	−16.30	<.0001*
Quarter[1]	0.3042508	0.011209	39.00	27.14	<.0001*
Quarter[2]	−0.149787	0.011446	39.00	−13.09	<.0001*
Quarter[3]	−0.084844	0.011433	39.00	−7.42	<.0001*

Summary

- A basic trend (linear, perhaps quadratic) plus dummy variables is a good starting model for many time series that show increasing levels.
- Log transformations stabilize the variation, are easily interpreted, and avoid more complicated trends and interactions.
- Dummy variables can model a "trend break".
  - Models do not anticipate the time of another trend break in the future.
  - Special "broken line" variable models shift in slope with one parameter, forcing continuity.
- R^2 is misleading when you see the prediction intervals when fitting on a log scale.