Regression Models for Time Trends

INSR 260, Spring 2009
Bob Stine
Overview

- Review categorical variables
- Polynomial trends
- Seasonal patterns via indicators
- Testing for omitted patterns: Durbin-Watson
- Prediction

Example (from Bowerman, Ch 6)
- Planning staffing levels for a seasonal business: Hotel occupancy
- Other examples in Chapter 6 Time Series Regression
Categorical Variables

Two special types of explanatory variables

- Indicators
  - Shift the regression line up or down by altering the intercept of the fitted model for cases in a subset

- Interactions
  - Alter the slope for a particular group, capturing different levels of association between y and x within subsets

Particularly relevant test: Partial F-test

- Used in general to test whether a subset of slopes in a regression model are zero
- Test whether the slopes (interaction) or the intercepts (categorical slopes) differ among the groups
Forecasting Problem

- Predict occupancy rates for hotel
  - 14 years of monthly data, n = 168
  - Forecast occupancy during the next year
  - Provide a measure of the forecast accuracy

Evident patterns
- Growth
- Seasonal
- Variation

Color-coding can also help verify the seasonality
Modeling Approach

 Decomposition (also in Ch 7)
 Data = Trend + Seasonal + Irregular

 Trend
 Simple functions of time that are easily forecasted, such as linear or quadratic growth

 Seasonal
 Repeating patterns, such as those related to weather or holidays

 Irregular
 May be dependent and predictable
Initial Modeling

Linear trend + Monthly seasonal pattern

Multiple regression with time trend (month = 1, 2, 3...) and monthly dummy variables (11 indicators, dec omitted)

Overall fit is highly statistically significant

Specific coefficients by-and-large differ
Residual Diagnostics

- Substantial pattern was missed
- Big $R^2$ does not guarantee a “good” model

Two residual plots are essential when have time series data:
- familiar plot of $e$ on $\hat{y}$
- sequence plot of the residuals
Two Ways to Fix

Two approaches
- Add interactions that allow slopes to differ by season
- Transform the response to stabilize the variance

Log transformation
- Natural log (base e)

Can also show original on log scale (better for presenting)
Revised Model

Very impressive fit overall (on log scale)

Do NOT compare $R^2$ statistic to prior model since the response variable is not the same as in the prior model

Interpretation of slope for time

- Rate of growth: about 0.3% per month

Interpretation of dummy variables

- Shift intercept relative to December
Residual Diagnostics

Pattern remaining?

How should the model be improved – if at all?
- What types of variables are missing from the model?
- What is a simple revision of the model?

Note: text does not revise the model
Revised Model

- Model with an additional quadratic component
- Suggests rate of growth is slowing
- Statistically significant improvement?

Further structure?

Summary of Fit

| Term                      | Estimate  | Std Error | DFDen | t Ratio | Prob>|t| |
|---------------------------|-----------|-----------|-------|---------|------|
| Intercept                 | 6.2724878 | 0.007035  | 154.00| 891.55  | <.0001* |
| Time                      | 0.0032592 | 0.000129  | 154.00| 25.35   | <.0001* |
| Time*Time                 | -3.159e-6 | 7.369e-7  | 154.00| -4.29   | <.0001* |
| Month[Jan]                | -0.041606 | 0.007601  | 154.00| -5.47   | <.0001* |
| Month-Feb                  | -0.112111 | 0.007599  | 154.00| -14.75  | <.0001* |
| Month[Mar]                | -0.084516 | 0.007598  | 154.00| -11.12  | <.0001* |
| Month[Apr]                | 0.0397572 | 0.007598  | 154.00| 5.23    | <.0001* |
| Month[May]                | 0.0203067 | 0.007597  | 154.00| 2.67    | 0.0083* |
| Month[Jun]                | 0.1468146 | 0.007596  | 154.00| 19.33   | <.0001* |
| Month[Jul]                | 0.2889278 | 0.007595  | 154.00| 38.04   | <.0001* |
| Month[Aug]                | 0.3111061 | 0.007594  | 154.00| 40.97   | <.0001* |
| Month[Sep]                | 0.0559114 | 0.007594  | 154.00| 7.36    | <.0001* |
| Month[Oct]                | 0.039487  | 0.007593  | 154.00| 5.20    | <.0001* |
| Month[Nov]                | -0.112247 | 0.007593  | 154.00| -14.78  | <.0001* |

Indicator Function Parameterization

Residual by Predicted Plot

Residual by Row Plot
Testing Residual Dependence

- Durbin-Watson test
  - Test whether adjacent residuals appear dependent
  - Test related to autocorrelation between residuals
    - Autocorrelation is correlation between “rows” in the data table, whereas the usual correlation is between “columns”

- Lag plot of residuals

![Lag plot of residuals](image)

- Regression summary

<table>
<thead>
<tr>
<th>Linear Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual Log Occupied = 0.000194 + 0.3257149*Lag Residuals</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Summary of Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSquare</td>
</tr>
<tr>
<td>RSquare Adj</td>
</tr>
<tr>
<td>Root Mean Square Error</td>
</tr>
<tr>
<td>Mean of Response</td>
</tr>
<tr>
<td>Observations (or Sum Wgts)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Lag Residuals</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Durbin-Watson</th>
</tr>
</thead>
<tbody>
<tr>
<td>Durbin-Watson</td>
</tr>
<tr>
<td>Number of Obs.</td>
</tr>
<tr>
<td>AutoCorrelation</td>
</tr>
<tr>
<td>Prob&lt;DW</td>
</tr>
</tbody>
</table>
Adjusting for Autocorrelation

Two reasons to adjust
- Improves short-term forecast accuracy
- Corrects errors in claimed statistical significance

Comparison of forecast errors
- Do not model dependence
  \[ y_{n+1} = \beta_0 + \beta_1 x_{n+1,1} + \ldots + \beta_k x_{n+1,k} + \varepsilon_{n+1} \]
  \[ \hat{y}_{n+1} = b_0 + b_1 x_{n+1,1} + \ldots + b_k x_{n+1,k} + 0 \]
- Modeling dependence
  \[ \varepsilon_t = \varphi \varepsilon_t + a_t, \quad \text{Var}(a_t) = (1-\varphi^2) \text{Var}(\varepsilon_t) \leq \text{Var}(\varepsilon_t) \]
  \[ \hat{y}_{n+1} = b_0 + b_1 x_{n+1,1} + \ldots + b_k x_{n+1,k} + \hat{\varphi} e_n \]

Dependence distorts standard error estimates
- Failure to recognize the presence of dependence produces spurious claims of accuracy.
Simple Adjustment

Add the lagged residuals from the current model as an explanatory variable

Text describes more elaborate methods (p 311)

Residual plots show little remaining structure

Other variables are still missing. Are these important?

We’ll ignore them for the moment and build forecasts.

Durbin-Watson is always OK after this correction
Forecasting

Forecast log occupancy several periods out

\[ \hat{y}_{n+j} = (6.2736 + b_j) + 0.00322 \, (n+j) - 0.00000293 \, (n+j)^2 + 0.328^j \, (e_n) \]

- seasonal
- time trend
- autocorr

Autocorrelation effect drops off geometrically, having little influence past a few terms

Point estimates for January, February

\[ \hat{y}_{168+1} = (6.2736 - 0.0390) + 0.00322 \, (169) - 0.00000293 \, (169)^2 + 0.328^2 \, (0.0456) \approx 6.2346 + 0.4605 + 0.0150 = 6.7101 \]

\[ \hat{y}_{168+2} = (6.2736 - 0.1121) + 0.00322 \, (170) - 0.00000293 \, (170)^2 + 0.328^2 \, (0.0456) \approx 6.1615 + 0.4627 + 0.0049 = 6.6291 \]
Forecast Accuracy

- More accurate in the near term because of the dependence between adjacent errors
- Benefit of autocorrelation decreases as extrapolate out
- Must trick JMP into making the correct intervals
- Following are approximate intervals; JMP shown next

One period out: use RMSE of fitted model
- \( \hat{y}_{168+1} \pm t_{0.025,152} \) RMSE = 6.7101 ± 1.98 (0.0191)
  \( \approx 6.6723 \) to 6.7479

Two periods out: inflate RMSE by \( \sqrt{1+\hat{\phi}^2} \)
- \( \hat{y}_{168+2} \pm t_{0.025,152} \) RMSE\((1+\hat{\phi}^2)^{1/2} = 6.6291\pm1.98(0.0191)(1+.328^2)^{1/2} \)
  \( \approx 6.589 \) to 6.669

m periods out: inflate RMSE by
- \( \sqrt{1 + \hat{\phi}^2 + \hat{\phi}^4 + \ldots + \hat{\phi}^{2(m-1)} \approx \sqrt{1/(1-\hat{\phi}^2)}} \)
JMP Calculations

Prediction interval
\[ \hat{y} \pm t_{0.025} \text{RMSE (Extrapolation) (Autocorrelation)} \]
“distance value”

Four components determine width of interval
1. t-percentile... \( \approx 2 \) for 95% coverage
2. RMSE... SD of unexplained factors
3. Extrapolation... increases as forecast farther from data
4. Autocorrelation... extrapolate residuals beyond 1 period

JMP adjusts for the first 3, but not the fourth
- Software “does not know” that we’ve plugged in predicted values of residuals rather than using known residuals
- Increase in length of interval is very small unless autocorrelation \( \phi \) is close to 1.
JMP Calculations, cntd

The autocorrelation adjustment is the square root of the expression on the bottom of slide 16

\[ \sqrt{1 + \hat{\phi}^2 + \hat{\phi}^4 + \ldots + \hat{\phi}^{2(m-1)}} \]

This portion of the data table for hotel occupancy shows the data and columns.

<table>
<thead>
<tr>
<th>Lag Residuals</th>
<th>Pred Formula Log</th>
<th>StdErr Indiv Log</th>
<th>Lower 95% Indiv Log</th>
<th>Upper 95% Indiv Log</th>
<th>RMSE Adjustment</th>
<th>Corrected 95% PI,</th>
<th>Corrected 95% PI,</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.0146867</td>
<td>6.73155</td>
<td>0.02007</td>
<td>6.69189</td>
<td>6.77121</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>-0.0106001</td>
<td>6.86167</td>
<td>0.02007</td>
<td>6.82202</td>
<td>6.90133</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>-0.0238339</td>
<td>7.00171</td>
<td>0.02015</td>
<td>6.96190</td>
<td>7.04152</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>0.00338923</td>
<td>7.03509</td>
<td>0.02010</td>
<td>6.99538</td>
<td>7.07479</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>-0.0084815</td>
<td>6.77825</td>
<td>0.02011</td>
<td>6.73851</td>
<td>6.81799</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>-0.0139638</td>
<td>6.76228</td>
<td>0.02014</td>
<td>6.72248</td>
<td>6.80207</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>-0.0090122</td>
<td>6.61441</td>
<td>0.02015</td>
<td>6.57461</td>
<td>6.65421</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>0.01952859</td>
<td>6.73826</td>
<td>0.02023</td>
<td>6.69830</td>
<td>6.77823</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>0.04564133</td>
<td>6.71004</td>
<td>0.02069</td>
<td>6.66917</td>
<td>6.75091</td>
<td>1</td>
<td>6.66949</td>
<td>6.75059</td>
</tr>
<tr>
<td>0.015</td>
<td>6.62912</td>
<td>0.02029</td>
<td>6.58903</td>
<td>6.66920</td>
<td>1.0525</td>
<td>6.58726</td>
<td>6.67097</td>
</tr>
</tbody>
</table>

Estimated future residual

sqrt(1+phi^2)

Slightly wider
Prediction Intervals

We need predictions of the occupancy, not the log of the occupancy
- Predictions from model are on a log scale

Conversion
- Form interval as we have done on transformed scale
- Then “undo” the transformation (here, exponentiate)

\[ 6.6695 \text{ to } 6.7479 \Rightarrow e^{6.6723} \text{ to } e^{6.7506} \]
\[ 790 \text{ to } 855 \text{ rooms} \]

Interval is much wider than you may have expected from the $R^2$ and RMSE of model
- Differences get far larger when exponentiate
Summary

- Polynomial trends are useful when lack other, substantive explanatory variables
  - Be cautious extrapolating a trend

- Dummy variables model regular seasonal effects, but the magnitude of the variation often increases with the level

- Log transformation stabilizes the variation and captures geometric growth

- Durbin-Watson statistic tests for presence of autocorrelation in underlying model errors