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

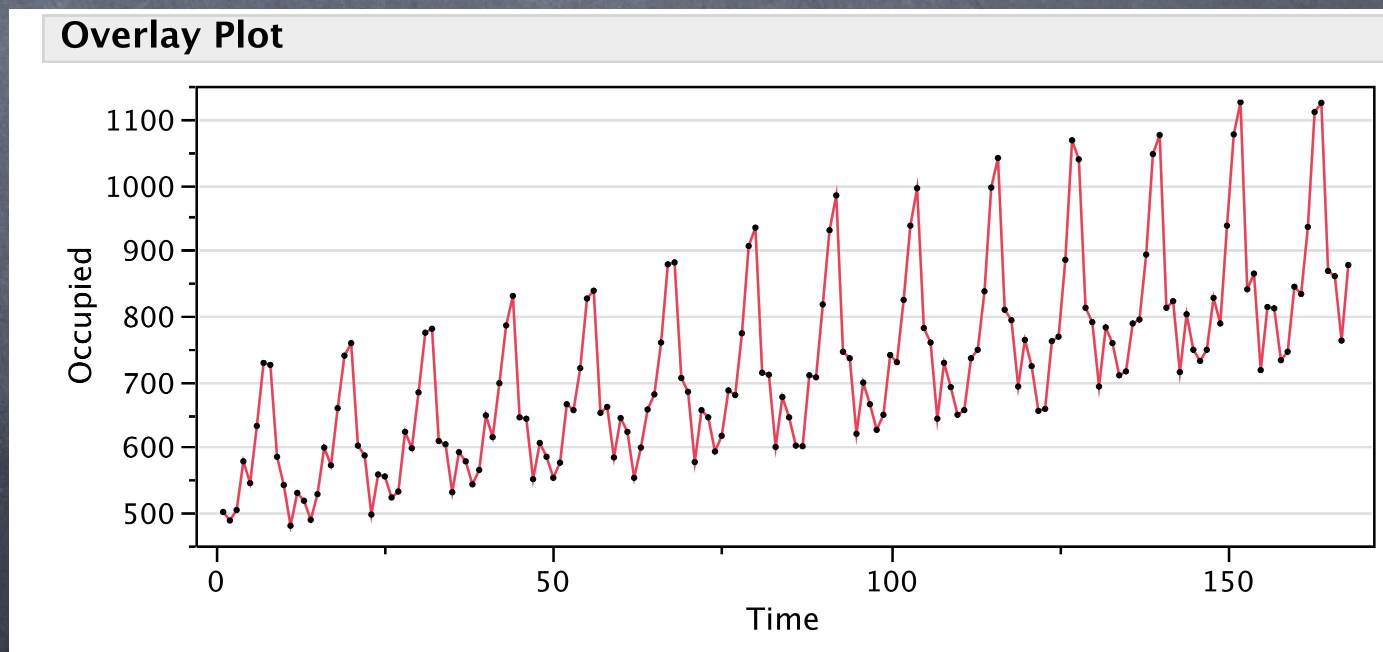
Categorical Variables

- 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

Table 6.4

- 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

Summary of Fit

RSquare	0.978941
RSquare Adj	0.977311
Root Mean Square Error	21.48822
Mean of Response	722.2976
Observations (or Sum Wgts)	168

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	12	3327046.9	277254	600.4501
Error	155	71570.2	462	**Prob > F**
C. Total	167	3398617.1		<.0001*

- Specific coefficients by-and-large differ

Indicator Function Parameterization

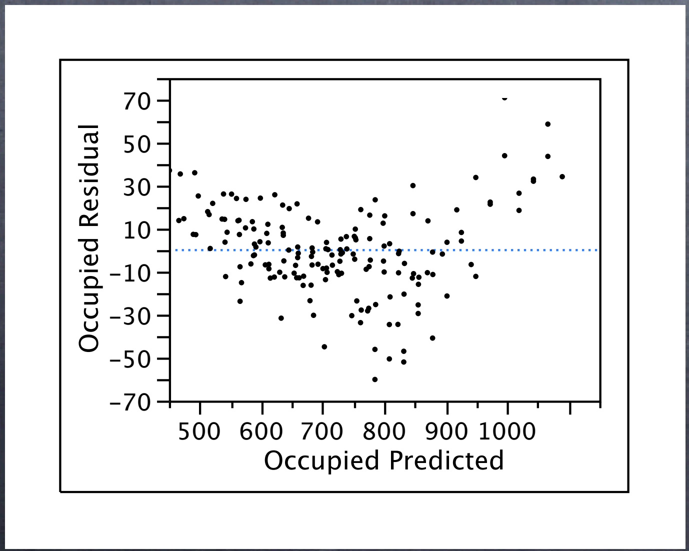
Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	518.86538	6.518866	155.00	79.59	<.0001*
Time	1.953083	0.034272	155.00	56.99	<.0001*
Month[Jan]	-27.01609	8.130527	155.00	-3.32	0.0011*
Month[Feb]	-71.82631	8.12901	155.00	-8.84	<.0001*
Month[Mar]	-56.13654	8.127637	155.00	-6.91	<.0001*
Month[Apr]	25.267521	8.126409	155.00	3.11	0.0022*
Month[May]	12.671581	8.125325	155.00	1.56	0.1209
Month[Jun]	106.43278	8.124385	155.00	13.10	<.0001*
Month[Jul]	229.19399	8.12359	155.00	28.21	<.0001*
Month[Aug]	250.66947	8.122939	155.00	30.86	<.0001*
Month[Sep]	38.216392	8.122433	155.00	4.71	<.0001*
Month[Oct]	27.406166	8.122072	155.00	3.37	0.0009*
Month[Nov]	-74.11835	8.121855	155.00	-9.13	<.0001*

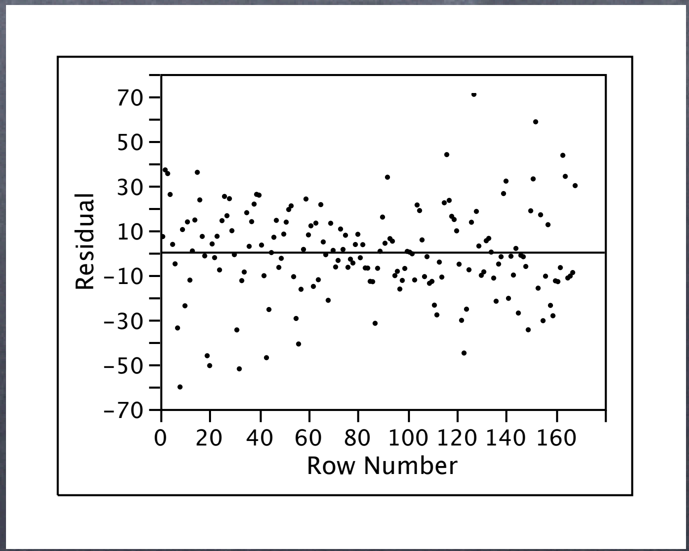
Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Time	1	1	1499569.3	3247.624	<.0001*
Month	11	11	1771253.7	348.7284	<.0001*

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

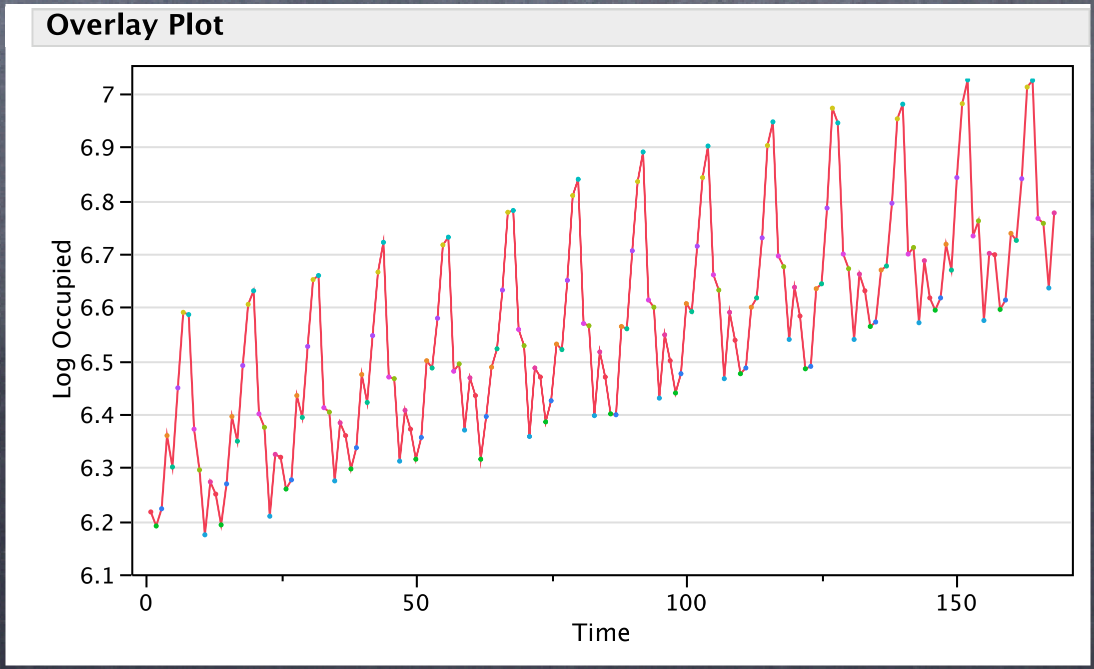




Two Ways to Fix

- 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)

Summary of Fit	
RSquare	0.988665
RSquare Adj	0.987787
Root Mean Square Error	0.021186
Mean of Response	6.563887
Observations (or Sum Wgts)	168

Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	6.2875573	0.006427	155.00	978.26	<.0001*
Time	0.0027253	3.379e-5	155.00	80.65	<.0001*
Month[Jan]	-0.041606	0.008016	155.00	-5.19	<.0001*
Month[Feb]	-0.112079	0.008015	155.00	-13.98	<.0001*
Month[Mar]	-0.084459	0.008013	155.00	-10.54	<.0001*
Month[Apr]	0.0398331	0.008012	155.00	4.97	<.0001*
Month[May]	0.0203951	0.008011	155.00	2.55	0.0119*
Month[Jun]	0.1469094	0.00801	155.00	18.34	<.0001*
Month[Jul]	0.2890226	0.008009	155.00	36.09	<.0001*
Month[Aug]	0.3111946	0.008009	155.00	38.86	<.0001*
Month[Sep]	0.0559872	0.008008	155.00	6.99	<.0001*
Month[Oct]	0.0395438	0.008008	155.00	4.94	<.0001*
Month[Nov]	-0.112215	0.008008	155.00	-14.01	<.0001*

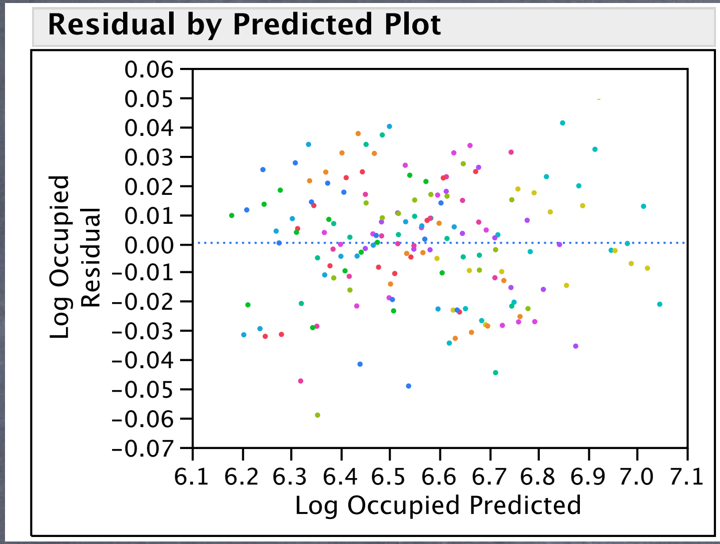
  - Do NOT compare R^2 statistic to prior model since the response variable is not the same as in the prior model

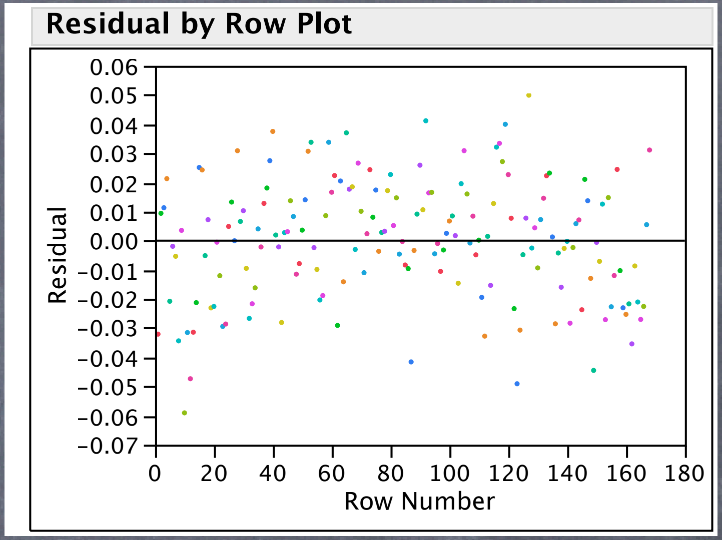
- Interpretation of slope for time
  - Rate of growth: about 0.3% per month

- 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?

Summary of Fit

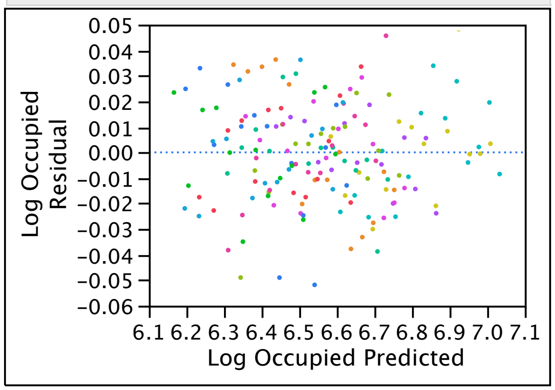
RSquare	0.989874
RSquare Adj	0.989019
Root Mean Square Error	0.02009
Mean of Response	6.563887
Observations (or Sum Wgts)	168

Indicator Function Parameterization

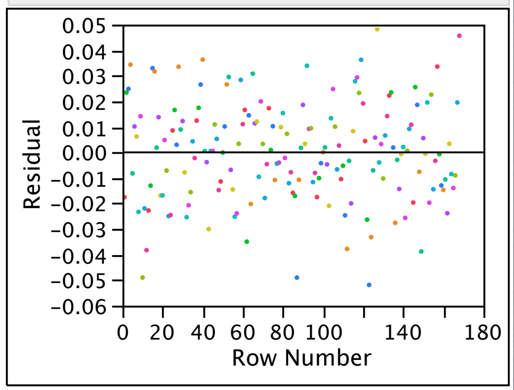
Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	6.2724878	0.007035	154.00	891.55	<.0001*
Time	0.0032592	0.000129	154.00	25.35	<.0001*
Time*Time	−3.159e−6	7.369e−7	154.00	−4.29	<.0001*
Month[Jan]	−0.041606	0.007601	154.00	−5.47	<.0001*
Month[Feb]	−0.112111	0.0076	154.00	−14.75	<.0001*
Month[Mar]	−0.084516	0.007599	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*

- Further structure?

Residual by Predicted Plot

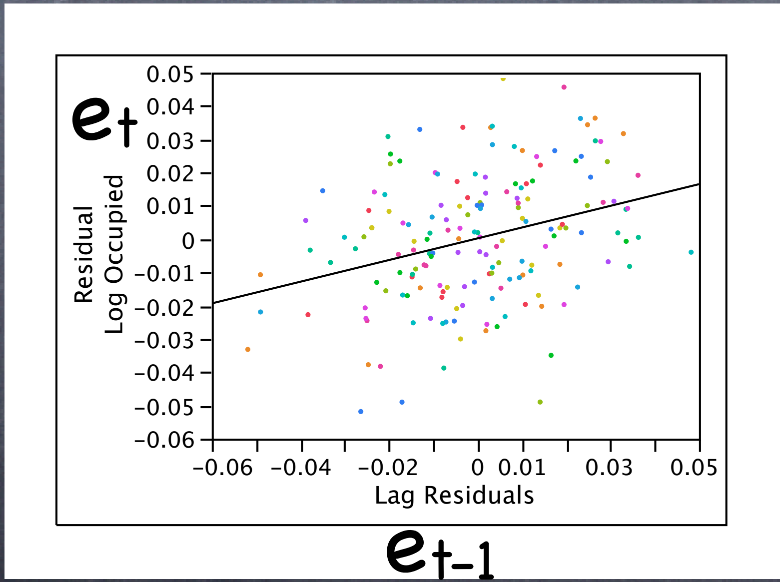


Residual by Row Plot



Testing Residual Dependence

- Durbin-Watson test
  - Test whether adjacent residuals appear dependent
  - Test related to <u>autocorrelation</u> between residuals
    - Autocorrelation is correlation between "rows" in the data table, whereas the usual correlation is between "columns"
- Lag plot of residuals



Linear Fit

Residual Log Occupied = 0.000194 + 0.3257149*Lag Residuals

Summary of Fit

RSquare	0.103026
RSquare Adj	0.09759
Root Mean Square Error	0.018336
Mean of Response	0.000105
Observations (or Sum Wgts)	167

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.000194	0.001419	0.14	0.8914
Lag Residuals	0.3257149	0.074819	4.35	<.0001*

- Regression summary

Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
1.3322276	168	0.3147	<.0001*

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

    $\varepsilon_t = \varphi\, \varepsilon_t + a_t,\qquad Var(a_t) = (1-\varphi^2)\, Var(\varepsilon_t) \le Var(\varepsilon_t)$

    $\hat{y}_{n+1} = b_0 + b_1\, x_{n+1,1} + \ldots + b_k\, x_{n+1,k} + \hat{\varphi}\, e_n$
- 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)

Summary of Fit

RSquare	0.990798
RSquare Adj	0.989951
Root Mean Square Error	0.019085
Mean of Response	6.565967
Observations (or Sum Wgts)	167

$\hat{\varphi}$

Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	6.2736436	0.006752	152.00	929.09	<.0001*
Time	0.0032199	0.000125	152.00	25.80	<.0001*
Time*Time	-2.932e-6	7.116e-7	152.00	-4.12	<.0001*
Month[Jan]	-0.039028	0.007362	152.00	-5.30	<.0001*
Month[Feb]	-0.112117	0.00722	152.00	-15.53	<.0001*
Month[Mar]	-0.084519	0.007219	152.00	-11.71	<.0001*
Month[Apr]	0.0397564	0.007217	152.00	5.51	<.0001*
Month[May]	0.0203076	0.007216	152.00	2.81	0.0055*
Month[Jun]	0.1468168	0.007216	152.00	20.35	<.0001*
Month[Jul]	0.2889308	0.007215	152.00	40.05	<.0001*
Month[Aug]	0.3111094	0.007214	152.00	43.12	<.0001*
Month[Sep]	0.0559145	0.007214	152.00	7.75	<.0001*
Month[Oct]	0.0394895	0.007214	152.00	5.47	<.0001*
Month[Nov]	-0.112245	0.007213	152.00	-15.56	<.0001*
Lag Residuals	0.328304	0.078043	152.00	4.21	<.0001*

- 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
  - $\hat{y}_{n+j} = (6.2736 + b_j) +$ seasonal
    $0.00322\,(n+j) - 0.00000293(n+j)^2 +$ time trend
    $0.328^j\,(e_n)$ autocorr
  - Autocorrelation effect drops off geometrically, having little influence past a few terms

- Point estimates for January, February
  - $\hat{y}_{168+1} = (6.2736-0.0390) +$
    $0.00322\,(169) - 0.00000293(169)^2 +$
    $0.328\,(0.0456)$
    $\approx 6.2346 + 0.4605 + 0.0150 = \underline{6.7101}$
  - $\hat{y}_{168+2} = (6.2736-0.1121) +$
    $0.00322\,(170) - 0.00000293(170)^2 +$
    $0.328^2\,(0.0456)$
    $\approx 6.1615 + 0.4627 + 0.0049 = \underline{6.6291}$

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 <u>approximate</u> intervals; JMP shown next

- One period out: use RMSE of fitted model
  - $\hat{y}_{168+1} \pm t_{.025,152}$ RMSE $= 6.7101 \pm 1.98\ (0.0191)$
    ≈ 6.6723 to 6.7479

- Two periods out: inflate RMSE by sqrt$(1+\hat{\varphi}^2)$
  - $\hat{y}_{168+2} \pm t_{.025,152}\text{RMSE}(1+\hat{\varphi}^2)^{1/2} = 6.6291 \pm 1.98(0.0191)(1+.328^2)^{1/2}$
    ≈ 6.589 to 6.669

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

JMP Calculations

- Prediction interval

  $\hat{y} \pm t_{.025}$ RMSE (Extrapolation) (Autocorrelation)

  "distance value"

- Four components determine width of interval

1. t-percentile…	≈ 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 φ is close to 1.

JMP Calculations, cntd

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

$$\sqrt{1+\hat{\varphi}^2+\hat{\varphi}^4+\ldots+\hat{\varphi}^{2(m-1)}}$$

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

	Lag Residuals	Pred Formula Log	StdErr Indiv Log	Lower 95% Indiv Log	Upper 95% Indiv Log	RMSE Adjustment	Corrected 95% PI,	Corrected 95% PI,
161	-0.0146867	6.73155	0.02007	6.69189	6.77121	•	•	•
162	-0.0106001	6.86167	0.02007	6.82202	6.90133	•	•	•
163	-0.0238339	7.00171	0.02015	6.96190	7.04152	•	•	•
164	0.00338923	7.03509	0.02010	6.99538	7.07479	•	•	•
165	-0.0084815	6.77825	0.02011	6.73851	6.81799	•	•	•
166	-0.0139638	6.76228	0.02014	6.72248	6.80207	•	•	•
167	-0.0090122	6.61441	0.02015	6.57461	6.65421	•	•	•
168	0.01952859	6.73826	0.02023	6.69830	6.77823	•	•	•
169	0.04564133	6.71004	0.02069	6.66917	6.75091	1	6.66949	6.75059
170	0.015	6.62912	0.02029	6.58903	6.66920	1.0525	6.58726	6.67097

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)

$$6.6695 \text{ to } 6.7479 \quad \Rightarrow \quad e^{6.6723} \text{ to } e^{6.7506}$$

790 to 855 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