Stat 102

Introduction to Business Statistics
Class 21

Richard P. Waterman

Wharton

Table of contents I

Today's module

Topics to be covered in this module:

- Last time
- Regression models for time series
- Correlated residuals
- The Durbin Watson test
- One-step ahead forecast with lagged residuals
- Summary
- Next time

Last time

- Introduction to logistic regression
- Quality of fit using the false positive and false negative rates
- RoC curves
- Lift curves

Time series

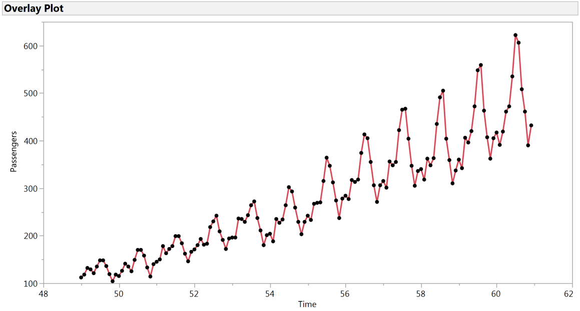
- We think of data for regression models being of one of three types:
    1. Cross-sectional: Data that is collected on many subjects at a single point in time (or ignoring time)
    2. Time series: collected on one subject, over the course of time
    3. Panel data: data collected on many subjects at multiple points in time
- We have been treating all data as cross-sectional
- This class will discuss regression for time-series data (Stine Chapter 27-2)

Time series data

- Time series data often requires additional thought because:
  1. It can display seasonality
  2. This independence assumption on the error terms is often broken.

Prediction airline passenger demand

- The file IntlAir-reg.JMP contains monthly passenger data (in units of 1,000 passengers) from 1949 to 1960, a period of rapid growth.
- A time series graph can be produced using the Graph→Overlay Plot in JMP:

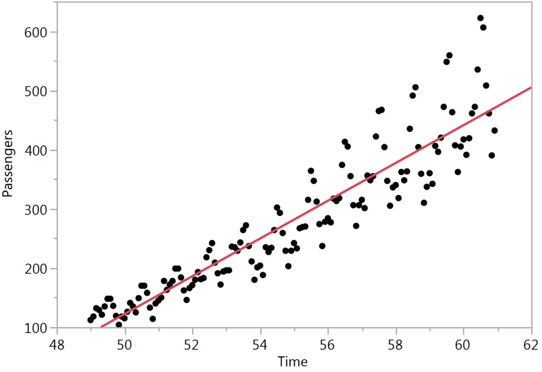


- Multiple regression provides an approach for modeling this kind of data

A simple regression of passengers against time

We begin with a simple regression of Passengers on Time, a simple linear trend model that captures much of the variation in Passengers.

Bivariate Fit of Passengers By Time



Linear Fit

Passengers = -1472.114 + 31.886207*Time

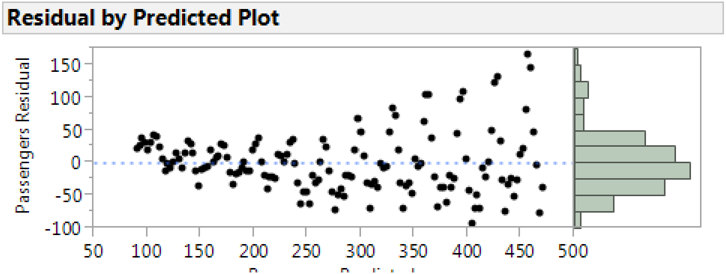
Summary of Fit

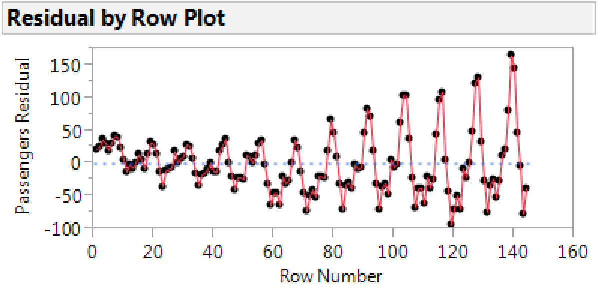
RSquare	0.854
RSquare Adj	0.853
Root Mean Square Error	46.057
Mean of Response	280.299
Observations (or Sum Wgts)	144.000

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1472.114	61.014	-24.13	<.0001*
Time	31.886	1.108	28.78	<.0001*

What about the residuals?



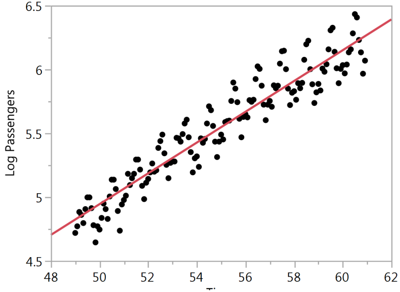


Issues with the residuals

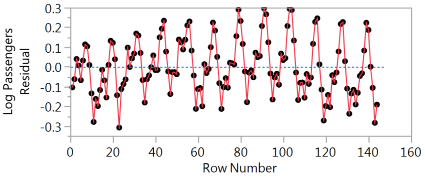
- Problems:
  1. Lack of constant variance
  2. Curvature, indicating a lack of fit of the model to the data
  3. A very strong cyclical pattern, that is, positive autocorrelation
- We will attempt to fix these problems:
- The linear model doesn't make a lot of sense with its implied constant increase of 32,000 passengers a year.
- An exponential growth process may be more reasonable, implying a constant proportional change (rather than additive change).
- We model exponential growth by taking the log of Y.

The log passengers model

Bivariate Fit of Log Passengers By Time



Residual by Row Plot



Diagnostics Plots

Log Passengers = -1.084732 + 0.1205806*Time

Summary of Fit

RSquare	0.9015
RSquare Adj	0.900807
Root Mean Square Error	0.139037
Mean of Response	5.542176
Observations (or Sum Wgts)	144

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.084732	0.184189	-5.89	<.0001*
Time	0.1205806	0.003345	36.05	<.0001*

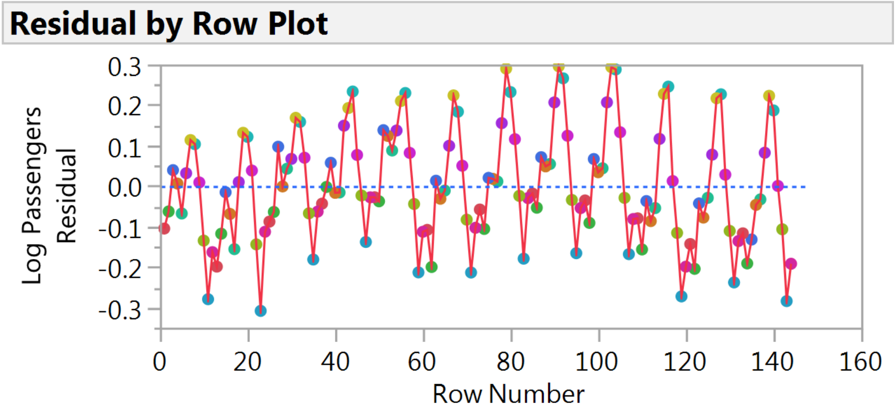
Improvements

Things are looking better:

- Much of the curvature has gone
- The variance has been stabilized
- The model implies constant 12% growth rate a year

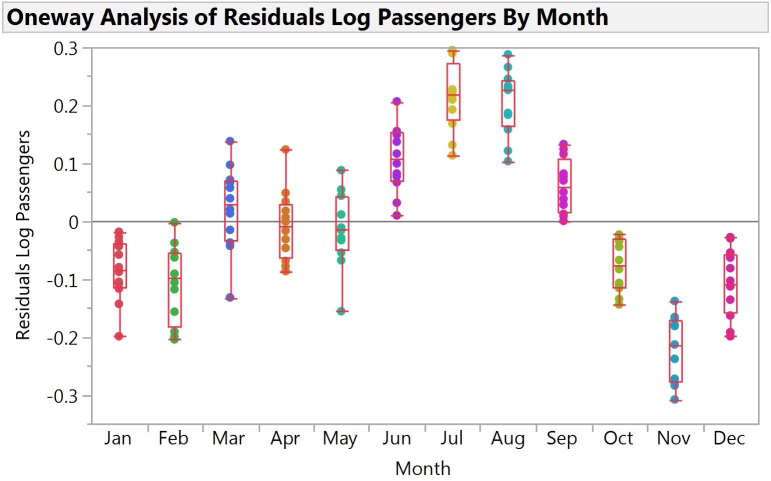
Color coding the residual by month

The color coding reveals strong seasonality (it will help to label by month):



Comparison box plots of the residuals against month

Plot the residuals over the level of the categorical variable, month, and the story is much clearer. There is a strong monthly **seasonal effect**:



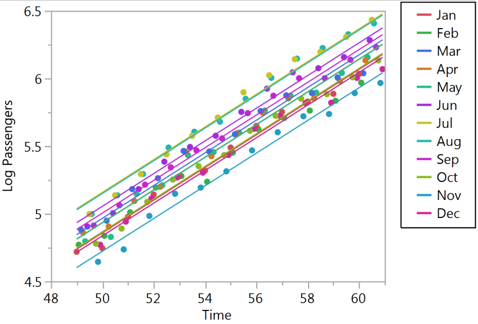
Using a categorical variable to capture a seasonal effect

Dropping the seasonal effect into the model as a multi-level categorical will capture most of the systematic variation in the residuals:

Response Log Passengers

Whole Model

Regression Plot



Summary of Fit

RSquare	0.983468
RSquare Adj	0.981954
Root Mean Square Error	0.059304
Mean of Response	5.542176
Observations (or Sum Wgts)	144

Indicator Function Parameterization

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.2049	0.0812	-14.85	<.0001*
Time	0.1208	0.0014	84.40	<.0001*
Month[Jan]	0.0213	0.0242	0.88	0.3808
Month[Feb]	-0.0007	0.0242	-0.03	0.9759
Month[Mar]	0.1295	0.0242	5.34	<.0001*
Month[Apr]	0.0982	0.0242	4.05	<.0001*
Month[May]	0.0959	0.0242	3.96	0.0001*
Month[Jun]	0.2180	0.0242	9.00	<.0001*
Month[Jul]	0.3219	0.0242	13.29	<.0001*
Month[Aug]	0.3126	0.0242	12.91	<.0001*
Month[Sep]	0.1680	0.0242	6.94	<.0001*
Month[Oct]	0.0299	0.0242	1.23	0.2198
Month[Nov]	-0.1139	0.0242	-4.70	<.0001*

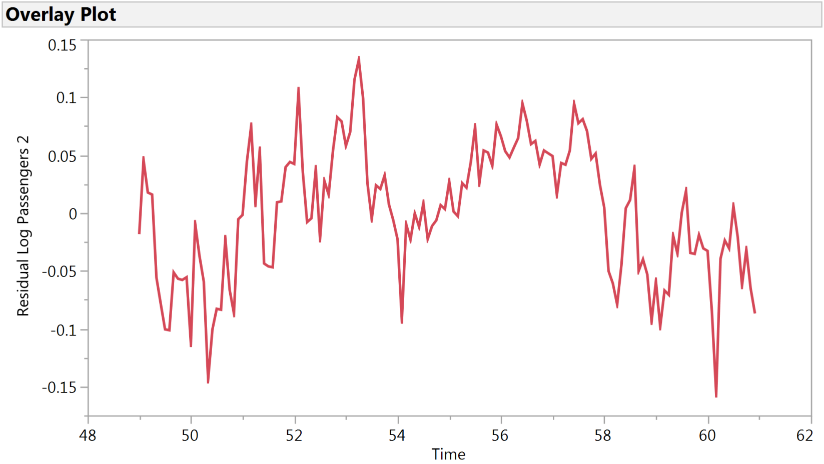
As the lines are parallel, the slope is the same for all months. That is we still have the same constant 12% annual growth rate, for all months.

Criticism's

1. The model is getting better
2. R^2 went up by a significant amount
3. That is, the added month term was significant (review the Effect test table)
4. RMSE has dropped from 0.14 to 0.06.
5. But we still need to check the residuals

Positive autocorrelation

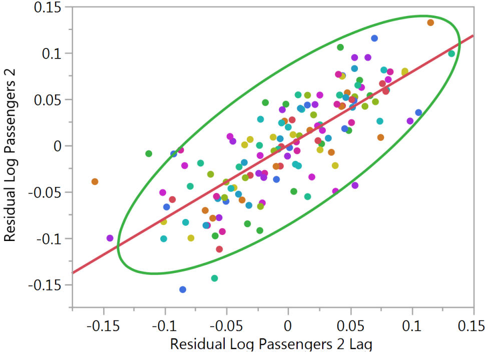
Plotting the residuals over time shows a meandering pattern



Positive autocorrelation

The scatterplot of adjacent residuals shows the positive autocorrelation:

Bivariate Fit of Residual Log Passengers 2 By Residual Log Passengers 2 Lag



Bivariate Normal Ellipse P=0.950

Variable	Mean	Std Dev	Correlation	Signif. Prob	Number
Residual Log Passengers 2 Lag	0.000608	0.056488	0.785619	<.0001*	143
Residual Log Passengers 2	0.000128	0.056939			

The Durbin Watson test

This [1] test looks to see if the lag 1 autocorrelation is statistically significant. In this case, it is:

Durbin-Watson			
Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
0.4251836	144	0.7788	<.0001*

[1] From the regression output, click on the title bar, row diagnostics, then Durbin Watson Test

The Durbin Watson test

- This test on the adjacent residuals looks to see if there is sequential (serial/auto) correlation between the adjacent error terms ϵ_i.
- The form of the test is:

$$D = \frac{\sum_{i=2}^{n}(e_t - e_{t-1})^2}{\sum_{i=1}^{n} e_t^2}.$$

- When the correlation is 0, D is expected to be 2. Values of D between 0 and 2 imply positive autocorrelation and values of D between 2 and 4 negative autocorrelation.

Adding the lagged residuals to the model

- Adding the lagged residuals to the model can be an effective way of dealing with autocorrelation
- It is appropriate when the goal is forecasting, not so much interpretation
- The idea is that if the residuals (forecast errors) are correlated you can use last times residual to predict this times residual.
- If you have a prediction for what the forecast error will be, you can then directly adjust the forecast, by using this prediction of the forecast error
- But you will not know whether it works well until you try it

The multiple regression, including the lagged residual

Summary of Fit

RSquare	0.993531
RSquare Adj	0.992879
Root Mean Square Error	0.036922
Mean of Response	5.547936
Observations (or Sum Wgts)	143

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Time	1	1	24.434696	17924.16	<.0001*
Month	11	11	2.284770	152.3637	<.0001*
Residual Log Passengers 2 Lag	1	1	0.284485	208.6847	<.0001*

Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
2.191914	143	-0.1089	0.8478

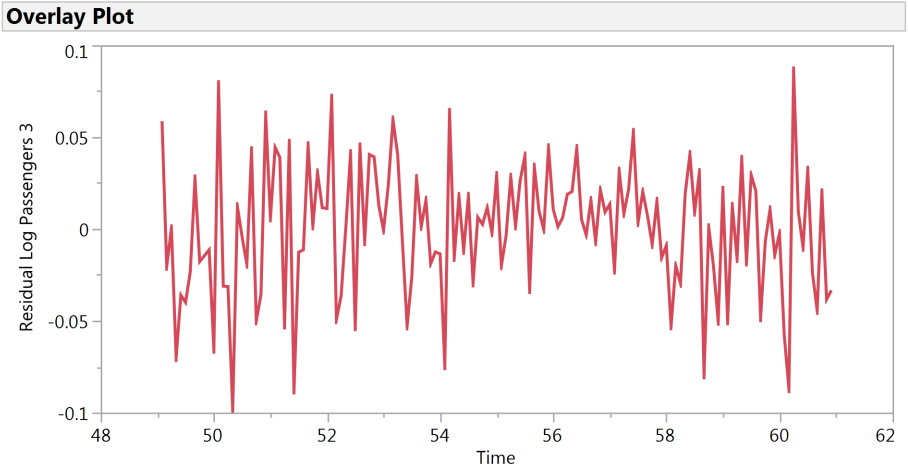
Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	-1.187682	0.05101	129.00	-23.28	<.0001*
Time	0.1205144	0.0009	129.00	133.88	<.0001*
Month[Jan]	0.0165913	0.015423	129.00	1.08	0.2840
Month[Feb]	-0.000993	0.015092	129.00	-0.07	0.9476
Month[Mar]	0.1292599	0.015088	129.00	8.57	<.0001*
Month[Apr]	0.098017	0.015085	129.00	6.50	<.0001*
Month[May]	0.0956703	0.015082	129.00	6.34	<.0001*
Month[Jun]	0.2178425	0.01508	129.00	14.45	<.0001*
Month[Jul]	0.3218107	0.015078	129.00	21.34	<.0001*
Month[Aug]	0.3125418	0.015076	129.00	20.73	<.0001*
Month[Sep]	0.1679331	0.015075	129.00	11.14	<.0001*
Month[Oct]	0.0298008	0.015074	129.00	1.98	0.0502
Month[Nov]	-0.113891	0.015073	129.00	-7.56	<.0001*
Residual Log Passengers 2 Lag	0.7930716	0.054899	129.00	14.45	<.0001*

- The Durbin Watson statistic is no longer significant
- The lagged residuals coefficient is highly significant
- Both the monthly seasonal effect and the overall time trend (Year) are highly significant
- RMSE has decreased to 0.037

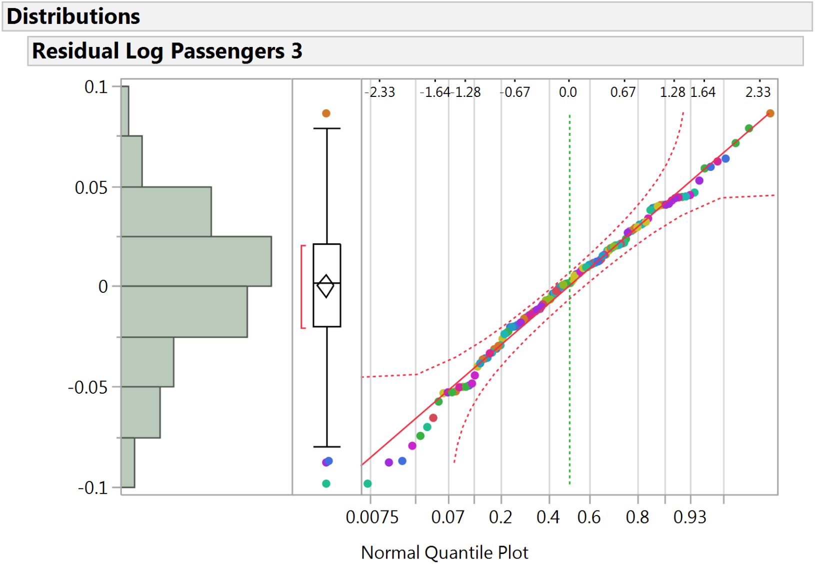
A time-series plot of the residuals

This is much better looking. No evident meandering as confirmed by the DW-statsitic.



The NQP of the residuals

This looks good too:



Forecasting

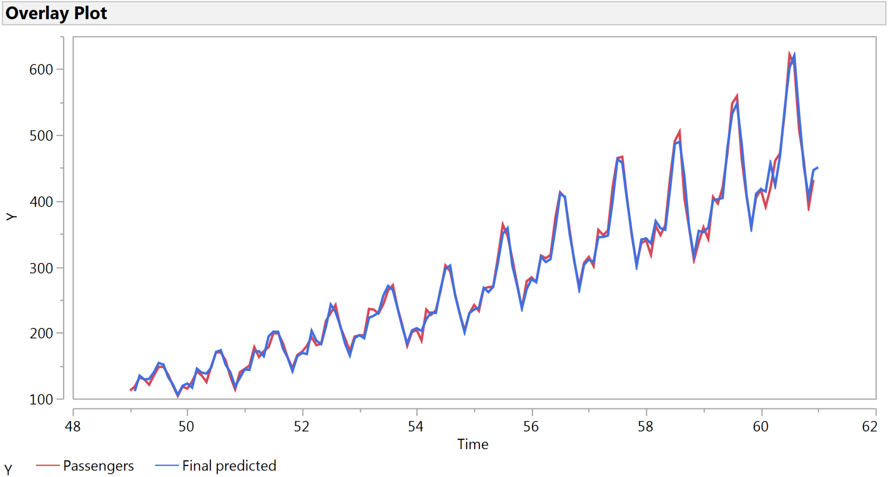
- We will use this model to predict the number of passengers in January 1961
- We will do everything on the log-scale first, then at the very last step, back-transform to the original scale
- The components of the prediction:

$$\begin{aligned}\hat{y}_{Jan,1961} &= \underbrace{-1.188}_{Intercept} + \underbrace{0.12 \times 61}_{Year\ effect} + \underbrace{0.0166}_{Jan\ effect} + \underbrace{0.7931 \times (-0.0335)}_{Previous\ residual} \\ &= 6.111\end{aligned}$$

- Add on +/- 2 RMSE to get on the log-scale: (6.0372, 6.1848)
- Exponeniate to finally get the 95% prediction interval of: (418.72, 485.32), which is in thousands

Plotting the observed and expected frequencies

Using the time-series Overlay plot we can plot the actual and forecast values together:



Summary

- Regression models for time series
- Correlated residuals
- The Durbin Watson test
- One-step ahead forecast with lagged residuals