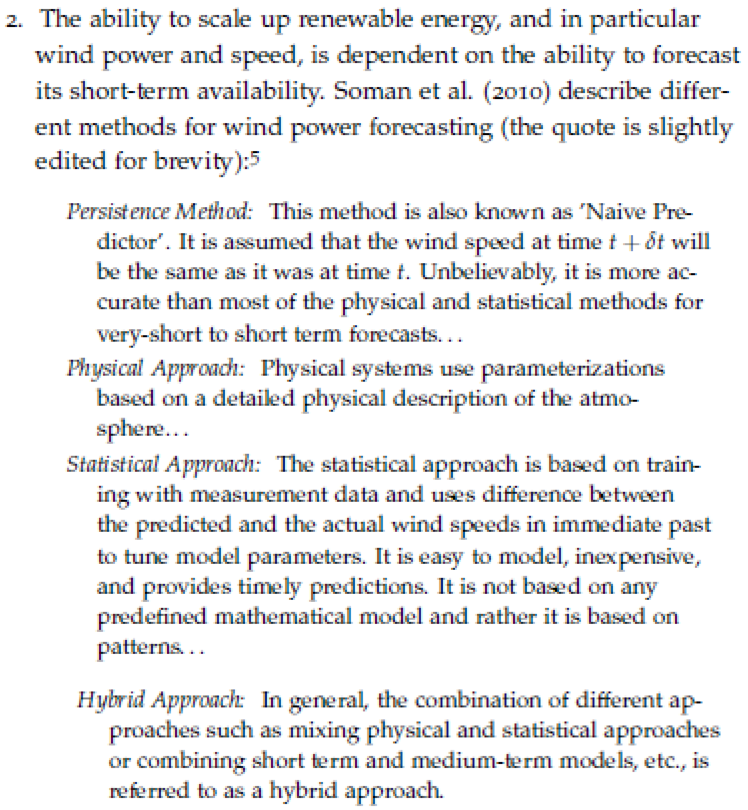
Statistics.com

Forecasting: Assignment #2 Solutions

Chapter 4 Problem #2



2. The ability to scale up renewable energy, and in particular wind power and speed, is dependent on the ability to forecast its short-term availability. Soman et al. (2010) describe different methods for wind power forecasting (the quote is slightly edited for brevity):[5]

    Persistence Method: This method is also known as 'Naive Predictor'. It is assumed that the wind speed at time $t + \delta t$ will be the same as it was at time t. Unbelievably, it is more accurate than most of the physical and statistical methods for very-short to short term forecasts...

    Physical Approach: Physical systems use parameterizations based on a detailed physical description of the atmosphere...

    Statistical Approach: The statistical approach is based on training with measurement data and uses difference between the predicted and the actual wind speeds in immediate past to tune model parameters. It is easy to model, inexpensive, and provides timely predictions. It is not based on any predefined mathematical model and rather it is based on patterns...

    Hybrid Approach: In general, the combination of different approaches such as mixing physical and statistical approaches or combining short term and medium-term models, etc., is referred to as a hybrid approach.

(a) For each of the four types of methods, describe whether it is model-based, data-driven, or a combination.

The "persistent method" is naïve forecasting, and is data-driven.

The "physical approach" is a model-based method, which rely on a physical model.

The "statistical approach" as described here is a model-based approach, where a statistical model (such as linear regression) is estimated from the training data

The “hybrid approach” combines different methods and is therefore an ensemble of model-based and data-driven methods.

(b) For each of the four types of methods, describe whether it is based on extrapolation, causal modeling, correlation modeling or a combination.

“Persistent method” is based on extrapolation

“Physical approach” is causal based

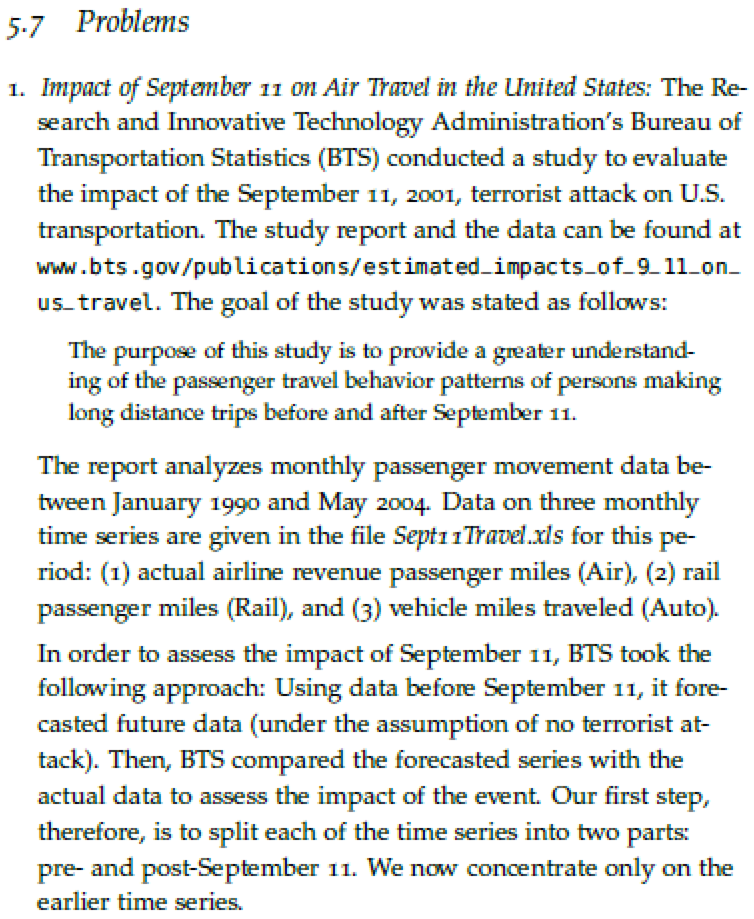
“Statistical approach” is based on extrapolation and possibly correlation

“Hybrid approach” combines all

(c) Describe the advantages and disadvantages of the hybrid approach.

The hybrid approach is more complicated to build, as it requires fitting all the different methods and models. Its potential advantage is that it will produce more precise forecasts and will be more robust over time.

Chapter 5 Problem #1



5.7 Problems

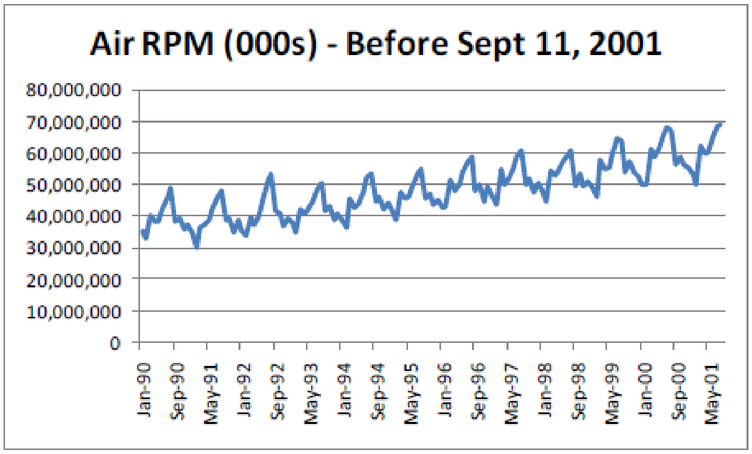
1. *Impact of September 11 on Air Travel in the United States:* The Research and Innovative Technology Administration's Bureau of Transportation Statistics (BTS) conducted a study to evaluate the impact of the September 11, 2001, terrorist attack on U.S. transportation. The study report and the data can be found at `www.bts.gov/publications/estimated_impacts_of_9_11_on_us_travel`. The goal of the study was stated as follows:

   > The purpose of this study is to provide a greater understanding of the passenger travel behavior patterns of persons making long distance trips before and after September 11.

   The report analyzes monthly passenger movement data between January 1990 and May 2004. Data on three monthly time series are given in the file *Sept11Travel.xls* for this period: (1) actual airline revenue passenger miles (Air), (2) rail passenger miles (Rail), and (3) vehicle miles traveled (Auto).

   In order to assess the impact of September 11, BTS took the following approach: Using data before September 11, it forecasted future data (under the assumption of no terrorist attack). Then, BTS compared the forecasted series with the actual data to assess the impact of the event. Our first step, therefore, is to split each of the time series into two parts: pre- and post-September 11. We now concentrate only on the earlier time series.

(a) Plot the pre-event Air time series. Which time series components appear from the plot?



Level, trend, seasonality and noise

(b) The figure below is a time plot of the seasonally adjusted pre-September-11 Air series. Which of the following methods would be adequate for forecasting this series?

- Linear regression model with dummy variables
- [✓ Linear regression model with trend]
- Linear regression model with dummy variables and trend

(c) Specify a linear regression model for the Air series that would produce a seasonally adjusted series similar to the one shown in (b), with multiplicative seasonality. What is the output variable? What are the predictors?

Output variable = log Y

Predictor = 11 dummies, for 11 of the 12 months (one reference category)

(d) Run the regression model from (c). Remember to create dummy variables for the months (XLMiner will create 12 dummy variables - use 11 only and drop the April dummy) and to use only pre-event data. Check the option in XLMiner to obtain fitted values and unstandardized residuals.

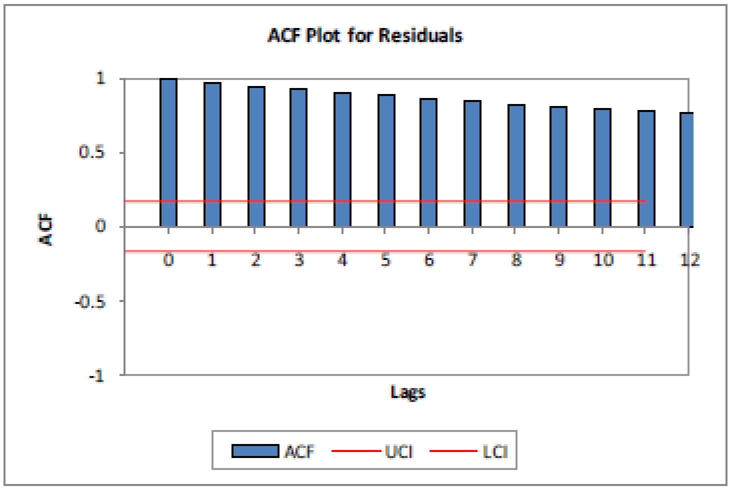
i. What can we learn from the statistical insignificance of the coefficients for October and September?

We learn that Oct and Sept are similar to April (the reference category) in terms of AIR RPM.

ii. The actual value of Air (air revenue passenger miles) in January 1990 was 35.153577 billion. What is the residual for this month, using the regression model? Report the residual in terms of air revenue passenger miles.

The fitted value is exp(17.55979347)= 42278678.81 billion, so the residual is -7125101.814

(e) Create an ACF (autocorrelation) plot of the regression residuals.



i. What does the ACF plot tell us about the regression model's forecasts?

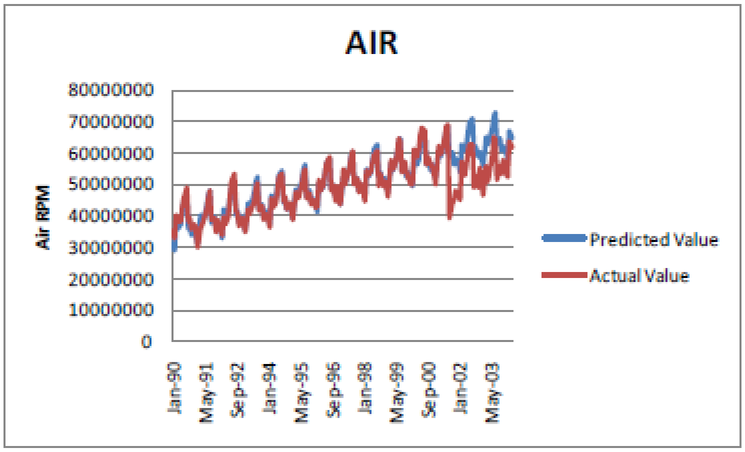
The positive autocorrelation at many lags tells us that the forecast errors from this model are positively correlated. If we over-predict in a certain month, we are likely to over-predict the following month as well.

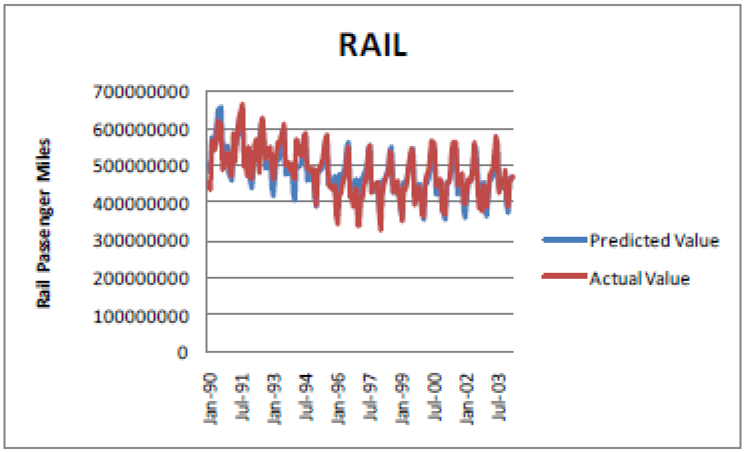
ii. How can this information be used to improve the model?

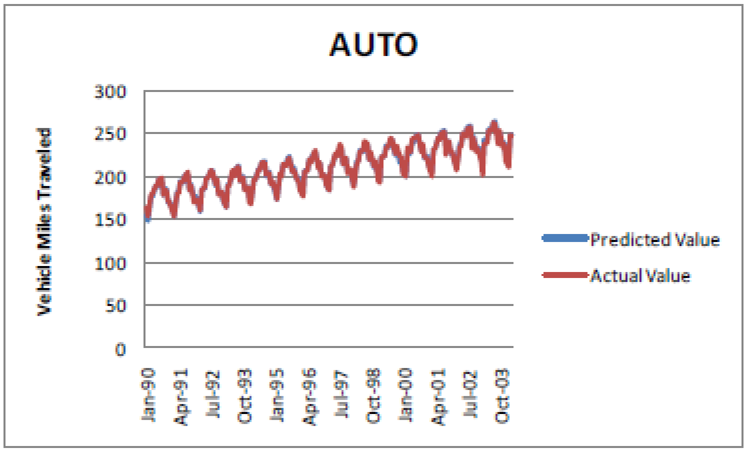
By creating a second-level where we model the residuals using an AR model. (Alternatively, by incorporating lagged periods as predictors, within an ARIMA model).

(f) Fit linear regression models to Air, Rail and Auto with additive seasonality and an appropriate trend. For Air and Auto, fit a linear trend. For Rail, use a quadratic trend. Remember to use only pre-event data. Once the models are estimated, use them to forecast each of the three post-event series.

i. For each series (Air, Rail, Auto), plot the complete prevent and post-event actual series overlaid with the predicted series.


AIR


RAIL


AUTO

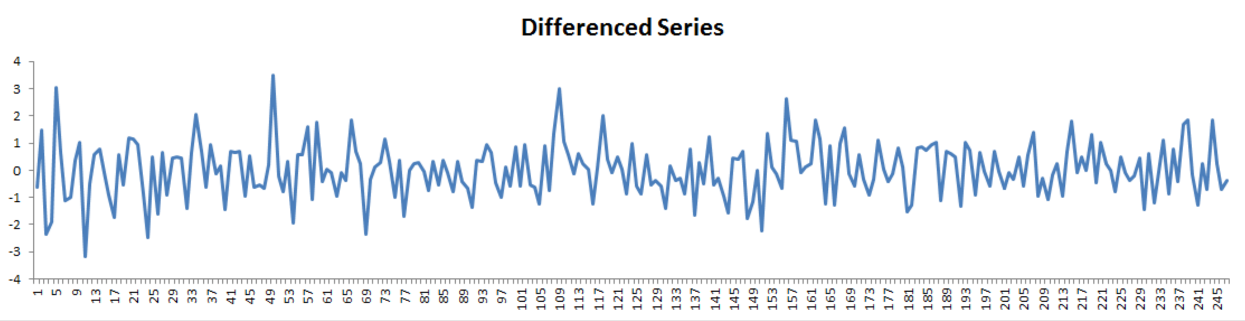
ii. What can be said about the effect of the September 11 terrorist attack on the three modes of transportation? Discuss the magnitude of the effect, its time span, and any other relevant aspect.

For Rail and Auto, there is hardly any change. For Air, there is an immediate drop after Sept 11, 2001 of approximately 30%, a modest portion of which is recovered by spring of 2002.

Chapter 5 Problem #4

4. *Forecasting Wal-Mart Stock:* Figure 5.32 shows a time plot of Wal-Mart daily closing prices between February 2001 and February 2002. (Thanks to Chris Albright for suggesting the use of these data, publicly available at `finance.yahoo.com` and in *WalMartStock.xls*.). Figure 5.33 shows the output from fitting an AR(1) model to the series of closing prices and to the series of differences. Use all the information to answer the following questions.

(a) Create a time plot of the differenced series.



(b) Which of the following is/are relevant for testing whether this stock is a random walk?

- The autocorrelations of the closing price series
- [✓ The AR(1) slope coefficient for the closing price series]
- The AR(1) constant coefficient for the closing price series
- [✓ The autocorrelations of the differenced series]
- The AR(1) slope coefficient for the differenced series
- The AR(1) constant coefficient for the differenced series

(c) Recreate the AR(1) model output for the Close price series shown in the left panel of Figure 5.33. Figure 5.34 shows how to set parameters in XLMiner's ARIMA screen to fit an AR(1) model. Does the AR model indicate that this is a random walk? Explain how you reached your conclusion.

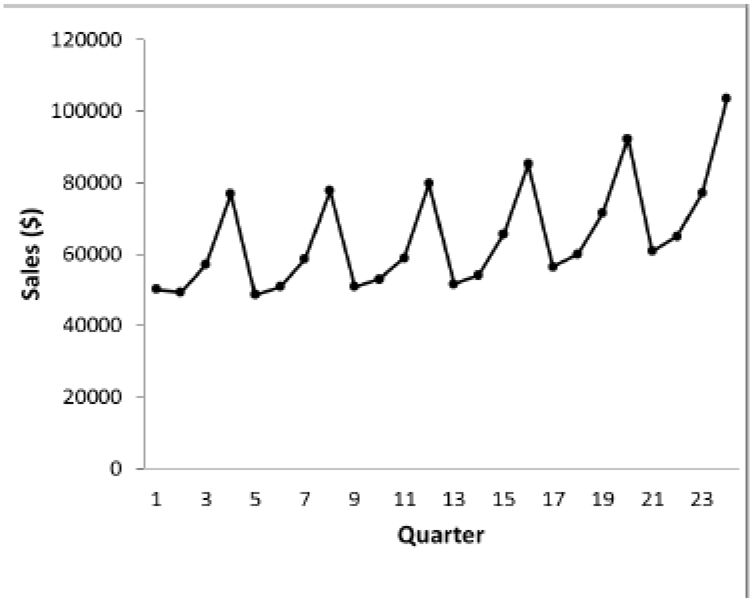
First, note that the ACF plot indicates no substantial autocorrelation for the differences. Fitting an AR(1) model to the close price series yields a slope coefficient of 0.956 with standard error 0.019. The slope coefficient is 2.36 standard errors away from the value 1. At a 5% significance level we'd say that this is not a random walk, but at 1% it is. In conclusion, the series appears to be close to a random walk.

(d) What are the implications of finding that a time series is a random walk? Choose the correct statement(s) below.

- [✓ It is impossible to obtain useful forecasts of the series.]
- The series is random.
- [✓ The changes in the series from one period to the other are random.]

Chapter 5 Problem #5

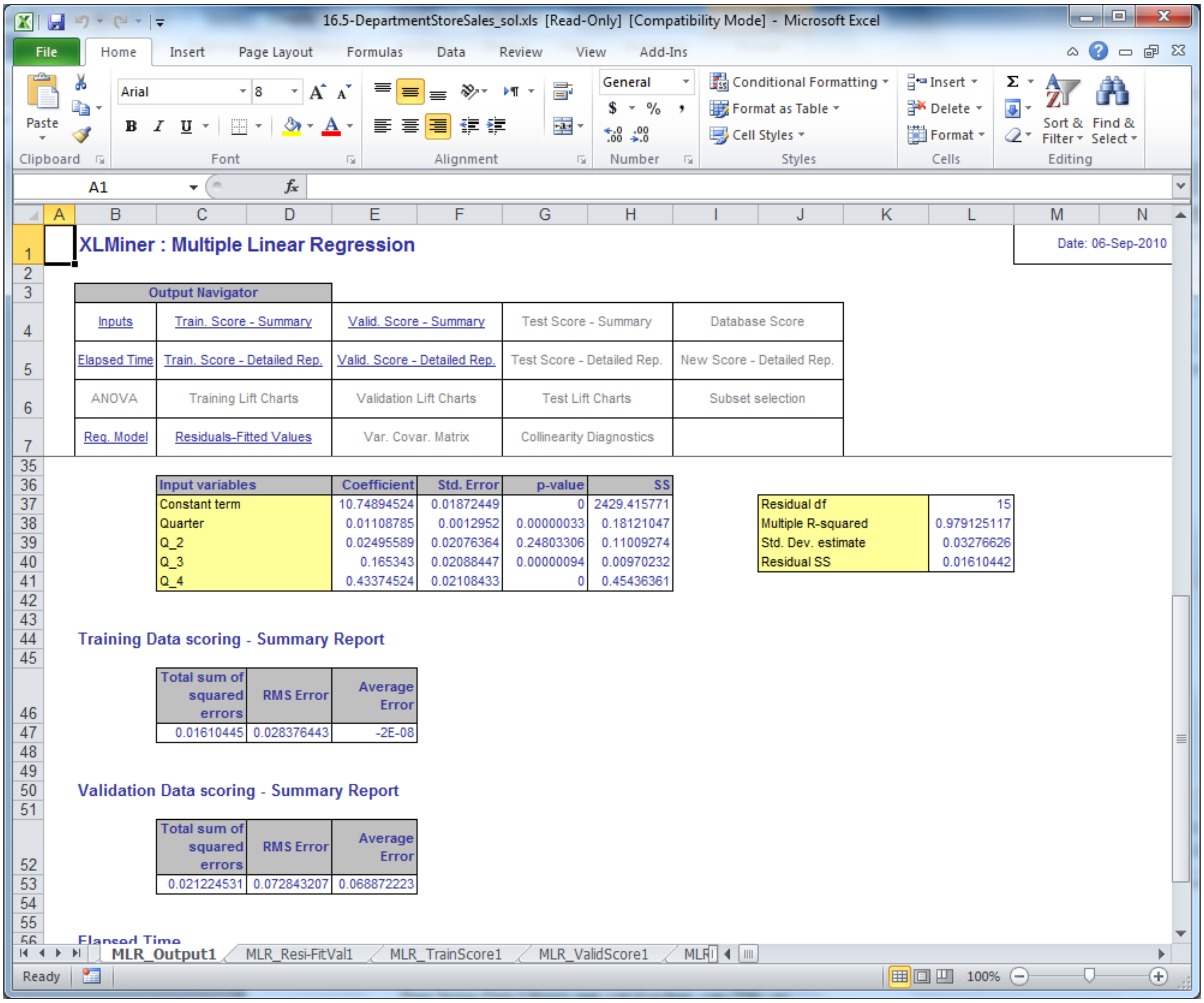
5. *Forecasting Department Store Sales:* The time series plot shown in Figure 5.36 describes actual quarterly sales for a department store over a 6-year period. (Data available in *DepartmentStoreSales.xls*, courtesy of Chris Albright.)



(a) The forecaster decided that there is an exponential trend in the series. In order to fit a regression-based model that accounts for this trend, which of the following operations must be performed?

- Take a logarithm of the Quarter index
- [✓ Take a logarithm of sales]
- Take an exponent of sales
- Take an exponent of Quarter index

(b) Fit a regression model with an exponential trend and seasonality, using only the first 20 quarters as the training period (remember to first partition the series into training and validation periods).



XLMiner : Multiple Linear Regression

Date: 06-Sep-2010

Output Navigator

Inputs	Train. Score - Summary	Valid. Score - Summary	Test Score - Summary	Database Score
Elapsed Time	Train. Score - Detailed Rep.	Valid. Score - Detailed Rep.	Test Score - Detailed Rep.	New Score - Detailed Rep.
ANOVA	Training Lift Charts	Validation Lift Charts	Test Lift Charts	Subset selection
Reg. Model	Residuals-Fitted Values	Var. Covar. Matrix	Collinearity Diagnostics	

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	10.74894524	0.01872449	0	2429.415771
Quarter	0.01108785	0.0012952	0.00000033	0.18121047
Q_2	0.02495589	0.02076364	0.24803306	0.11009274
Q_3	0.165343	0.02088447	0.0000094	0.00970232
Q_4	0.43374524	0.02108433	0	0.45436361

Residual df	15
Multiple R-squared	0.979125117
Std. Dev. estimate	0.03276626
Residual SS	0.01610442

Training Data scoring - Summary Report

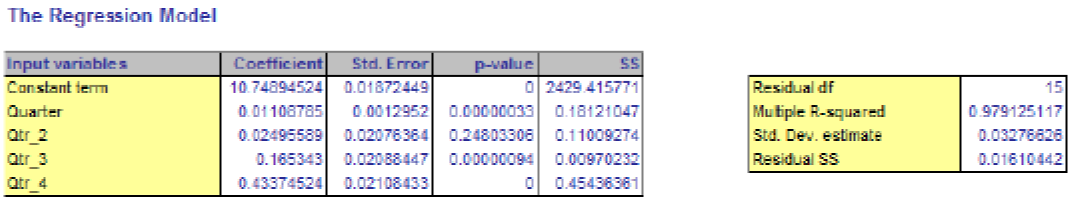
Total sum of squared errors	RMS Error	Average Error
0.01610445	0.028376443	-2E-08

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
0.021224531	0.072843207	0.068872223

Elapsed Time

(c) A partial output is shown in Figure 5.37. From the output, after adjusting for trend, are Q2 average sales higher, lower, or approximately equal to the average Q1 sales?



The Regression Model

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	10.74894524	0.01872449	0	2429.415771
Quarter	0.01108785	0.0012952	0.00000033	0.18121047
Qtr_2	0.02495589	0.02076364	0.24803306	0.11009274
Qtr_3	0.165343	0.02088447	0.00000094	0.00970232
Qtr_4	0.43374524	0.02108433	0	0.45436361

Residual df	15
Multiple R-squared	0.979125117
Std. Dev. estimate	0.03276626
Residual SS	0.01610442

(d) Use this model to forecast sales in quarters 21 and 22.

By plugging in the appropriate predictors into the regression model (or using the regression output) we get the predicted values:

Q 21: 10.98179009 , Q 22: 11.01783383

As we have fitted an exponential seasonal model it produced forecast for log(sales).

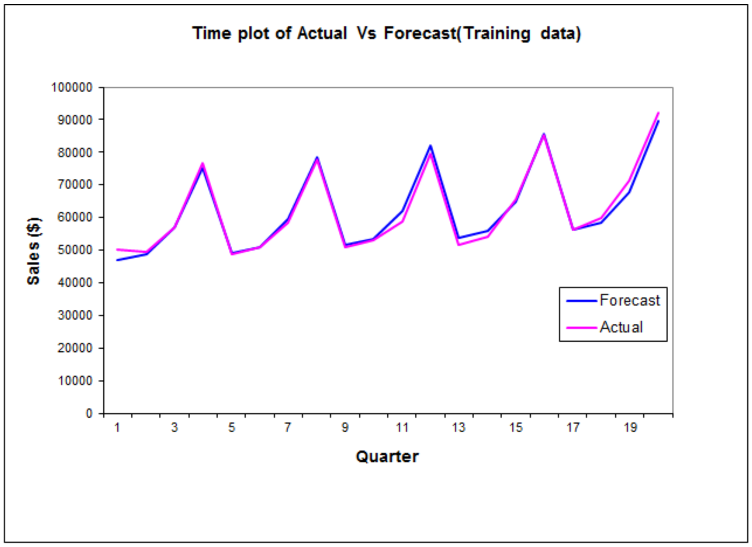
Forecast for the store sales of quarters 21 and 22 are:

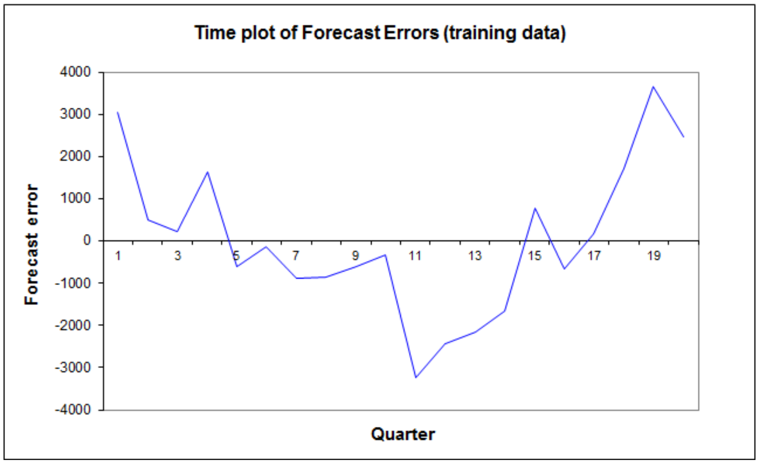
F_{21} = exp(10.9818) = $ 58,793.70

F_{22} = exp(11.0178) = $ 60,951.51

(e) The plots shown in Figure 5.38 describe the fit (top) and forecast errors (bottom) from this regression model.

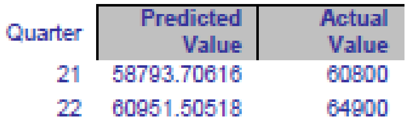
i. Recreate these plots.





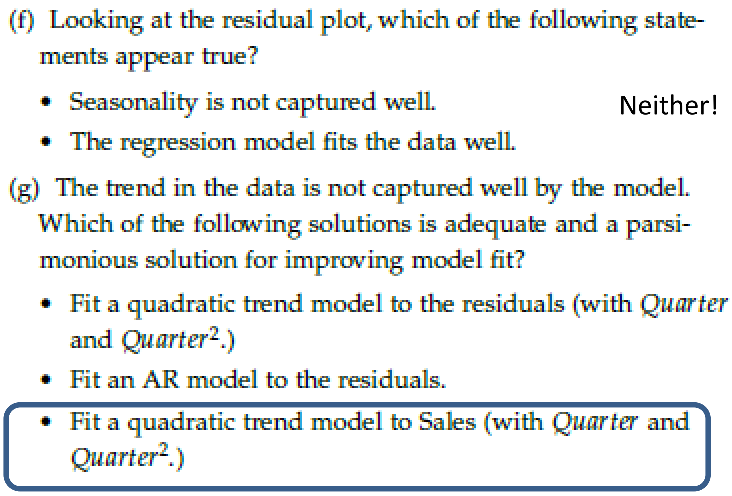
ii. Based on these plots, what can you say about your forecasts for quarters Q21 and Q22? Are they likely to overforecast, under-forecast, or be reasonably close to the real sales values?

Based on the U-shape of the residual plot at Q 20, you would expect Q21 and Q22 estimates to be too low (actual - forecast > 0). If we look at the actual data (below), it turns out that the predictions were indeed too low, indicating that the fitted exponential trend is inadequate.



Quarter	Predicted Value	Actual Value
21	58793.70616	60800
22	60951.50518	64900

From the above table we can see that the predicted forecast values for quarter 21 and 22 are below the actual values.



(f) Looking at the residual plot, which of the following statements appear true?

- Seasonality is not captured well.
- The regression model fits the data well.

Neither!

(g) The trend in the data is not captured well by the model. Which of the following solutions is adequate and a parsimonious solution for improving model fit?

- Fit a quadratic trend model to the residuals (with *Quarter* and *Quarter*^{2}.)
- Fit an AR model to the residuals.
- [✓ Fit a quadratic trend model to Sales (with *Quarter* and *Quarter*^{2}.)]