Homework #3 Questions

Directions: *Be sure to show your work and explain your answer for each question, even if the question seems to require only a Yes or No answer. Your homework solutions are to be entirely your own effort. You may not communicate with anyone about the homework, except for the TA and/or the instructor. You may use the Canvas postings, in-class discussion, any of the recommended textbooks, and computer software, if necessary, but no other resources. In writing up your solutions, you are required to support your answers with cut-and-pasted output; your answers must be clearly labeled and circled or highlighted. The grader will not search through unlabeled computer output to try to find your answers.*

10 points per question.

Note: *There is an "Answer Sheet Summary" at the end of these questions. After you have prepared your solutions, you must fill out the Answer Sheet Summary to provide a summary of your answers; attach the Summary to your solutions and submit both.*

The Excel workbook "Web analytics sales.xlsx" contains actual monthly sales in dollars for an anonymous web analytics firm. Use the data in this file to help answer the questions.

In the following questions, log means logarithm to the base e.

For Q1-Q6, the response variable is Y = logSales = log(Sales) instead of Sales.

Let MONTH = the month number, in chronological sequence from 1 – 69.

1. Run the model logSales = MONTH (this notation means the same as in the MODEL statement in SAS – namely, the response variable is on the left of the "=", the predictor(s) on the right.) Suppose that this model is valid. How would you interpret the meanings of the numerical values of the intercept and slope coefficients?

2. Again assuming the model logSales = MONTH is valid, provide the values of the two most commonly cited statistics to assess how well this model fits the data, and interpret the meanings of the values of those two statistics.

3. Diagnose the validity of the model logSales = MONTH by quantitatively testing the L,H,I,N regression specifications.

Consider the potential gain in explanatory power that may be achievable by adding seasonality predictors to the model logSales = MONTH: Create 0-1 indicator variables M1-M12 for each of the twelve months January-December.

4. Test whether the addition of all seasonal indicators M1-M12 to the model logSales = MONTH adds significant explanatory power.
[Hint: Run the original and augmented models. Get the change in the SS (Sum of Squares) explained by the regressions. Divide the change in SS by the number of added predictors. Divide that by the MSE (mean square error) of the augmented model. The result is the Wald statistic, which has an F distribution with k and m degrees of freedom, where k = number of added predictors and m = degrees of freedom for the MSE. Reject Ho: all added predictors have zero coefficients (hence no added explanatory power) vs Ha: at least one added predictor has nonzero coefficient (hence some added explanatory power) if the Wald statistic exceeds a critical point in the $F_{k,m}$ distribution. You may use the 0.05 critical point.]
[The intuition is that the added explanatory power is significant if the mean increase in explanatory power per added predictor is large in relation to the mean remaining available explanatory power per unused degree of freedom.]
*[**Caution**! Can you write the augmented* MODEL *statement as* logSales = MONTH M1-M12 ?*]*

Consider the potential gain in explanatory power that may be achievable by adding the past of the time series to the model logSales = MONTH: Create 12 lag variables: lagLSales1 = lag1(logSales); lagLSales2 = lag2(logSales); … , lagLSales12 = lag12(lagSales).

5. Test whether the addition of all lag predictors lagLSales1-lagLSales12 to the model logSales = MONTH adds significant explanatory power.
[Hint: See the hint for Q4.]
[Can you write the augmented MODEL *statement as* logSales = MONTH lagLSales1-lagLSales12 ?*]*

6. Determine the "best" subset of lagLSales1-lagLSales12 to add to the model logSales = MONTH:
Run a regular stepwise regression, in which a predictor is eligible for addition to the model if its coefficient in the next step of the model would be significant at the 0.05 level if it were added, and in which a predictor is eligible for removal from the model if its coefficient in the next step of the model fails to be significant at the 0.10 level.

a) What is the equation of your final model?
b) How much does the explained SS increase from the model logSales = MONTH by the addition of your "best" subset?
c) What proportion of the increase in explained SS that is achievable by the model logSales = MONTH lagLSales1-lagLSales12 *[see Q5]* is in fact achieved by your "best" subset?

[Hint: The SAS **MODEL** *statement option* **INCLUDE=1** *forces the first predictor in your* **MODEL** *statement to be included in your stepwise model, regardless of whether that predictor meets selection criteria.]*

To answer the remaining 4 questions, your general assignment is to develop a good model to forecast sales. See Q7-Q9 for the specific criteria you should meet. Provide your SAS program and output organized and labeled by question number to follow your program. Then, any question about how you got your result can be quickly resolved. The grader will record 0 points rather than spend significant time trying to figure out what you did. Be transparent.

Suggestions: You may wish to look at a timeplot of the data to see what structural features are present in the data – you may want to choose predictors that represent those features. You may wish to consider functional transformations of sales, trend variables, monthly indicators, lags, etc. If you transform sales, it is OK to meet your R-square, parsimony, and validity requirement [Q7-Q9] in terms of the transformed sales, instead of original scale sales. However, Q10 requires a forecast in original scale (so if you transform, transform your forecast back to original scale).

The following 4 questions award points for how good your model is.

6.5. You must begin this section by stating the equation of your model after you have developed it, including the estimated values of coefficients. *[See the Answer Sheet Summary.]* You earn no points by answering this question. However, you will lose up to 10 points if you do not answer this question satisfactorily.

Then your model will be scored according to the following criteria:

7. Explanatory power.
10 points if your R-square > 0.80; 9 points if 0.80 > R-square > 0.79; 8 points if 0.79 > R-square > 0.78; etc. but no negative points. You must provide the output and ***conspicuously label*** the output by question number next to the R-square that your model achieves. The TA will record 0 points rather than search through unlabeled output.

8. Parsimony.
10 points if your model has 3 or fewer predictors (excluding the intercept); 9 points if 4 predictors; 8 points if 5 predictors; etc. but no negative points. You must provide the output and ***conspicuously label*** the output by question number next to the predictors that your model has. The TA will record 0 points rather than search through unlabeled output.

9. Validity.

a) +1 for each of the three powers of RESET that your model passes at the 0.05 significance level.
b) +3 for passing White's homoscedasticity test at the 0.05 significance level.
c) +1 for each side (for positive and for negative autocorrelation – both for order 1 only) of the Durbin-Watson test that is passed at the 0.05 significance level.
d) +2 for passing the Shapiro-Wilk test at the 0.10 significance level.

You must provide the output and ***conspicuously*** label the output by question number next to the indicated test results. The TA will record 0 points rather than search through unlabeled output.

10. Forecast.
I have held out the actual sales figure for October, 2008. Use your model to forecast the value of sales for October, 2008. 10 points for forecasting actual sales for October, 2008 to within ±$30,000. 9 points for missing actual sales by more than $30,000 but less than $40,000; 8 points for missing actual sales by more than $40,000 but less than $50,000; etc. but no negative points.

Time Series
Homework #3
Answer Sheet Summary

NAME: _____

Q1. Intercept = _____ Interpretation: _____ <*add the space you need*>

Slope = _____ Interpretation: _____<*add the space you need*>

Q2. Statistic1 <name and value> = _____ Interpretation: _____

Statistic2 <name and value> = _____ Interpretation: _____

Q3. Test of L: _____

Test of H: _____

Test of I: _____

Test of N: _____

Q4. Change in SS (Sum of squares) = _____

Numerator of Wald statistic = _____

Denominator of Wald statistic = _____

Value of Wald statistic = _____

Critical point for Wald test = _____

Decision: _____

Q5. Change in SS (Sum of squares) = _____

Numerator of Wald statistic = _____

Denominator of Wald statistic = _____

Value of Wald statistic = _____

Critical point for Wald test = _____

Decision: _____

Q6. (a) Equation: _____

(b) Increase in SS = _____

(c) Proportion = _____

Q6.5 Equation: _____

_____

Q7. R-square = (4 digits) _____

Q8. Number of predictors = _____

Q9.

(a) RESET

Power	Statistic	p-value	p < 0.05? (Yes/No)
2			
3			
4			

(b) White

d.f.	Statistic	p-value	p < 0.05? (Yes/No)

(c) Durbin-Watson

Order	DW	Pr < DW	p-value < 0.05? (Yes/No)	Pr > DW	p-value < 0.05? (Yes/No)
1					

(d) Shapiro-Wilk

Test	Statistic	p-value	p < 0.05? (Yes/No)
Shapiro-Wilk			

10. Forecast = _____