**STA380 Time Series Analytics | Homework on Model Building**

Submitted by: Ruchi Sharma [rs58898]

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.

*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.*

**Ques 01:** Run the model logSales ~ MONTH (this notation refers to the formula in an appropriate "lm" statement in R) Suppose that this model is valid. How would you interpret the meanings of the numerical values of the intercept and slope coefficients?

Intercept: 10.88
Slope: 0.037

In the log scale, we get the values as above. This Implies that a unit change in Month represents about 3% change in Sales. The intercept can be interpreted as the average value of logSales in absence of any predictor.

```
Call:
lm(formula = logSales ~ MONTH, data = WebSales)

Residuals:
     Min       1Q   Median       3Q      Max
-0.75907 -0.22180  0.05064  0.20387  0.67003

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.88873    0.08134  133.87   <2e-16 ***
MONTH        0.03742    0.00202   18.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3342 on 67 degrees of freedom
Multiple R-squared:  0.8367,    Adjusted R-squared:  0.8342
F-statistic: 343.2 on 1 and 67 DF,  p-value: < 2.2e-16
```

**Ques 02:** 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.

2 commonly cited statistics are R Squared: 83.67% and RSE: 0.3342 as seen in the regression summary above. The residual standard error tells the net error in residuals and the R squared explains the percentage of variability in the target explained by the predictors.

**Ques 03:** Diagnose the validity of the model logSales ~ MONTH by quantitatively testing the L,I,N,E regression specifications.

Test of L:

```
> library(lmtest)
> resettest(lm1,power=2:4,type="fitted")

        RESET test

data:  lm1
RESET = 12.459, df1 = 3, df2 = 64, p-value = 1.629e-06
```

Since p-value is less than 0.05, we reject null hypothesis. L is not satisfied.

Test of I:

```
> library(car)
> durbinWatsonTest(lm1,max.lag=4,method="normal",alternative="positive")
 lag Autocorrelation D-W Statistic p-value
   1       0.5812223     0.8175203   0.000
   2       0.4453346     1.0412416   0.000
   3       0.3458671     1.1883674   0.000
   4       0.1480322     1.5723054   0.106
 Alternative hypothesis: rho[lag] != 0
```

Since p-value is less than 0.05, we reject the null hypothesis. I is not satisfied.

Test of N:

```
> ks.test(lm1$residuals,"pnorm",0,0.3315651)

        Exact one-sample Kolmogorov-Smirnov test

data:  lm1$residuals
D = 0.088187, p-value = 0.6245
alternative hypothesis: two-sided
```

```
> shapiro.test(lm1$residuals)

        Shapiro-Wilk normality test

data:  lm1$residuals
W = 0.9745, p-value = 0.1699
```

Looking at the Shapiro Test, we get p-value more than 0.05, we fail to reject the null hypothesis. So N is satisfied.

Test of E:

```
> library(whitestrap)
> white_test(lm1)
White's test results

Null hypothesis: Homoskedasticity of the residuals
Alternative hypothesis: Heteroskedasticity of the residuals
Test Statistic: 1.6
P-value: 0.449036
```

Since p-value is more than 0.05, we fail to reject the null hypothesis. So E is satisfied.

*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 corresponding twelve months January-December.*

**Ques 04:** Test whether the addition of all seasonal indicators M1-M12 to the model logSales ~ MONTH adds significant explanatory power.

[Caution! Can you write the augmented formula as logSales ~ MONTH M1-M12 ?] [Hint: The R command "lrtest(lmObject1, lmObject2)", where "lmObject1" and "lmObject2" are the R objects created by two "lm" commands, performs a likelihood ratio test of the difference in explanatory power of two models. You may use the 0.05 critical point.]

```
Call:
lm(formula = logSales ~ MONTH + Jan + Feb + Mar + Apr + May +
    Jun + Jul + Aug + Sep + Oct + Nov, data = WebSales)

Residuals:
     Min       1Q   Median       3Q      Max
-0.84954 -0.14344  0.05208  0.21602  0.53898

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.148270   0.162396  68.649   <2e-16 ***
MONTH        0.037764   0.001985  19.029   <2e-16 ***
Jan         -0.163669   0.197712  -0.828   0.4113
Feb         -0.377153   0.197623  -1.908   0.0615 .
Mar         -0.335520   0.197553  -1.698   0.0950 .
Apr         -0.046938   0.197503  -0.238   0.8130
May         -0.353596   0.197473  -1.791   0.0788 .
Jun         -0.362308   0.197463  -1.835   0.0718 .
Jul         -0.329113   0.197473  -1.667   0.1012
Aug         -0.473943   0.197503  -2.400   0.0198 *
Sep         -0.426237   0.197553  -2.158   0.0353 *
Oct         -0.077322   0.206282  -0.375   0.7092
Nov         -0.229302   0.206253  -1.112   0.2710
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3261 on 56 degrees of freedom
Multiple R-squared:   0.87,     Adjusted R-squared:  0.8421
F-statistic: 31.23 on 12 and 56 DF,  p-value: < 2.2e-16
```

Since the R-Squared increased from 83% to 87%, we can say that the new model enhances the explanatory power as compared to the old model.

```
> lrtest(lm1, lm2)
Likelihood ratio test

Model 1: logSales ~ MONTH
Model 2: logSales ~ MONTH + Jan + Feb + Mar + Apr + May + Jun + Jul +
    Aug + Sep + Oct + Nov
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   3 -21.257
2  14 -13.387 11 15.741      0.151
```

From the test above, p value is not less 0.05, which implies both models fit the data equally value. Since additional parameters don’t add much value to the model we can use the simpler model instead.

*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).*

**Ques 05.** 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” ?]

```
> lm3_2 = lm(logSales ~ MONTH + lagLSales1 + lagLSales2 + lagLSales3 + lagLSales4 +
+            lagLSales5 + lagLSales6 + lagLSales7 + lagLSales8 + lagLSales9 + lagLSales10 +
+            lagLSales11 + lagLSales12, data=WebSales2)
> summary(lm3_2)

Call:
lm(formula = logSales ~ MONTH + lagLSales1 + lagLSales2 + lagLSales3 +
    lagLSales4 + lagLSales5 + lagLSales6 + lagLSales7 + lagLSales8 +
    lagLSales9 + lagLSales10 + lagLSales11 + lagLSales12, data = WebSales2)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5498 -0.1019 -0.0112  0.1550  0.4705

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.398995   2.685362   2.011   0.0507 .
MONTH        0.008959   0.008865   1.011   0.3179
lagLSales1   0.273840   0.149531   1.831   0.0740 .
lagLSales2   0.120533   0.152311   0.791   0.4331
lagLSales3   0.076546   0.152622   0.502   0.6186
lagLSales4  -0.135269   0.153021  -0.884   0.3816
lagLSales5  -0.256346   0.153345  -1.672   0.1018
lagLSales6   0.228076   0.166733   1.368   0.1784
lagLSales7  -0.075400   0.164697  -0.458   0.6494
lagLSales8  -0.068737   0.163538  -0.420   0.6763
lagLSales9  -0.006212   0.166731  -0.037   0.9705
lagLSales10  0.110462   0.162134   0.681   0.4993
lagLSales11  0.126038   0.159761   0.789   0.4345
lagLSales12  0.154776   0.154209   1.004   0.3211
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2583 on 43 degrees of freedom
Multiple R-squared:  0.8383,    Adjusted R-squared:  0.7894
F-statistic: 17.15 on 13 and 43 DF,  p-value: 6.122e-13
```

The R Squared does not improve much.

```
> lrtest(lm1_2, lm3_2)
Likelihood ratio test

Model 1: logSales ~ MONTH
Model 2: logSales ~ MONTH + lagLSales1 + lagLSales2 + lagLSales3 + lagLSales4 +
    lagLSales5 + lagLSales6 + lagLSales7 + lagLSales8 + lagLSales9 +
    lagLSales10 + lagLSales11 + lagLSales12
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   3 -8.2417
2  15  4.3047 12 25.093    0.01439 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the test above, p value is not less 0.05, which implies both models fit the data equally value. Since additional parameters don’t add much value to the model we can use the simpler model instead.

**Ques 06.** See if you can find a "best" subset of lagLSales1-lagLSales12 to add to the model logSales ~ MONTH: Try not to include predictors that are not significant at the 0.05 level. But try to maximize your R-square. In other words, balance explanatory power with a parsimonious predictor set.

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 potentially achievable by the model logSales ~ MONTH lagLSales1-lagLSales12 [see Q5] is in fact achieved by your "best" subset?

```
> summary(step_model)

Call:
lm(formula = logSales ~ MONTH + lagLSales12 + lagLSales1 + lagLSales5,
    data = WebSales2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.63130 -0.13703  0.01412  0.14211  0.46545

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.191505   2.022173   3.556 0.000812 ***
MONTH        0.014116   0.006917   2.041 0.046373 *
lagLSales12  0.309949   0.116903   2.651 0.010601 *
lagLSales1   0.280642   0.121690   2.306 0.025118 *
lagLSales5  -0.204244   0.105968  -1.927 0.059400 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2461 on 52 degrees of freedom
Multiple R-squared:  0.8225,    Adjusted R-squared:  0.8088
F-statistic: 60.23 on 4 and 52 DF,  p-value: < 2.2e-16
```

From the regression summary above, it can be noted that MONTH, lagLSales12, and lagLSales1 are best predictors.

```
> lm3_final = lm(logSales ~ MONTH + lagLSales1 + lagLSales12, data=WebSales2)
> summary(lm3_final)

Call:
lm(formula = logSales ~ MONTH + lagLSales1 + lagLSales12, data = WebSales2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.71604 -0.08929  0.01487  0.11049  0.46792

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.717160   1.601923   2.945  0.00479 **
MONTH       0.006883   0.005958   1.155  0.25317
lagLSales1  0.308240   0.123901   2.488  0.01603 *
lagLSales12 0.302902   0.119801   2.528  0.01447 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2524 on 53 degrees of freedom
Multiple R-squared:  0.8098,    Adjusted R-squared:  0.799
F-statistic: 75.21 on 3 and 53 DF,  p-value: < 2.2e-16
```

a) The equation of the final model is logSales = 4.717 + (0.006)*MONTH + (0.302)*lagLSales12 + (0.308)*lagLSales1

b) & c) The output below explains the SS increase for best model

```
> ss = sum(WebSales2$logSales^2)-sum(bestmodel$residuals^2)
> ss2 = sum(WebSales2$logSales^2)-sum(nullmodel$residuals^2)
> ss3 = sum(WebSales2$logSales^2)-sum(fullmodel$residuals^2)
> ss
[1] 8895.371
> ss2
[1] 8894.291
> ss3
[1] 8895.878
> ss-ss2
[1] 1.080853
> ss-ss3
[1] -0.5061383
> ss/ss2
[1] 1.000122
> ss/ss3
[1] 0.9999431
>
> (ss-ss2)/(ss3-ss2) # proportion of increased model
[1] 0.6810705
```

*To answer the remaining 4 questions, your general assignment is to develop a good model to forecast sales – not necessarily limited to the predictors already considered in the preceding questions in this assignment. See Q7-Q9 for the specific criteria you should meet. Provide your 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.*

**Ques 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. Then your model will be scored according to the following rubrics:**

```
final_model2 <- lm(logSales~MONTH+I(MONTH^2)+lagLSales1, data = WebSales3)
```

**Ques 07: 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.

```
> summary(final_model2)

Call:
lm(formula = logSales ~ MONTH + I(MONTH^2) + lagLSales1, data = WebSales3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.81147 -0.13532  0.01111  0.16451  0.56724

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.5434779  1.2034895   5.437 9.05e-07 ***
MONTH        0.0479474  0.0111279   4.309 5.77e-05 ***
I(MONTH^2)  -0.0003521  0.0001109  -3.176  0.00230 **
lagLSales1   0.3738332  0.1158234   3.228  0.00197 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2558 on 64 degrees of freedom
Multiple R-squared:  0.9035,    Adjusted R-squared:  0.899
F-statistic: 199.8 on 3 and 64 DF,  p-value: < 2.2e-16
```

R Squared = 90.35%

**Ques 08: 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.

Yes, the model has 3 predictors.

**Ques 09: Validity:**

a) +1 for each of the three powers 2,3,4 of RESET that your model passes at the 0.05 significance level.
b) +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.
c) +2 for passing the Shapiro-Wilk test at the 0.10 significance level.
d) +3 for passing White's homoscedasticity test (lag 1) at the 0.05 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.

```
> resettest(final_model2,power=2,type="fitted")

        RESET test

data:  final_model2
RESET = 1.888, df1 = 1, df2 = 63, p-value = 0.1743

> resettest(final_model2,power=3,type="fitted")

        RESET test

data:  final_model2
RESET = 2.1914, df1 = 1, df2 = 63, p-value = 0.1438

> resettest(final_model2,power=4,type="fitted")

        RESET test

data:  final_model2
RESET = 2.1556, df1 = 1, df2 = 63, p-value = 0.147
```

```
> durbinWatsonTest(final_model2)
 lag Autocorrelation D-W Statistic p-value
   1     -0.01930332       2.02586   0.842
 Alternative hypothesis: rho != 0
```

```
> shapiro.test(final_model2$residuals)

        Shapiro-Wilk normality test

data:  final_model2$residuals
W = 0.98553, p-value = 0.6188
```

```
> white_test(final_model2)
White's test results

Null hypothesis: Homoskedasticity of the residuals
Alternative hypothesis: Heteroskedasticity of the residuals
Test Statistic: 1.33
P-value: 0.515258
```

Looking at the above tests, we can see that L, I, N, E are satisfied.

**Ques 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.

```
  Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
1       13.10305 12.74858 13.45753 12.55619 13.64992
> exp(c(f1$mean,f1$lower,f1$upper))
[1] 490438.3 344062.7 283846.7 699087.1 847393.2
```