MTH 325, Final Exam, At Home, Spring 2023 Name KEY

**Instructions**: This exam is in two parts: Part I is to be completed partly at home using the materials posted in the course for the at-home portion and you will answer questions about that work during the in-class portion of the exam; Part II is to be completed entirely in class. You may not use cell phones, and you may only access internet resources you are specifically directed to use.

At home, prepare for questions in Part I using R. Open the data file entitled **325final_data.xlsx** posted in Blackboard. Complete the calculations noted below. You will be asked for additional analysis and interpretation of this data in the in-class portion of the test. Print out the results of your analysis and code, and bring the pages with you to the exam. You will submit all this work along with the in-class exam.

Use the data on motels to complete the following tasks. Sheet 1 has data on existing motels and information about their location, relationship to competitors, specific amenities and operating margin. Sheet 2 has data on possible locations for building a new motel.

1. Import the data in the file into R and remove the Motel and Site columns (they are not a variable). Separate your data into two dataframes. One for the existing locations and one for the possible building sites (to be used later).

2. Convert the Indoor Pool Variable to a binary dummy variable.

3. Create a correlation table of the variables. Make a correlation plot (type is of your choice), or a pairplot.

4. Which variable has the strongest relationship with whether or not the motel has an indoor pool? Create a logistic regression model to predict the indoor pool variable. Create a graph of the model in ggplot. Create appropriate graphs for diagnostic testing of assumptions, and identify potential outliers. Create a confusion matrix for your model.

5. Create a multiple variable model of operating margins using all remaining available variables. Use appropriate automated selection techniques. Compare the result to manual backward selection. In your backward selection, stop only when all the coefficients are significant at the 0.05 level.

6. Construct diagnostic plots for your machine selected model and your manually selected model (these may be the same). Identify any potential problems with model assumptions, outliers and influential points.

7. Using the data on Sheet 2, predict the operating margins for all locations if a) the motel does not have an indoor pool, b) if it does.

To complete the calculations below, use the time series Seatbelts.

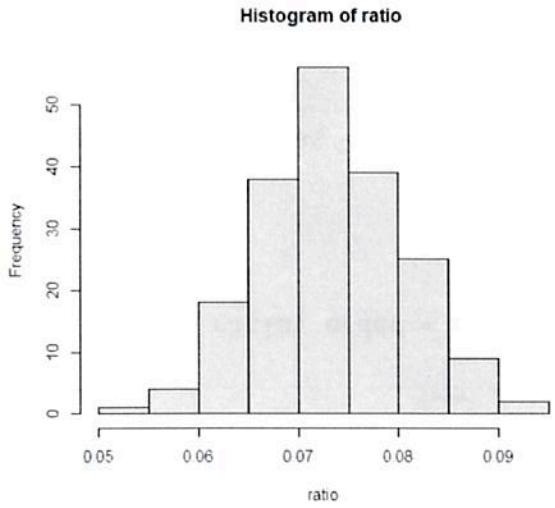
8. Create a new column in the dataset (or a separate vector) that represents the ratio of drivers killed to total drivers. Construct appropriate one-variable numerical plots to describe the overall data set.

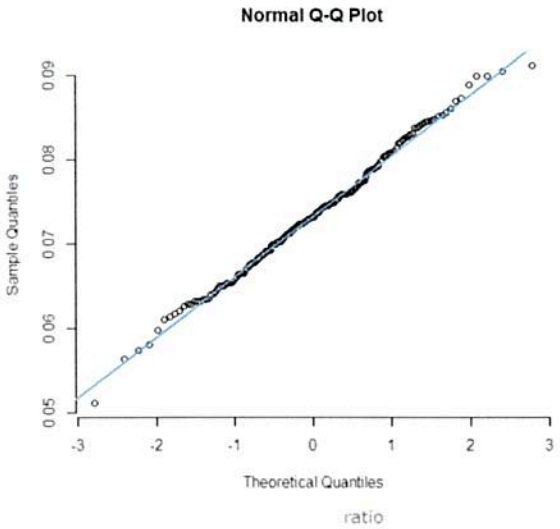
9. Create a plot of the new time series. Perform seasonal decomposition and plot the resulting graph.

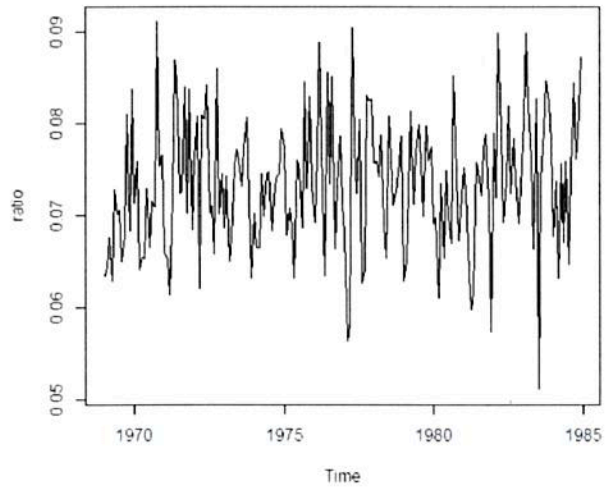
10. Create an ACF graph for the time series.

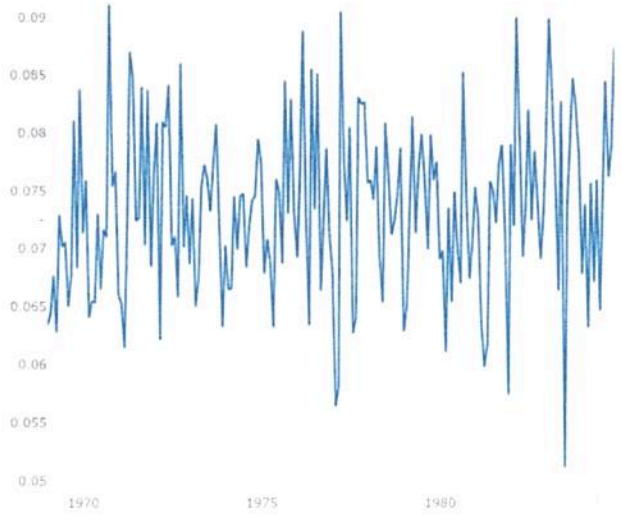
11. Construct an ARIMA model. Plot the model against the original time series.

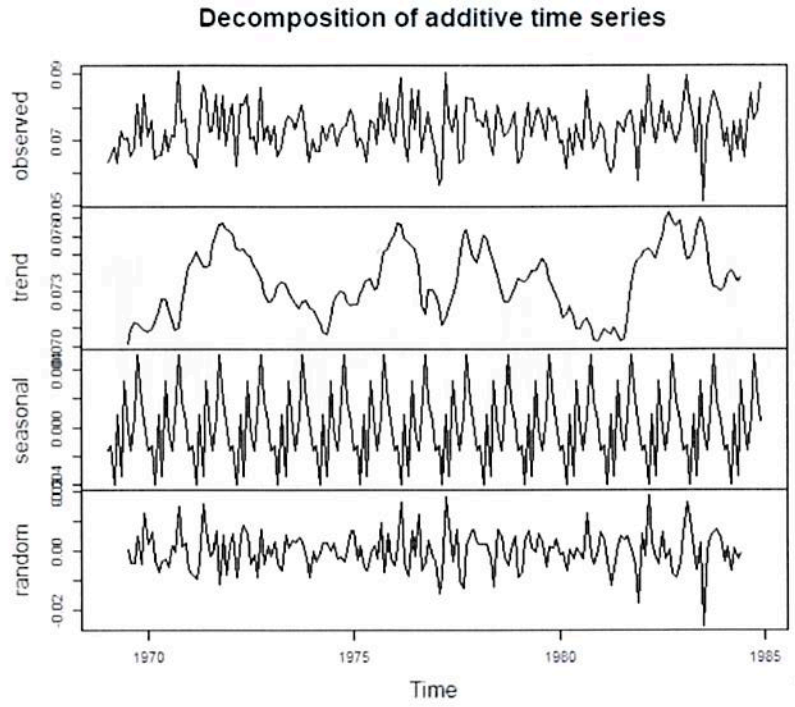
325 Final Exam Spring 2023 at-home analysis

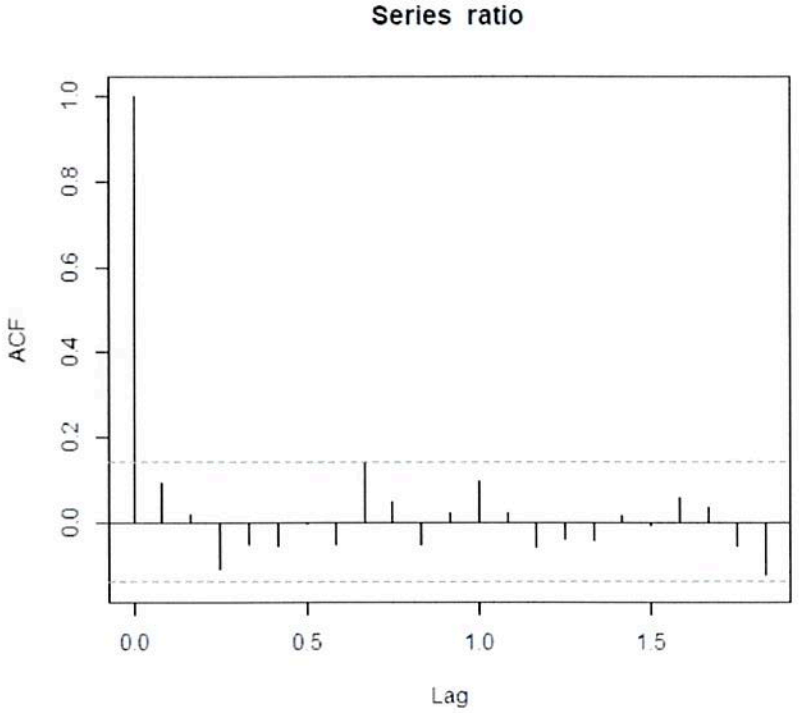


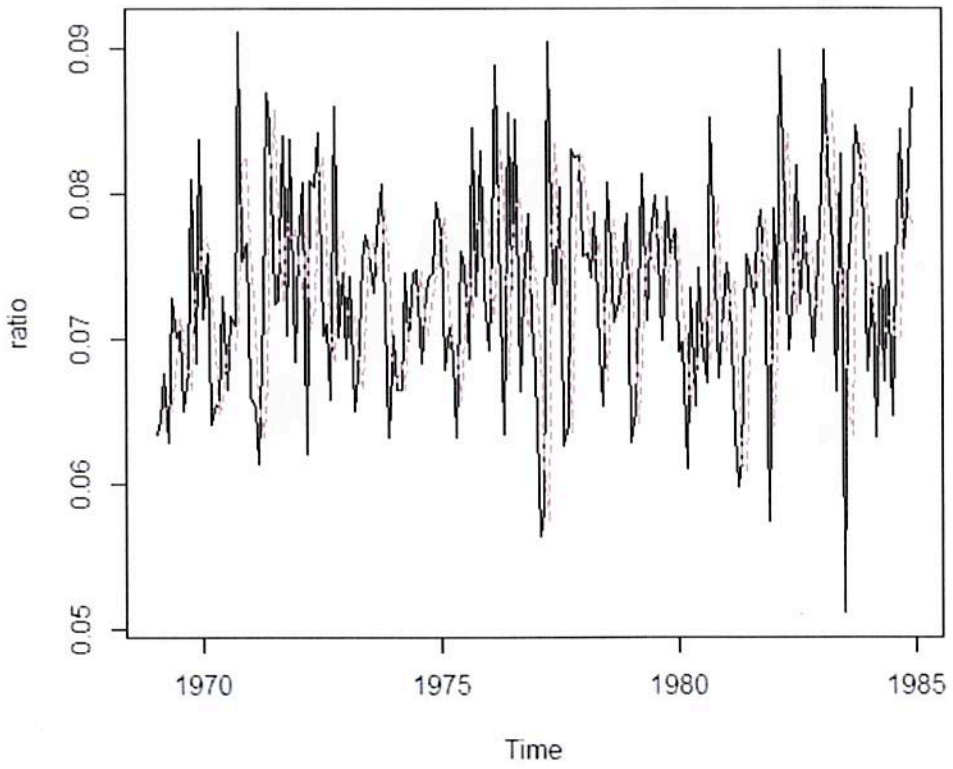






```
Call:rimax = ratio, order = c(1, 1, 0))

Coefficients:
          ar1
      -0.4582
s.e.   0.0642

sigma^2 estimated as 7.383e-05:  log likelihood = 637.42,  aic = -1270.85
```

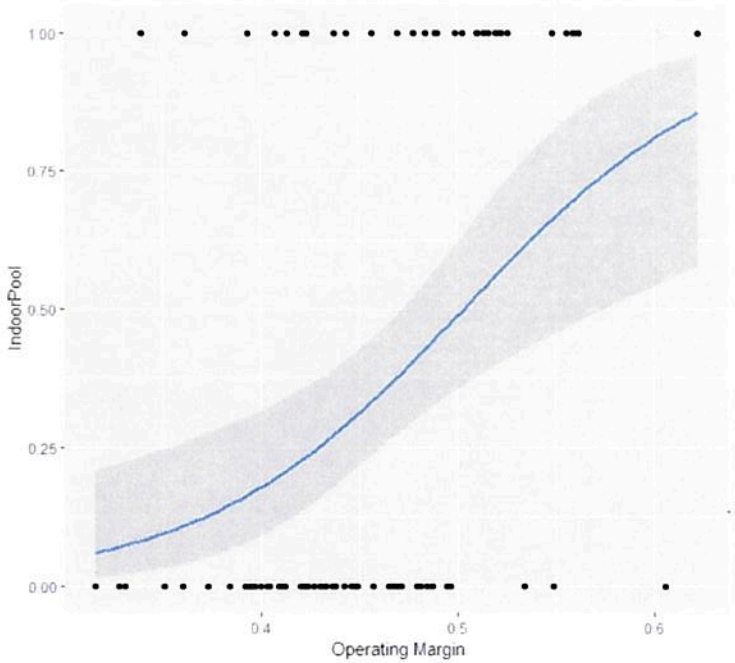



BIC = -1264.344

| | Competitor rooms | Distance to competitor | Office space |
|---|---|---|---|
| Competitor rooms | 1.00000000 | -0.10584448 | -0.08089589 |
| Distance to competitor | -0.10584448 | 1.00000000 | -0.08350897 |
| Office space | -0.08089589 | -0.08350897 | 1.00000000 |
| Median income | -0.04324780 | -0.07847425 | -0.15526820 |
| Distance to downtown | -0.18020406 | 0.14943471 | -0.03189481 |
| Operating Margin | -0.57232036 | -0.27515942 | 0.61802716 |
| IndoorPool | -0.02709271 | -0.17450757 | 0.23352704 |

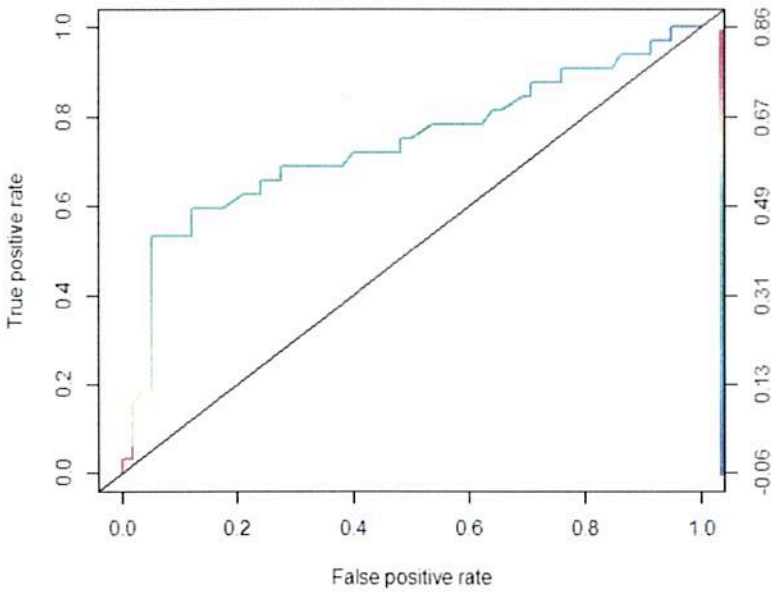
| | Median income | Distance to downtown | Operating Margin |
|---|---|---|---|
| Competitor rooms | -0.04324780 | -0.18020406 | -0.57232036 |
| Distance to competitor | -0.07847425 | 0.14943471 | -0.27515942 |
| Office space | -0.15526820 | -0.03189481 | 0.61802716 |
| Median income | 1.00000000 | -0.15535453 | -0.09630002 |
| Distance to downtown | -0.15535453 | 1.00000000 | 0.14943946 |
| Operating Margin | -0.09630002 | 0.14943946 | 1.00000000 |
| IndoorPool | -0.14059902 | 0.03553107 | 0.36923156 |

| | IndoorPool |
|---|---|
| Competitor rooms | -0.02709271 |
| Distance to competitor | -0.17450757 |
| Office space | 0.23352704 |
| Median income | -0.14059902 |
| Distance to downtown | 0.03553107 |
| Operating Margin | 0.36923156 |
| IndoorPool | 1.00000000 |

```
Call:
glm(formula = IndoorPool ~ `Operating Margin`, data = data1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7930  -0.3503  -0.1994   0.4584   0.9868

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.9834     0.3624  -2.714 0.008002 **
`Operating Margin`  2.9313     0.7865   3.727 0.000342 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.2023949)

    Null deviance: 20.622  on 89  degrees of freedom
Residual deviance: 17.811  on 88  degrees of freedom
AIC: 115.61

Number of Fisher Scoring iterations: 2


Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 55  3
         1 17 15

               Accuracy : 0.7778
                 95% CI : (0.6779, 0.8587)
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 0.74971

                  Kappa : 0.4624

 Mcnemar's Test P-Value : 0.00365

            Sensitivity : 0.7639
            Specificity : 0.8333
         Pos Pred Value : 0.9483
         Neg Pred Value : 0.4687
             Prevalence : 0.8000
         Detection Rate : 0.6111
   Detection Prevalence : 0.6444
```

```
      Balanced Accuracy : 0.7986

       'Positive' Class : 0
Auc = 0.7349138
```




Backward Selection:

```
Call:
lm(formula = `Operating Margin` ~ ., data = data1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.055718 -0.021494 -0.002689  0.019477  0.064003

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               5.930e-01  4.087e-02  14.509  < 2e-16 ***
`Competitor rooms`       -6.841e-05  6.707e-06 -10.200 2.63e-16 ***
`Distance to competitor` -1.878e-02  3.640e-03  -5.160 1.66e-06 ***
`Office space`            1.851e-04  1.947e-05   9.504 6.42e-15 ***
```

```
`Median income`          -1.753e-04  4.442e-04  -0.395  0.69420
`Distance to downtown`    1.886e-03  1.007e-03   1.874  0.06448 .
IndoorPool                2.284e-02  6.789e-03   3.364  0.00116 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02931 on 83 degrees of freedom
Multiple R-squared:  0.7821,   Adjusted R-squared:  0.7664
F-statistic: 49.66 on 6 and 83 DF,  p-value: < 2.2e-16


Call:
lm(formula = `Operating Margin` ~ `Competitor rooms` + `Distance to competitor` +
    `Office space` + `Distance to downtown` + IndoorPool, data = data1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.054847 -0.020963 -0.001815  0.020932  0.065592

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              5.808e-01  2.659e-02  21.840  < 2e-16 ***
`Competitor rooms`      -6.815e-05  6.640e-06 -10.263  < 2e-16 ***
`Distance to competitor` -1.864e-02  3.604e-03  -5.172 1.55e-06 ***
`Office space`           1.862e-04  1.916e-05   9.717 2.15e-15 ***
`Distance to downtown`   1.949e-03  9.890e-04   1.971  0.05201 .
IndoorPool               2.316e-02  6.707e-03   3.453  0.00087 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02916 on 84 degrees of freedom
Multiple R-squared:  0.7817,   Adjusted R-squared:  0.7687
F-statistic: 60.17 on 5 and 84 DF,  p-value: < 2.2e-16
```

Final backward selection model:

```
Call:
lm(formula = `Operating Margin` ~ `Competitor rooms` + `Distance to competitor` +
    `Office space` + IndoorPool, data = data1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.059943 -0.024734 -0.002763  0.020760  0.066817

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              6.013e-01  2.490e-02  24.150  < 2e-16 ***
```

```
`Competitor rooms`       -7.035e-05  6.656e-06 -10.570  < 2e-16 ***
`Distance to competitor` -1.766e-02  3.629e-03  -4.865 5.21e-06 ***
`Office space`            1.843e-04  1.946e-05   9.472 5.98e-15 ***
IndoorPool                2.403e-02  6.805e-03   3.531 0.000671 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02965 on 85 degrees of freedom
Multiple R-squared:  0.7716,   Adjusted R-squared:  0.7609
F-statistic:  71.8 on 4 and 85 DF,  p-value: < 2.2e-16
```

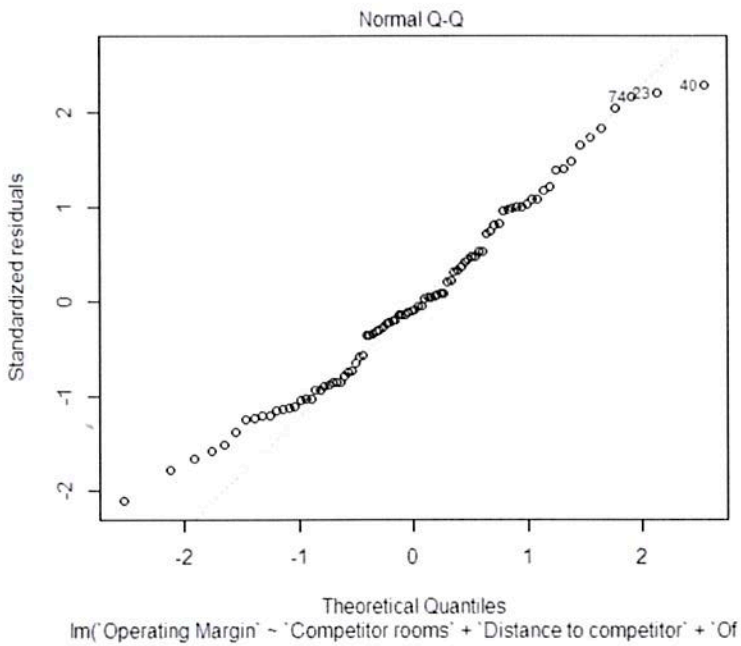
Best subsets

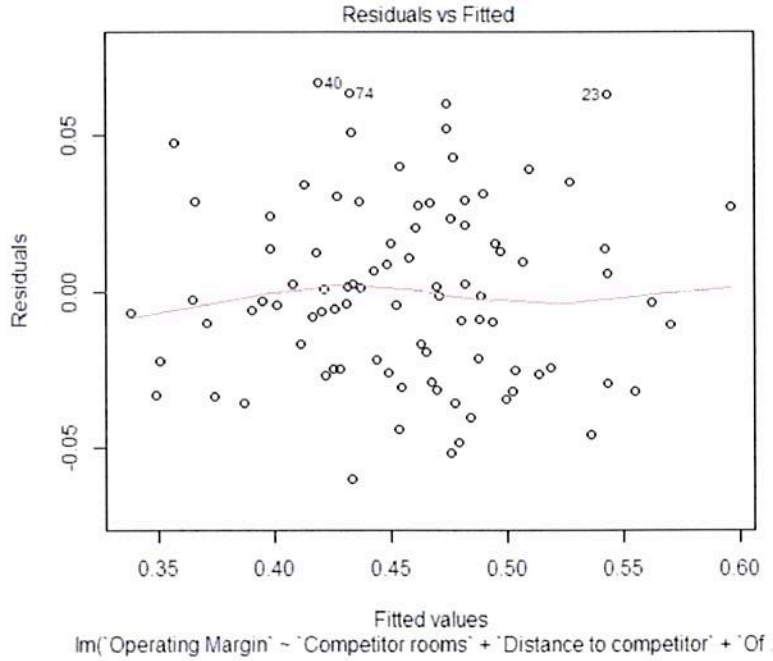
```
Subset selection object
Call: regsubsets.formula(`Operating Margin` ~ ., data = data1, nvmax = 6)
6 Variables  (and intercept)
                         Forced in Forced out
`Competitor rooms`           FALSE      FALSE
`Distance to competitor`     FALSE      FALSE
`Office space`               FALSE      FALSE
`Median income`              FALSE      FALSE
`Distance to downtown`       FALSE      FALSE
IndoorPool                   FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: exhaustive
         `Competitor rooms` `Distance to competitor` `Office space` `Median income`
1  ( 1 ) " "                " "                      "*"            " "
2  ( 1 ) "*"                " "                      "*"            " "
3  ( 1 ) "*"                "*"                      "*"            " "
4  ( 1 ) "*"                "*"                      "*"            " "
5  ( 1 ) "*"                "*"                      "*"            " "
6  ( 1 ) "*"                "*"                      "*"            "*"
         `Distance to downtown` IndoorPool
1  ( 1 ) " "                    " "
2  ( 1 ) " "                    " "
3  ( 1 ) " "                    " "
4  ( 1 ) " "                    "*"
5  ( 1 ) "*"                    "*"
6  ( 1 ) "*"                    "*"
```

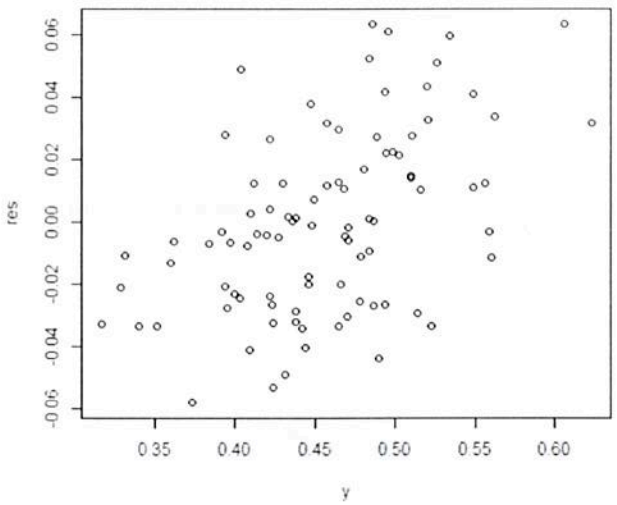
LASSO model:

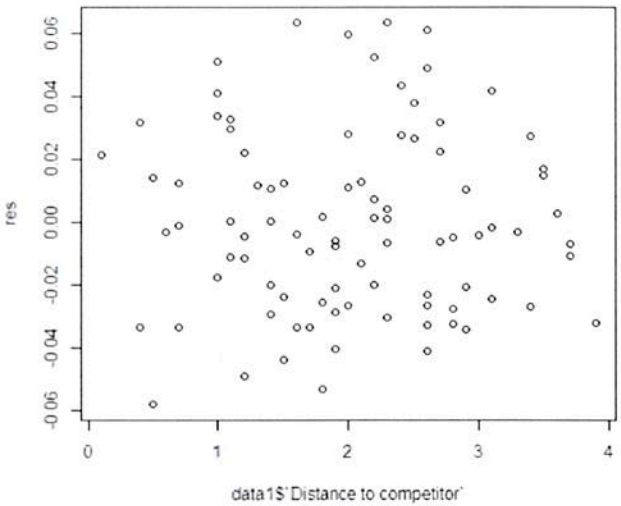
```
7 x 1 sparse Matrix of class "dgCMatrix"
                              s0
(Intercept)         6.189800e-01
(Intercept)                    .
`Competitor rooms` -7.029538e-05
```

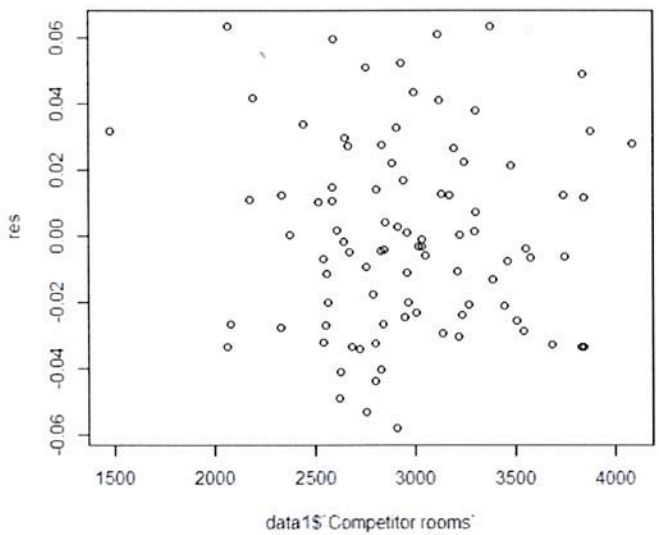
```
`Distance to competitor` -1.775691e-02
`Office space`            1.818834e-04
`Median income`          -2.844955e-04
IndoorPool                2.323760e-02

      RMSE        s0 = R^2
1 0.0287342 0.7729143
```

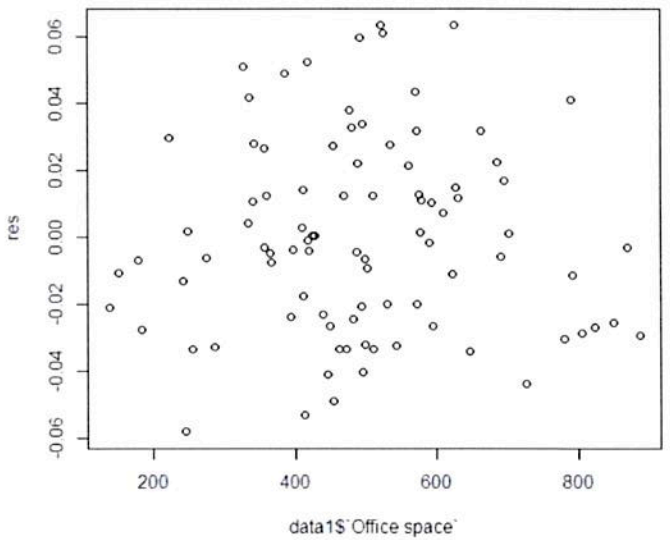


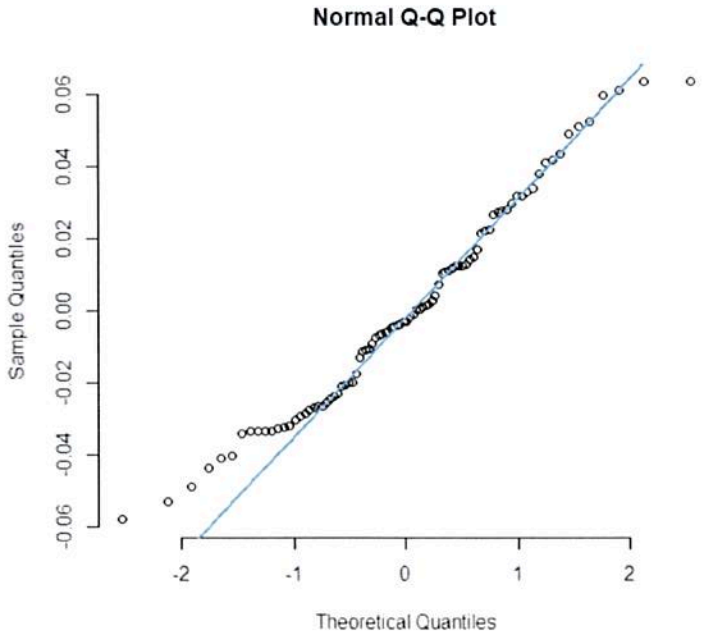




res
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
0.35
0.40
0.45
0.50
0.55
0.60
y


res
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
0
1
2
3
4
data1$`Distance to competitor`


res
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
1500
2000
2500
3000
3500
4000
data1$`Competitor rooms`


res
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
200
400
600
800
data1$`Office space`




Backward selection/best subsets predictions: no pool

```
        1         2         3         4         5
0.4742227 0.3703893 0.4694931 0.4954500 0.4307804
```

With pool

```
        1         2         3         4         5
0.4982514 0.3944180 0.4935217 0.5194786 0.4548091
```

MTH 325, Final Exam, Spring 2023 Name KEY

Instructions: Answer each question thoroughly. For questions in Part 1, use the work you did at home to answer the questions. Be sure to answer each part of each question. In Part 2, report exact answers unless directed to round.

Part I:
Use the work you did at home to answer these questions about motels.

1. Based on your correlation table or correlation plot, identify the variable that has the highest negative correlation with Operating Margin. What is the (approximate) correlation value?

Competitor rooms ≈ −0.5 or so
−0.572...

2. What is the best variable we can use to predict whether or not the motel has an indoor pool?

Operating margin

3. Write the equation of your logistic model below. You can write it in the form $\ln\left(\frac{p}{1-p}\right) = linear\ model.$

$$\ln\left(\frac{p}{1-p}\right) = -0.9834 + 2.9313(\text{OperatingMargin})$$

4. Write your confusion matrix here, and state the accuracy of your model.

| Pred \ True | 0 | 1 |
|---|---|---|
| 0 | 55 | 3 |
| 1 | 17 | 15 |

77.78%

5. Write the equation you obtained from your backward selection process for predicting operating expenses. Be sure to clearly indicate what each variable in the equation represents.

Operating Margin = 0.6013 − $7.035\times10^{-5}$(competitor rooms) − $1.766\times10^{-2}$(dist. to competitor) + $1.843\times10^{-4}$(office space) + $2.403\times10^{-2}$(IndoorPool)

6. What is the (approximate) area under your ROC curve?

73%

7. Describe how your other model selection methods differed (or were similar to) the results obtained from the backward selection process. answers may vary.

best subset regression got the same set of variables

LASSO kept all variables, but coeff of similar variables were also similar the $R^2$ was slightly higher

8. What percentage of the variability in Operating Margin can be explained by the relationship to the other model variables?

77.16% (backward selection)

77.29 for my LASSO model

9. ~~Answer this question and the remaining questions in Part 1 using the backward selection~~ model ~~you found by hand. Write the equation of your model that describes your multiple regression model~~

do your diagnostic plots suggest any outliers or model problems.

residuals plots look fine. QQplot suggests there may be 1 or 2 outliers

10. How do your predictions for the 5 possible locations differ? Should they have an indoor pool or not? What is the best location based on the predicted operating margins? (Give the location number.)

W/ no pool range from 0.37 to 0.495

W/ pool range from 0.39 to 0.519

W/ pool are all higher

worst location is #2, best is #4

11. Interpret the meaning of the Indoor Pool coefficient in the context of the problem.

the presence of an indoor pool improves operating margin by 2.4%.

12. ~~Test your model assumptions using your residual plots and other diagnostic plots. Do~~ they ~~appear to be approximately satisfied? Identify any potential outliers.~~ if you could only model operating margin w/ a single variable, what would it be?

office space

Use the work you did at home to answer these questions about the time series model.

13. Does the model appear approximately stationary or does there appear to be a trend? Consider any boxplots or histograms here, as well as any time series plots or decompositions you may have done.

yes, other than the mean not being zero
the histogram & qqplot suggest normality
decomp suggests trend is a random walk.

14. Based on your ACF graph, how many lags should be included in your time series model? Why?

one or none. other than 0 lags, all correlations are below significance

15. What settings did you use for your ARIMA model? Why? What diagnostics did you use to select these settings?

answers may vary

(1,1,0) was best

based on lowest |AIC|

16. Write the equation of your final time series model.

$$t_n = -0.4582\, t_{n-1}$$

17. What is the AIC of your final model? How good does the model appear to fit?

$-1270.85$

pretty well

Part II:

18. Recall that $Cov(X,Y) = E(XY) - E(X)E(Y)$. For the probability density function $f(x,y) = \frac{25}{192}x^{3/2}y^{2/3}, y \in [0,1], x \in [0,4]$, find the covariance.

$$E(XY) = \frac{25}{192}\int_0^1\int_0^4 x^{5/2}y^{5/3}\,dx\,dy = \frac{25}{192}\left(\frac{2}{7}x^{7/2}\Big|_0^4\right)\left(\frac{3}{8}x^{8/3}\Big|_0^1\right) =$$

$$\frac{25}{192}\cdot\frac{2}{7}\cdot 128\left(\frac{3}{8}(1)\right) = \frac{25}{14}$$

$$E(X) = \frac{25}{192}\int_0^1\int_0^4 x^{5/2}y^{2/3}\,dx\,dy = \frac{25}{192}\left(\frac{2}{7}x^{7/2}\Big|_0^4\right)\left(\frac{3}{5}x^{5/3}\Big|_0^1\right) =$$

$$\frac{25}{192}\left(\frac{2}{7}\cdot 128\right)\left(\frac{3}{5}\cdot 1\right) = \frac{20}{7}$$

$$E(Y) = \frac{25}{192}\int_0^1\int_0^4 x^{3/2}y^{5/3}\,dx\,dy = \frac{25}{192}\left(\frac{2}{5}x^{5/2}\Big|_0^4\right)\left(\frac{3}{8}y^{8/3}\Big|_0^1\right) =$$

$$\frac{25}{192}\left(\frac{2}{5}\cdot 32\right)\left(\frac{3}{8}(1)\right) = \frac{5}{8}$$

$$E(XY) - E(X)E(Y) = \frac{25}{14} - \frac{20}{7}\cdot\frac{5}{8} = \frac{25}{14} - \frac{25}{14} = \boxed{0}$$

19. Consider the small data set {(2,1), (5,3), (8,7)}. Find the value of the regression coefficients for $y = \beta_0 + \beta_1 x$, using the normal equation $(A^T A)^{-1} A^T Y = B$. Write the coefficients you find in the equation.

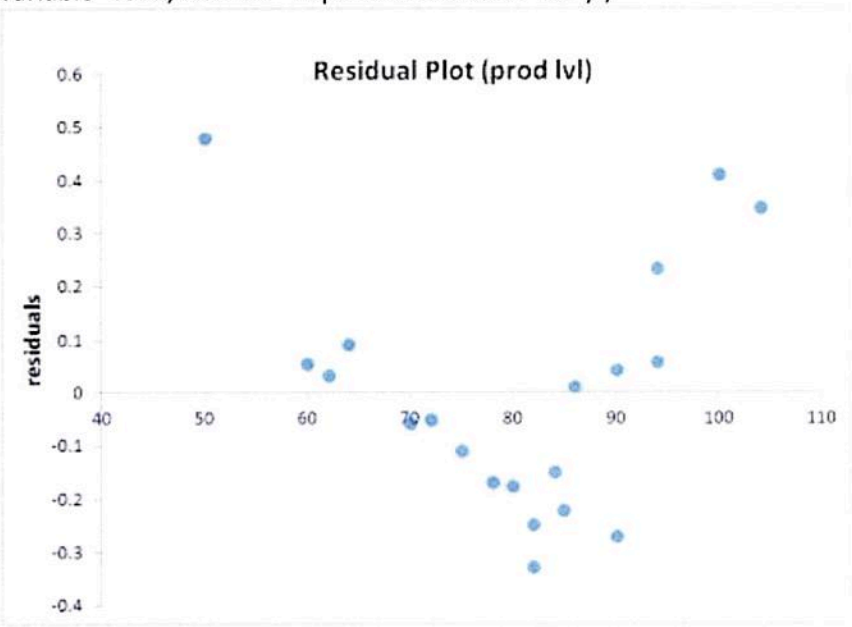
$\beta_0 + \beta_1(2) = 1$
$\beta_0 + \beta_1(5) = 3$
$\beta_0 + \beta_1(8) = 7$

$A = \begin{bmatrix} 1 & 2 \\ 1 & 5 \\ 1 & 8 \end{bmatrix}$ $B = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ $Y = \begin{bmatrix} 1 \\ 3 \\ 7 \end{bmatrix}$

$A^T = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 5 & 8 \end{bmatrix}$ $A^T A = \begin{bmatrix} 3 & 15 \\ 15 & 93 \end{bmatrix}$ $(A^T A)^{-1} = \begin{bmatrix} 31/18 & -5/18 \\ -5/18 & 1/18 \end{bmatrix}$

$(A^T A)^{-1} A^T Y = \begin{bmatrix} -4/3 \\ 1 \end{bmatrix}$ $\hat{y} = -4/3 + x$

20. Examine the residual plot below. Identify some potential issues with the linear model used to produce these residuals. (For example, without knowing the variables names, use "independent variable" for x, and the "dependent variable" for y.)




the independent variables appear to be related non-linearly to the dependent variable

21. Describe clustering (in machine learning), and give an example of a machine learning algorithm that implements this learning method.

this is an unsupervised learning method that groups data into related groups.

KNN, Kmeans, LDA

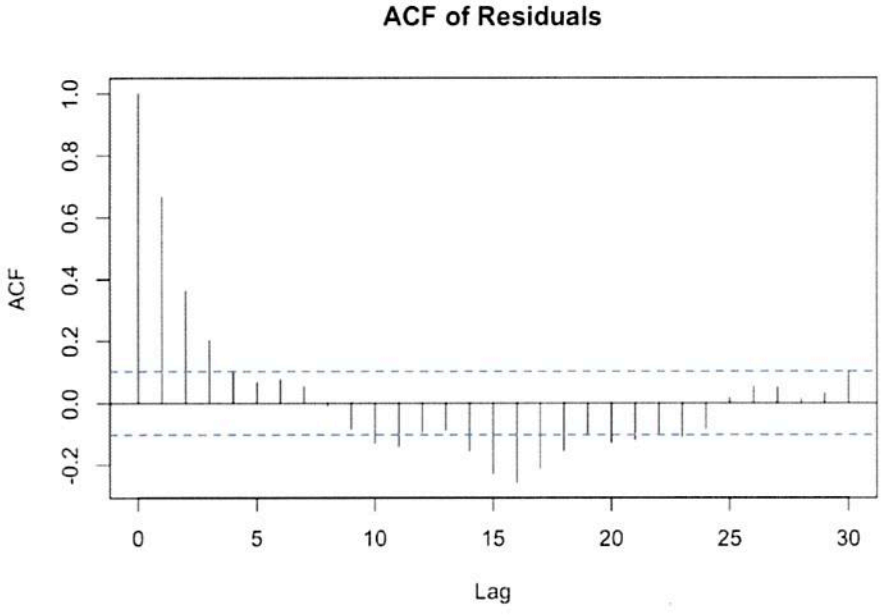
22. Describe how LOESS regression works in general terms.

makes a local polynomial model on a subset of the data (defined by a percent of the data nearby)

23. What are some reasons it might be beneficial to use a non-parametric nonlinear model for a regression problem rather than a parametric non-linear model?

parametric models make claims about the form of the model and can only adjust parameters. Nonlinear nonparametric models make no such assumptions, and are more flexible

24. An example of an ACF plot is shown below. How many lags should be used in an ARIMA model based on this graph?



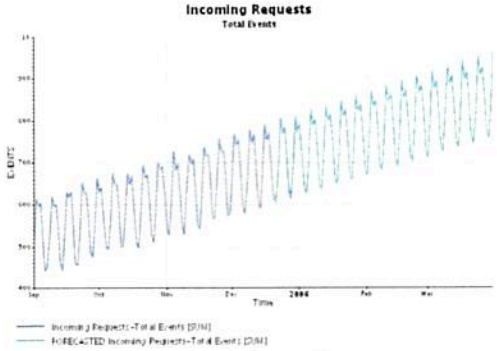

at least 3

25. What are some properties of seasonal time series?

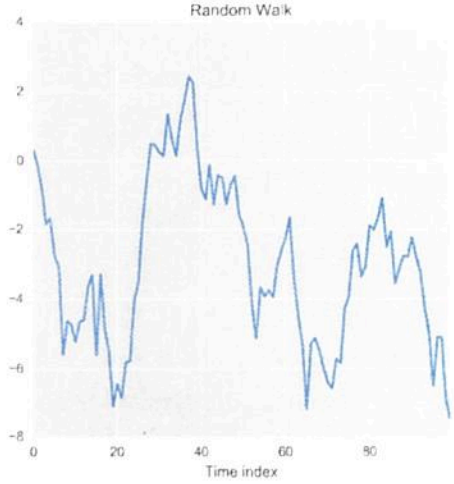
fixed pattern of repetitions or similar behaviors

26. For each of the graphs of time series below, identify the type of trend. Options include: stationary (no) trend, exponential trend, linear trend, polynomial trend, random walk, log trend.

linear

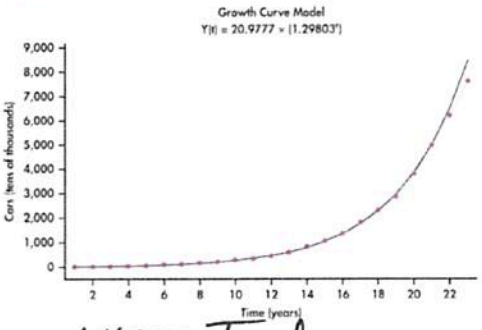



a.



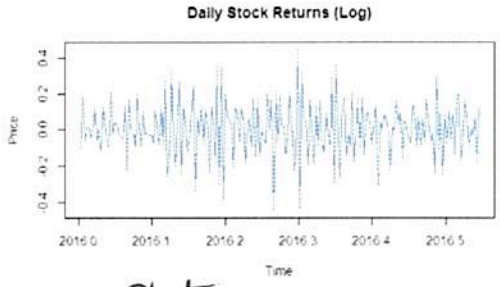

c.

random walk




b.

exponential




d.

Stationary

27. Explain why autocorrelation prevents us from using traditional regression to model some time series data.

if values are correlated errors may be too, but most regression methods assume independent errors.

28. Why are irregular time series so much more difficult to work with than regular time series? Describe some methods we can use to make irregular time series more regular.

most time series methods are based on the assumption of regular spacing, so unregular series don't match these conditions.

answers may vary -
regression & imputation
sampling

29. Describe each of the components of a SARIMA model.

Seasonal component - S

autoregression component - AR

integrated component - I

moving average component - MA

30. A confusion matrix is shown below. What is the accuracy of this model?

| | | Predicted: Yes | Predicted: No |
|---|---|---|---|
| Actual | Yes | 123 | 20 |
| | NO | 33 | 161 |

123 + 20 + 33 + 161 = 337

123 + 161 = 284

284 / 337 = 84.27%

31. How do we use the AUC (of an ROC curve) as a diagnostic for a classification model?

more accurate models produce higher AUC so we can use it to compare models.