# Regression Assignment 3

NWUDO CHIKAEZE FIDELIS JUNIOR

2022-11-10

```
## corrplot 0.92 loaded
```

1. Consider the data on commercial properties.

(a) Obtain the scatter plot matrix and the correlation matrix of the data. Interpret these and state your findings.

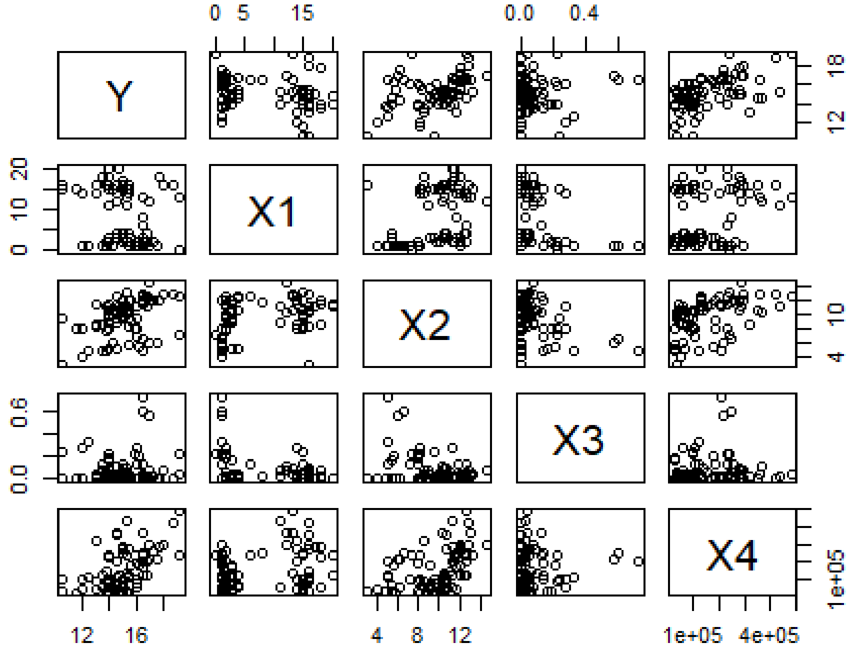
(b) Obtain the ANOVA table that decomposes the regression sum of squares into extra sum of squares associated with X4; with X1 given X4; with X2 given X1 and X4; and with X3 given X1, X2 and X4.

(c) There were three properties with no rental information available. The characteristics of the 3 properties are given in the dataset for this problem. Develop separate prediction intervals for the rental rates of these properties using a 95% confidence coefficient in each case. Can the rental rates of these three properties be predicted fairly precisely? Explain. What is the family confidence level for the set of three predictions?

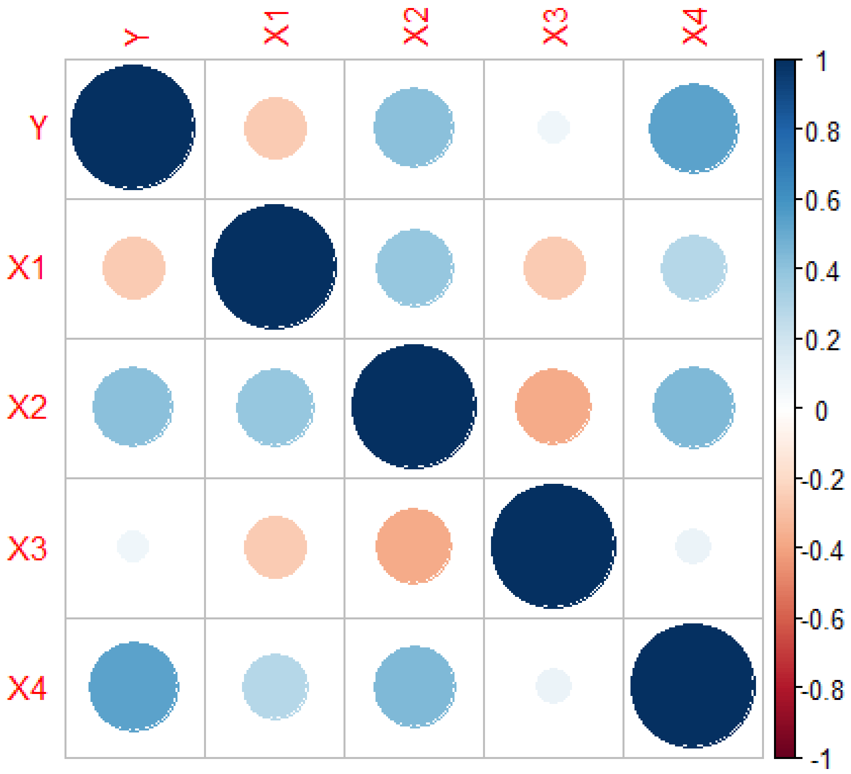
(d) Test whether B1 = 0.1, B2 = 0.4. Use alpha = 0.01. State the null and alternative hypotheses, the full and reduced estimated regression models, decision rule and conclusion.

QUESTION 1a

```
##             Y X1    X2   X3     X4
##  [1,] 13.500  1  5.02 0.14 123000
##  [2,] 12.000 14  8.19 0.27 104079
##  [3,] 10.500 16  3.00 0.00  39998
##  [4,] 15.000  4 10.70 0.05  57112
##  [5,] 14.000 11  8.97 0.07  60000
##  [6,] 10.500 15  9.45 0.24 101385
##  [7,] 14.000  2  8.00 0.19  31300
##  [8,] 16.500  1  6.62 0.60 248172
##  [9,] 17.500  1  6.20 0.00 215000
## [10,] 16.500  8 11.78 0.03 251015
## [11,] 17.000 12 14.62 0.08 291264
## [12,] 16.500  2 11.55 0.03 207549
## [13,] 16.000  2  9.63 0.00  82000
## [14,] 16.500 13 12.99 0.04 359665
## [15,] 17.225  2 12.01 0.03 265500
## [16,] 17.000  1 12.01 0.00 299000
## [17,] 16.000  1  7.99 0.14 189258
## [18,] 14.625 12 10.33 0.12 366013
## [19,] 14.500 16 10.67 0.00 349930
## [20,] 14.500  3  9.45 0.03  85335
## [21,] 16.500  6 12.65 0.13 235932
## [22,] 16.500  3 12.08 0.00 130000
## [23,] 15.000  3 10.52 0.05  40500
## [24,] 15.000  3  9.47 0.00  40500
## [25,] 13.000 14 11.62 0.00  45959
## [26,] 12.500  1  5.00 0.33 120000
## [27,] 14.000 15  9.89 0.05  81243
## [28,] 13.750 16 11.13 0.06 153947
## [29,] 14.000  2  7.96 0.22  97321
## [30,] 15.000 16 10.73 0.09 276099
## [31,] 13.750  2  7.95 0.00  90000
## [32,] 15.625  3  9.10 0.00 184000
## [33,] 15.625  3 12.05 0.03 184718
## [34,] 13.000 16  8.43 0.04  96000
## [35,] 14.000 16 10.60 0.04 106350
## [36,] 15.250 13 10.55 0.10 135512
## [37,] 16.250  1  5.50 0.21 180000
## [38,] 13.000 14  8.53 0.03 315000
## [39,] 14.500  3  9.04 0.04  42500
## [40,] 11.500 15  8.20 0.00  30005
## [41,] 14.250  1  6.13 0.00  60000
## [42,] 15.500 15  8.32 0.00  73521
## [43,] 12.000  1  4.00 0.00  50000
## [44,] 14.250 15 10.10 0.00  50724
## [45,] 14.000  3  5.25 0.16  31750
## [46,] 16.500  3 11.62 0.00 168000
## [47,] 14.500  4  5.31 0.00  70000
## [48,] 15.500  1  5.75 0.00  27000
## [49,] 16.750  4 12.46 0.03 129614
## [50,] 16.750  4 12.75 0.00 129614
## [51,] 16.750  2 12.75 0.00 130000
## [52,] 16.750  2 11.38 0.00 209000
## [53,] 17.000  1  5.99 0.57 220000
## [54,] 16.000  2 11.37 0.27  60000
## [55,] 14.500  3 10.38 0.00 110000
## [56,] 15.000 15 10.77 0.05 101206
## [57,] 15.000 17 11.30 0.00 288847
## [58,] 16.000  1  7.06 0.14 105000
## [59,] 15.500 14 12.10 0.05 276425
## [60,] 15.250  2 10.04 0.06  33000
## [61,] 16.500  1  4.99 0.73 210000
## [62,] 19.250  0  7.33 0.22 240000
## [63,] 17.750 18 12.11 0.00 281552
## [64,] 18.750 16 12.86 0.00 421000
## [65,] 19.250 13 12.70 0.04 484290
## [66,] 14.000 20 11.58 0.00 234493
## [67,] 14.000 18 11.58 0.03 230675
## [68,] 18.000 16 12.97 0.08 296966
## [69,] 13.750  1  4.82 0.00  32000
## [70,] 15.000  2  9.75 0.03  38533
## [71,] 15.500 16 10.36 0.02 109912
## [72,] 15.900  1  8.13 0.23 236000
## [73,] 15.250 15 13.23 0.05 243338
## [74,] 15.500  4 10.57 0.04 122183
## [75,] 14.750 20 11.22 0.00 128268
## [76,] 15.000  3 10.34 0.00  72000
## [77,] 14.500  3 10.67 0.00  43404
## [78,] 13.500 18  8.60 0.08  59443
## [79,] 15.000 15 11.97 0.14 254700
## [80,] 15.250 11 11.27 0.03 434746
## [81,] 14.500 14 12.68 0.03 201930
```



```
##              Y         X1         X2          X3         X4
## Y   1.00000000 -0.2502846  0.4137872  0.06652647 0.53526237
## X1 -0.25028456  1.0000000  0.3888264 -0.25266347 0.28858350
## X2  0.41378716  0.3888264  1.0000000 -0.37976174 0.44069713
## X3  0.06652647 -0.2526635 -0.3797617  1.00000000 0.08061073
## X4  0.53526237  0.2885835  0.4406971  0.08061073 1.00000000
```



- First, Y and X1 show a negative correlation when viewed in the correlation matrix. Y and X3 have a very weak positive association, but Y and X2 and Y and X4 have similar positive correlations.
- I can see some outliers in the scatter plot matrix, especially for X3. This plot also reveals to me that variables are abstract.

QUESTION 1b

```
##
## Call:
## lm(formula = Y ~ X4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1390 -0.7930  0.2890  0.9653  3.4415
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.378e+01  2.903e-01  47.482  < 2e-16 ***
## X4          8.437e-06  1.498e-06   5.632 2.63e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.462 on 79 degrees of freedom
## Multiple R-squared:  0.2865, Adjusted R-squared:  0.2775
## F-statistic: 31.72 on 1 and 79 DF,  p-value: 2.628e-07
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X4         1  67.775  67.775  31.723 2.628e-07 ***
## Residuals 79 168.782   2.136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Call:
## lm(formula = Y ~ X4 + X1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2032 -0.4593  0.0641  0.7730  2.5083
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.436e+01  2.771e-01  51.831  < 2e-16 ***
## X4           1.045e-05  1.363e-06   7.663 4.23e-11 ***
## X1          -1.145e-01  2.242e-02  -5.105 2.27e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.274 on 78 degrees of freedom
## Multiple R-squared:  0.4652, Adjusted R-squared:  0.4515
## F-statistic: 33.93 on 2 and 78 DF,  p-value: 2.506e-11

## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X4         1  67.775  67.775  41.788 8.076e-09 ***
## X1         1  42.275  42.275  26.065 2.275e-06 ***
## Residuals 78 126.508   1.622
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Call:
## lm(formula = Y ~ X4 + X1 + X2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0620 -0.6437 -0.1013  0.5672  2.9583
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.237e+01  4.928e-01  25.100  < 2e-16 ***
## X4           8.178e-06  1.305e-06   6.265 1.97e-08 ***
## X1           -1.442e-01  2.092e-02  -6.891 1.33e-09 ***
## X2            2.672e-01  5.729e-02   4.663 1.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.132 on 77 degrees of freedom
## Multiple R-squared:  0.583,  Adjusted R-squared:  0.5667
## F-statistic: 35.88 on 3 and 77 DF,  p-value: 1.295e-14
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X4         1 67.775  67.775  52.901 2.515e-10 ***
## X1         1 42.275  42.275  32.997 1.752e-07 ***
## X2         1 27.857  27.857  21.744 1.287e-05 ***
## Residuals 77 98.650   1.281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## lm(formula = Y ~ X4 + X1 + X2 + X3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1872 -0.5911 -0.0910  0.5579  2.9441
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.220e+01  5.780e-01  21.110  < 2e-16 ***
## X4           7.924e-06  1.385e-06   5.722 1.98e-07 ***
## X1          -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
## X2           2.820e-01  6.317e-02   4.464 2.75e-05 ***
## X3           6.193e-01  1.087e+00   0.570     0.57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.137 on 76 degrees of freedom
## Multiple R-squared:  0.5847, Adjusted R-squared:  0.5629
## F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X4         1 67.775  67.775 52.4369 3.073e-10 ***
## X1         1 42.275  42.275 32.7074 2.004e-07 ***
## X2         1 27.857  27.857 21.5531 1.412e-05 ***
## X3         1  0.420   0.420  0.3248    0.5704
## Residuals 76 98.231   1.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

final ANOVA

```
## Analysis of Variance Table
##
## Model 1: Y ~ X4
## Model 2: Y ~ X4 + X1
## Model 3: Y ~ X4 + X1 + X2
## Model 4: Y ~ X4 + X1 + X2 + X3
##   Res.Df     RSS Df Sum of Sq       F    Pr(>F)
## 1     79 168.782
## 2     78 126.508  1    42.275 32.7074 2.004e-07 ***
## 3     77  98.650  1    27.857 21.5531 1.412e-05 ***
## 4     76  98.231  1     0.420  0.3248    0.5704
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

QUESTION 1c

```
##   X1   X2   X3     X4
## 1  4 10.0 0.10  80000
## 2  6 11.5 0.00 120000
## 3 12 12.5 0.32 340000

##          [,1]
## [1,] 1.292508
```

confidence interval for property 1

```
##         [,1]
## [1,] 1.32895

## [1] 15.14785

##          [,1]
## [1,] 17.44385

##          [,1]
## [1,] 12.85185
```

- The prediction interval for property 1 is (12.85185, 17.44385)

Confidence interval for property 2

```
##          [,1]
## [1,] 1.330622

## [1] 15.54188

##          [,1]
## [1,] 17.83933
##          [,1]
## [1,] 13.24443
```

- The prediction interval for property 2 is (13.24443, 17.83933)

Confidence interval for property 3

```
##          [,1]
## [1,] 1.426945

## [1] 16.91334

##          [,1]
## [1,] 19.29248

##          [,1]
## [1,] 14.53419
```

- The prediction interval for property 3 is (14.53419, 19.29248)

- The family confidence interval for the set of three predictions include:

- Bonferroni; ynew ± t(1 − α/2r, n − 2)spred

```
##          [,1]
## [1,] 12.32569

##          [,1]
## [1,] 17.97001

##          [,1]
## [1,] 12.71794

##          [,1]
## [1,] 18.36582

##          [,1]
## [1,] 13.98897

##         [,1]
## [1,] 19.8377
```

- 85%
- Bonferroni confidence interval for property 1 is (12.32569, 17.97001)
- Bonferroni confidence interval for property 2 is (12.71794, 18.36582)
- Bonferroni confidence interval for property 3 is (13.98897, 19.8377)

QUESTION 1d

Ho: B1 = -0.1, B2 = 0.4, Ha: not both equalities in Ho holds

full model: y = Bo + B1X1 + B2X2 + B3X3 + B4X4 reduced model: y + 0.1X1 - 0.4X2 = Bo + B3X3 + B4X4

```
##   (Intercept)            X1            X2            X3            X4 
##  1.220059e+01 -1.420336e-01  2.820165e-01  6.193435e-01  7.924302e-06

## Analysis of Variance Table
## 
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)    
## X1         1 14.819  14.819 11.4649  0.001125 ** 
## X2         1 72.802  72.802 56.3262 9.699e-11 ***
## X3         1  8.381   8.381  6.4846  0.012904 *  
## X4         1 42.325  42.325 32.7464 1.976e-07 ***
## Residuals 76 98.231   1.293                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Variance Table
## 
## Response: Z
##           Df  Sum Sq Mean Sq F value    Pr(>F)    
## X3         1   9.205   9.205  6.5187   0.01263 *  
## X4         1  31.872  31.872 22.5713 9.058e-06 ***
## Residuals 78 110.141   1.412                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] 4.607303

## [1] 4.89584
```

Decision Rule:

- If the test statistic is below the critical value we accept the null hypothesis.
- Otherwise we reject.

Conclusion:

- Since the test statistic (4.607303) is less than critical value(4.89584), we cannot reject the null hypothesis.

2. A city tax assessor was interested in predicting residential home sales prices in a western city as a function of various characteristics of the home and surrounding property. Residential sales that occurred during the year 2002 were available from the city. Data from arms-length transactions include sales price, style, finished square feet, number of bedrooms, pool, lot size, year built, air conditioning, and whether or not the lot is adjacent to a highway. See Appendix C.7 of Applied Linear Statistical Models by Kutner et. al. for the description of the data set. Select the first 300 observations to use in the model-building data set.

(a) Develop a best subset model for predicting sales price. Justify your choice of model.

(b) Assess your model’s ability to predict and discuss its use as a tool for predicting sales price.

QUESTION 2a

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:MASS':
##
##     cement

## The following object is masked from 'package:datasets':
##
##     rivers

##               Best Subsets Regression
## ---------------------------------------------------
## Model Index    Predictors
## ---------------------------------------------------
##      1         C10
##      2         C3 C10
##      3         C3 C10 C11
##      4         C3 C10 C11 C12
##      5         C3 C9 C10 C11 C12
##      6         C3 C6 C9 C10 C11 C12
##      7         C3 C6 C7 C9 C10 C11 C12
##      8         C3 C6 C7 C9 C10 C11 C12 C13
##      9         C3 C4 C6 C7 C9 C10 C11 C12 C13
##     10         C3 C4 C5 C6 C7 C9 C10 C11 C12 C13
##     11         C3 C4 C5 C6 C7 C8 C9 C10 C11 C12 C13
## ---------------------------------------------------
##
##                                                                 Subsets
Regression Summary
## --------------------------------------------------------------------------
-----------------------------------------------------------------------------
---
##                        Adj.        Pred
## Model    R-Square    R-Square    R-Square      C(p)         AIC
SBIC          SBC           MSEP              FPE               HSP
APC
## --------------------------------------------------------------------------
-----------------------------------------------------------------------------
---
##   1        0.5783      0.5769      0.5721    239.8443    7745.4691
6891.9580    7756.5805    2.828662e+12    9491731139.6328    31747043.7475
0.4274
##   2        0.7073      0.7053      0.6993     77.9295    7637.9273
6785.3290    7652.7424    1.970007e+12    6632280856.9294    22184509.1548
0.2986
```

```
##   3        0.7322      0.7295      0.7225     48.2962    7613.2609
6760.9224    7631.7798    1.808564e+12    6108788176.8038    20435286.0509
0.2751
##   4        0.7439      0.7404      0.7318     35.4575    7601.8856
6749.7280    7624.1083    1.735585e+12    5881501422.1669    19677154.3063
0.2648
##   5        0.7621      0.7581      0.7485     14.2753    7581.7175
6730.2639    7607.6439    1.617461e+12    5499115605.5986    18400306.5168
0.2476
##   6        0.7653      0.7605      0.7497     12.2506    7579.6963
6728.4152    7609.3265    1.601391e+12    5462208161.4418    18279666.7756
0.2459
##   7        0.7686      0.7631      0.7519      9.9990    7577.3889
6726.3515    7610.7229    1.583987e+12    5420380933.0333    18142927.2092
0.2441
##   8        0.7708      0.7645      0.7529      9.2519    7576.5726
6725.7483    7613.6104    1.574598e+12    5405681339.6738    18097359.6909
0.2434
##   9        0.7722      0.7651      0.7526      9.4826    7576.7446
6726.1054    7617.4862    1.570448e+12    5408816533.4878    18111898.2034
0.2435
##  10        0.7732      0.7653      0.7507     10.2110    7577.4239
6726.9573    7621.8693    1.568978e+12    5421122664.8420    18157565.1957
0.2441
##  11        0.7733      0.7647      0.7489     12.0000    7579.2042
6728.8376    7627.3533    1.573293e+12    5453442287.3711    18270712.5682
0.2456
## ----------------------------------------------------------------------------
-----------------------------------------------------------------------------
---
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

- We see that Model 10 with C3, C4, C5, C6, C7, C9, C10, C11, C12, C13 as predictor variables is selected based on the R2 adjusted criterion because this model has the largest value of R2 adjusted. The C(p) criterion leads to model 8 with predictor variables C3, C6, C7, C9, C10, C11, C12, C13, because the C(p) value for this model is near k=9 and is small. This 8 predictor variable model is also selected by the AIC criterion because it has the smallest AIC value. Meanwhile, the SBC criterion selects model 5. Model diagnostics and the generalized linear test approach can then be used to select the best among these three competing models.

Comparing model 10 and 8, Testing significance of C4, C5 in model 10

```
##
## Call:
## lm(formula = C2 ~ C3 + C4 + C5 + C6 + C7 + C9 + C10 + C11 + C12 +
##     C13)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -194753  -40404   -4459   38616  260442
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.434e+06  6.002e+05  -4.055 6.46e-05 ***
## C3           1.134e+02  1.074e+01  10.560  < 2e-16 ***
## C4          -8.378e+03  5.138e+03  -1.630   0.1041
## C5           7.363e+03  6.520e+03   1.129   0.2598
## C6          -3.219e+04  1.677e+04  -1.919   0.0559 .
## C7           1.603e+04  7.847e+03   2.043   0.0420 *
## C9           1.345e+03  3.034e+02   4.432 1.32e-05 ***
## C10         -9.156e+04  1.199e+04  -7.634 3.34e-13 ***
## C11         -1.021e+04  2.045e+03  -4.994 1.03e-06 ***
## C12          1.616e+00  3.835e-01   4.214 3.36e-05 ***
## C13         -3.622e+04  2.314e+04  -1.565   0.1186
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72310 on 289 degrees of freedom
## Multiple R-squared:  0.7732, Adjusted R-squared:  0.7653
## F-statistic: 98.51 on 10 and 289 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: C2
##            Df     Sum Sq    Mean Sq  F value    Pr(>F)
## C3          1 3.7600e+12 3.7600e+12 719.0206 < 2.2e-16 ***
## C4          1 4.5527e+10 4.5527e+10   8.7061 0.0034311 **
## C5          1 1.0325e+11 1.0325e+11  19.7450 1.262e-05 ***
## C6          1 6.2188e+10 6.2188e+10  11.8920 0.0006476 ***
## C7          1 2.5061e+11 2.5061e+11  47.9242 2.867e-11 ***
## C9          1 1.6389e+11 1.6389e+11  31.3410 5.032e-08 ***
## C10         1 4.9674e+11 4.9674e+11  94.9911 < 2.2e-16 ***
## C11         1 1.6807e+11 1.6807e+11  32.1389 3.476e-08 ***
## C12         1 8.8243e+10 8.8243e+10  16.8745 5.205e-05 ***
## C13         1 1.2812e+10 1.2812e+10   2.4501 0.1186133
## Residuals 289 1.5113e+12 5.2294e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## lm(formula = C2 ~ C3 + C6 + C7 + C9 + C10 + C11 + C12 + C13)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -213764  -45487   -3915   39863  269527
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.580e+06  5.877e+05  -4.390 1.59e-05 ***
## C3           1.119e+02  9.445e+00  11.852  < 2e-16 ***
## C6          -3.633e+04  1.657e+04  -2.193   0.0291 *
## C7           1.612e+04  7.857e+03   2.052   0.0411 *
## C9           1.422e+03  2.964e+02   4.796 2.59e-06 ***
## C10         -9.521e+04  1.183e+04  -8.050 2.13e-14 ***
## C11         -1.005e+04  2.035e+03  -4.937 1.34e-06 ***
## C12          1.629e+00  3.801e-01   4.287 2.47e-05 ***
## C13         -3.830e+04  2.312e+04  -1.657   0.0987 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72440 on 291 degrees of freedom
## Multiple R-squared:  0.7708, Adjusted R-squared:  0.7645
## F-statistic: 122.3 on 8 and 291 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: C2
##            Df     Sum Sq    Mean Sq  F value    Pr(>F)
## C3          1 3.7600e+12 3.7600e+12 716.4374 < 2.2e-16 ***
## C6          1 6.5673e+10 6.5673e+10  12.5133 0.0004701 ***
## C7          1 2.8669e+11 2.8669e+11  54.6254 1.556e-12 ***
## C9          1 2.1453e+11 2.1453e+11  40.8762 6.439e-10 ***
## C10         1 5.4185e+11 5.4185e+11 103.2444 < 2.2e-16 ***
## C11         1 1.6036e+11 1.6036e+11  30.5553 7.216e-08 ***
## C12         1 9.1893e+10 9.1893e+10  17.5094 3.790e-05 ***
## C13         1 1.4405e+10 1.4405e+10   2.7447 0.0986575 .
## Residuals 291 1.5272e+12 5.2482e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ho: B4=B5=0 Ha: B4!=0, or B5!=0 or both !=0 SSE(full) = 1.5113e+12 DF(full) = 289 SSE(reduced) = 1.5416e+12 DF(reduced) = 292

```
## [1] 1.520247

## [1] 0.2204011

## [1] 3.027001
```

- Since Fo(1.520247) is not greater than F critical(3.027001) we cannot reject Ho.

Comparing model 8 & 5 testing significance of C6, C7, C13 on model 8

```
##
## Call:
## lm(formula = C2 ~ C3 + C6 + C7 + C9 + C10 + C11 + C12 + C13)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -213764  -45487   -3915   39863  269527
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.580e+06  5.877e+05  -4.390 1.59e-05 ***
## C3           1.119e+02  9.445e+00  11.852  < 2e-16 ***
## C6          -3.633e+04  1.657e+04  -2.193   0.0291 *
## C7           1.612e+04  7.857e+03   2.052   0.0411 *
## C9           1.422e+03  2.964e+02   4.796 2.59e-06 ***
## C10         -9.521e+04  1.183e+04  -8.050 2.13e-14 ***
## C11         -1.005e+04  2.035e+03  -4.937 1.34e-06 ***
## C12          1.629e+00  3.801e-01   4.287 2.47e-05 ***
## C13         -3.830e+04  2.312e+04  -1.657   0.0987 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72440 on 291 degrees of freedom
## Multiple R-squared:  0.7708, Adjusted R-squared:  0.7645
## F-statistic: 122.3 on 8 and 291 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: C2
##            Df     Sum Sq    Mean Sq  F value    Pr(>F)
## C3          1 3.7600e+12 3.7600e+12 716.4374 < 2.2e-16 ***
## C6          1 6.5673e+10 6.5673e+10  12.5133 0.0004701 ***
## C7          1 2.8669e+11 2.8669e+11  54.6254 1.556e-12 ***
## C9          1 2.1453e+11 2.1453e+11  40.8762 6.439e-10 ***
## C10         1 5.4185e+11 5.4185e+11 103.2444 < 2.2e-16 ***
## C11         1 1.6036e+11 1.6036e+11  30.5553 7.216e-08 ***
## C12         1 9.1893e+10 9.1893e+10  17.5094 3.790e-05 ***
## C13         1 1.4405e+10 1.4405e+10   2.7447 0.0986575 .
## Residuals 291 1.5272e+12 5.2482e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## lm(formula = C2 ~ C3 + C9 + C10 + C11 + C12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -215613  -46584   -5771   44152  273677
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -2.420e+06  5.661e+05  -4.275 2.58e-05 ***
## C3           1.166e+02  9.329e+00  12.494  < 2e-16 ***
## C9           1.333e+03  2.807e+02   4.748 3.21e-06 ***
## C10         -9.597e+04  1.148e+04  -8.360 2.54e-15 ***
## C11         -9.676e+03  2.029e+03  -4.768 2.93e-06 ***
## C12          1.848e+00  3.731e-01   4.953 1.23e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 73430 on 294 degrees of freedom
## Multiple R-squared:  0.7621, Adjusted R-squared:  0.7581 
## F-statistic: 188.4 on 5 and 294 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
## 
## Response: C2
##            Df     Sum Sq    Mean Sq F value    Pr(>F)    
## C3          1 3.7600e+12 3.7600e+12 697.427 < 2.2e-16 ***
## C9          1 3.9843e+11 3.9843e+11  73.902 4.952e-16 ***
## C10         1 6.0827e+11 6.0827e+11 112.824 < 2.2e-16 ***
## C11         1 1.7863e+11 1.7863e+11  33.133 2.163e-08 ***
## C12         1 1.3227e+11 1.3227e+11  24.534 1.234e-06 ***
## Residuals 294 1.5850e+12 5.3913e+09                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

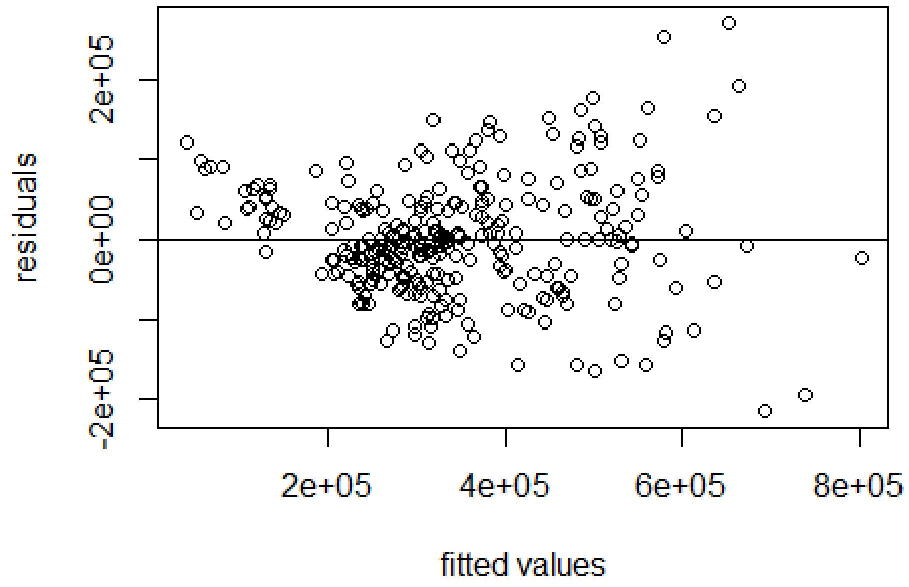
Ho: B6=B7=B13=0 Ha: B6!=0, or B7!=0, or B13!=0 or all three !=0 SSE(full) = 1.5272e+12 DF(full) = 291 SSE(reduced) = 1.5850e+12 DF(reduced) = 294

```
## [1] 3.671163

## [1] 0.01268928

## [1] 2.635629
```

- Since Fo(3.671163) is greater than F_critical(2.635629), we reject the null hypothesis. That means, there is evidence in the data in support of Ha, that, C6, C7and C13 cannot be dropped from model 8.

- Based on my statistical test perform on all three models, and comparing significance of C4, C5 in model 10 as well as the significance of C6, C7, C13 on model 8, i will chose model 8 as my choice of model.

QUESTION 2b



- The assumption of constant variance is not valid as the residuals are seen to fan out as we move along the fitted values axis. So i performed a log transformation on the C2 values as a remedial measure for constant variance problem.

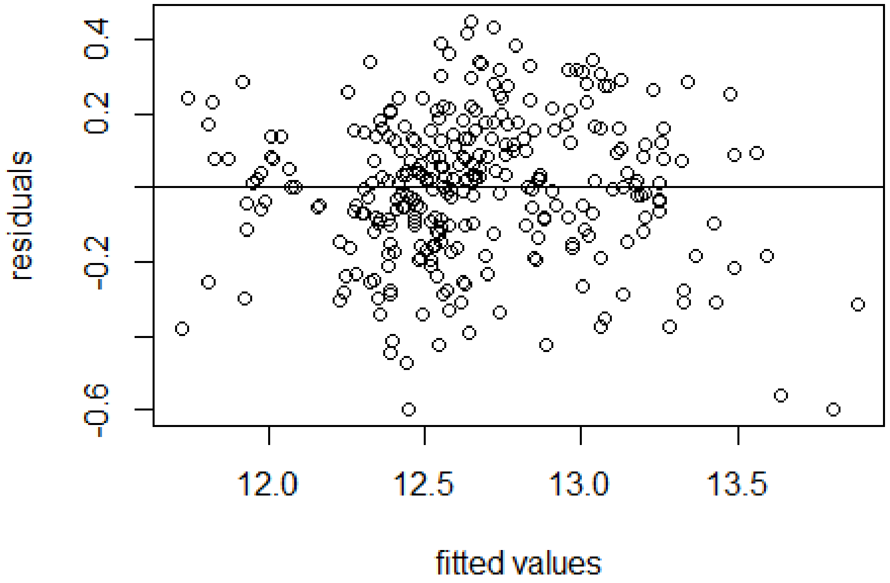
- As an alternative, Performing a Breusch-pagan test to check for constant variance.

Ho: C3=C6=C7=C9=C10=C11=C12=C13=0, Ha: At least one C != 0

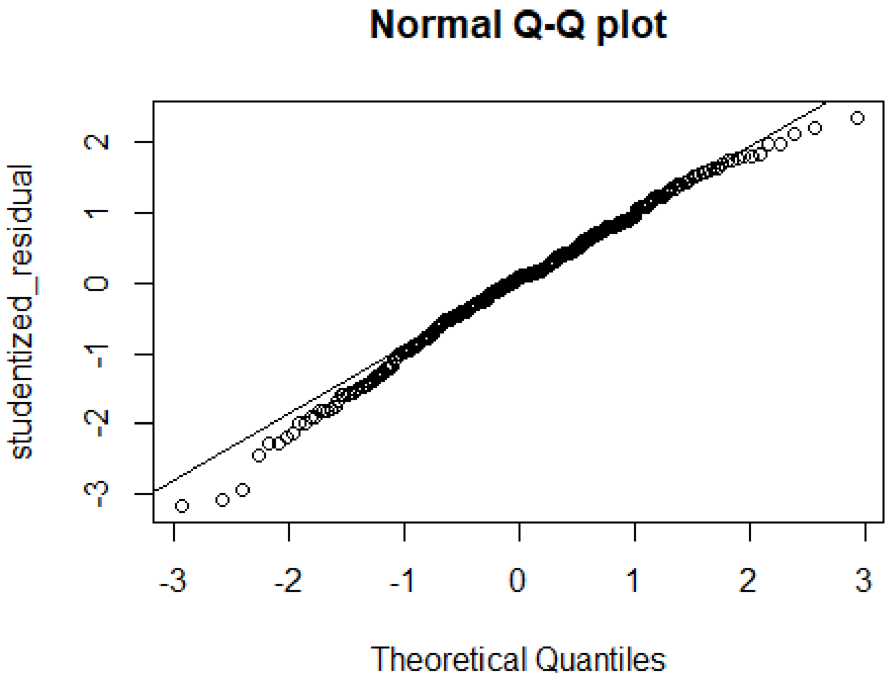
```
## Analysis of Variance Table
##
## Response: C2
##            Df     Sum Sq    Mean Sq  F value    Pr(>F)
## C3          1 3.7600e+12 3.7600e+12 716.4374 < 2.2e-16 ***
## C6          1 6.5673e+10 6.5673e+10  12.5133 0.0004701 ***
## C7          1 2.8669e+11 2.8669e+11  54.6254 1.556e-12 ***
## C9          1 2.1453e+11 2.1453e+11  40.8762 6.439e-10 ***
## C10         1 5.4185e+11 5.4185e+11 103.2444 < 2.2e-16 ***
## C11         1 1.6036e+11 1.6036e+11  30.5553 7.216e-08 ***
## C12         1 9.1893e+10 9.1893e+10  17.5094 3.790e-05 ***
## C13         1 1.4405e+10 1.4405e+10   2.7447 0.0986575 .
## Residuals 291 1.5272e+12 5.2482e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Analysis of Variance Table
##
## Response: xsq
##            Df     Sum Sq    Mean Sq F value    Pr(>F)
## C3          1 2.7948e+21 2.7948e+21 45.3547 8.771e-11 ***
## C6          1 4.0228e+19 4.0228e+19  0.6528   0.41976
## C7          1 1.2236e+19 1.2236e+19  0.1986   0.65621
## C9          1 3.1547e+19 3.1547e+19  0.5120   0.47487
## C10         1 3.8080e+20 3.8080e+20  6.1797   0.01348 *
## C11         1 1.2321e+21 1.2321e+21 19.9948 1.114e-05 ***
## C12         1 3.0930e+20 3.0930e+20  5.0193   0.02582 *
## C13         1 6.7508e+18 6.7508e+18  0.1096   0.74089
## Residuals 291 1.7932e+22 6.1621e+19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] 345.9787

## [1] 331.7856
```

- since Xo is greater than X-critical, we reject the null hypothesis and as a result, the assumption of constant variance is not valid.

- Below, i performed a log transformation on C2



- From here, the residuals are seen the to fall within a horizontal band centered around the 0 line, displaying no systematic pattern of positive and negative values. this shows that the assumption of linearity as well as constant variance is valid.



- Quantile-Quantile plot to check for normality of error terms. As we can see, the points at the center are approximately on the line indicating that the assumption of normality is valid.

3. The primary objective of the study on the efficacy of nosocomial (hospital-acquired) infection control (SENIC Project) was to determine whether infection surveillance and control programs have reduced the rates of nosocomial infection in United States hospitals.For predicting the average length of stay of patients in a hospital (Y ), it has been decided to include age (X1) and infection risk (X2) as predictor variables. The question now is whether an additional predictor variable would be helpful in the model and, if so, which variable would be most helpful. See Appendix C.1 of Applied Linear Statistical Models by Kutner et. al. for the description of the data set. Assume that a first-order multiple regression model is appropriate.

(a) For each of the following variables, calculate the coefficient of partial determination given that X1 and X2 are in the model: routine culturing ratio (X3), average daily census X4, number of nurses X5, and available facilities and services X6.

(b) On the basis of the results in part (a), which of the four additional predictor variables is best? Explain.

(c) Using the F test, test whether the variable determined to be best in part (b) is helpful in a regression model containing X1 and X2. Use alpha = 0:05. State the null and alternative hypotheses, the full and reduced models, decision rule and conclusion.

(d) It is of interest to examine whether the effect on length of stay for hospitals located in the western region differs from that for hospitals located in the other three regions. Regress length of stay on age, routine culturing ratio, average daily census, available facilities and region. Construct an appropriate confidence interval for each pairwise comparison. Take alpha=0.05 number of comparisons . Summarize your findings.

QUESTION 3a

```
## Analysis of Variance Table
## 
## Response: Y
##            Df  Sum Sq Mean Sq F value    Pr(>F)    
## X1          1  14.604  14.604  5.7734   0.01794 *  
## X2          1 116.356 116.356 45.9986 6.188e-10 ***
## Residuals 110 278.250   2.530                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
## 
## Response: Y
##            Df  Sum Sq Mean Sq F value  Pr(>F)    
## X1          1  14.604  14.604  5.7885 0.01781 *  
## X2          1 116.356 116.356 46.1188 6.1e-10 ***
## X3          1   3.248   3.248  1.2874 0.25902    
## Residuals 109 275.002   2.523                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
## 
## Response: Y
##            Df  Sum Sq Mean Sq F value    Pr(>F)    
## X1          1  14.604  14.604   6.623   0.01141 *  
## X2          1 116.356 116.356  52.768 5.928e-11 ***
## X4          1  37.899  37.899  17.187 6.722e-05 ***
## Residuals 109 240.352   2.205                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
## 
## Response: Y
##            Df  Sum Sq Mean Sq F value    Pr(>F)    
## X1          1  14.604  14.604   5.943   0.01639 *  
## X2          1 116.356 116.356  47.350 3.931e-10 ***
## X5          1  10.397  10.397   4.231   0.04208 *
## Residuals 109 267.853   2.457
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Variance Table
##
## Response: Y
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## X1          1  14.604  14.604  5.9370   0.01645 *
## X2          1 116.356 116.356 47.3017 3.998e-10 ***
## X6          1  10.125  10.125  4.1162   0.04491 *
## Residuals 109 268.125   2.460
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] 0.0116478

## [1] 0.1362049

## [1] 0.03736568

## [1] 0.03638814
```

QUESTION 3b

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2413 -0.7188 -0.1385  0.7706  7.8325
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.52455    1.91664   0.795   0.4281
## X1           0.09152    0.03497   2.617   0.0101 *
## X2           0.67105    0.13672   4.908 3.24e-06 ***
## X3           0.02086    0.01839   1.135   0.2590
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.588 on 109 degrees of freedom
## Multiple R-squared:  0.328,  Adjusted R-squared:  0.3095
## F-statistic: 17.73 on 3 and 109 DF,  p-value: 1.915e-09

##
## Call:
## lm(formula = Y ~ X1 + X2 + X4)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max 
## -2.6746 -0.9136 -0.1339  0.7457  7.7411 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.6243147  1.7430795   0.932   0.3535    
## X1          0.0884837  0.0315055   2.809   0.0059 ** 
## X2          0.5807868  0.1132304   5.129 1.27e-06 ***
## X4          0.0040999  0.0009889   4.146 6.72e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.485 on 109 degrees of freedom
## Multiple R-squared:  0.4126, Adjusted R-squared:  0.3965 
## F-statistic: 25.53 on 3 and 109 DF,  p-value: 1.387e-12
```

```
## 
## Call:
## lm(formula = Y ~ X1 + X2 + X5)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.0580 -0.8241 -0.1065  0.6724  7.9144 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.723830   1.843554   0.935   0.3518    
## X1          0.086906   0.033337   2.607   0.0104 *  
## X2          0.662303   0.120268   5.507 2.46e-07 ***
## X5          0.002390   0.001162   2.057   0.0421 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.568 on 109 degrees of freedom
## Multiple R-squared:  0.3454, Adjusted R-squared:  0.3274 
## F-statistic: 19.17 on 3 and 109 DF,  p-value: 4.67e-10
```

```
## 
## Call:
## lm(formula = Y ~ X1 + X2 + X6)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.9998 -0.8594 -0.1823  0.7109  7.7619 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.38646    1.86622   0.743   0.4591    
## X1           0.08371    0.03325   2.518   0.0133 *  
## X2           0.65845    0.12135   5.426 3.52e-07 ***
## X6           0.02174    0.01071   2.029   0.0449 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.568 on 109 degrees of freedom
## Multiple R-squared:  0.3448, Adjusted R-squared:  0.3267
## F-statistic: 19.12 on 3 and 109 DF,  p-value: 4.931e-10
```

- The model with X4 added to it is the best model based on it having the the highest R2 and R2 adjusted value.

QUESTION 3c

full model <- Y = Bo + B1X1 + B2X2 + B4X4, reduced model <- Y = Bo + B1X1 + B2X2, testing for the significance of X4 in full model, null hypothesis: Ho: B4 = 0, Alternative hypothesis: Ha: B4 != 0 SSE_red = 278.250 SSE_full = 240.352 DF_red = 110 DF_full = 109

```
## [1] 17.1868

## [1] 3.928195

## [1] 6.723399e-05
```

Decision Rule:

- If the test statistic is below the critical value we accept the null hypothesis.
- Otherwise we reject.

Conclusion:

- Since the test statistic (17.1868) is greater than critical value(3.928195), we reject the null hypothesis.Therefore, X4(average daily census) is significant in the prediction of Y(length of stay). Alternatively my P value (6.723399e-05) is less than alpha value 0.05

QUESTION 3d

- Geographic region, where: 1 =NE, 2=NC, 3=S, 4=W
- full model; y = Bo + B1X1 + B2X3 + B3X4 + B4X6 + B5D1 + B6D2 + B7D3
- $t(1 - \alpha/2r, n - 3)$

```
##
## Call:
## lm(formula = Y ~ X1 + X3 + X4 + X6 + D1 + D2 + D3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7938 -0.7304  0.0037  0.5388  7.7231
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.047830   1.812955   1.130  0.26124
## X1           0.103691   0.031459   3.296  0.00134 **
## X3           0.040302   0.014303   2.818  0.00578 **
## X4           0.006600   0.001404   4.700 7.92e-06 ***
## X6          -0.020761   0.014369  -1.445  0.15148
## D1           2.149988   0.461517   4.659 9.37e-06 ***
## D2           1.190333   0.437058   2.724  0.00757 **
## D3           0.633478   0.427554   1.482  0.14143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.399 on 105 degrees of freedom
## Multiple R-squared:  0.4981, Adjusted R-squared:  0.4647
## F-statistic: 14.89 on 7 and 105 DF,  p-value: 2.283e-13

## [1] 2.432633
```

- B = t(0.99166,105) = 2.432633
- standard error: B5=0.461517, B6=0.437058, B7= 0.427554
- estimates : B5=2.149988, B6=1.190333, B7=0.633478

confidence interval;

1) 2.149988 ± 2.432633(0.461517) -> (1.0273 <= B5 <= 3.2727)

2) 1.190333 ± 2.432633(0.437058) -> (0.1271 <= B6 <= 2.2535)
3) 0.633478 ± 2.432633(0.427554) -> (-0.4066 <= B7 <= 1.6736 )