HW#3

Halid Kopanski

2022-06-12

LMR 4.1

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

a)

```
##        fit      lwr      upr
## 1 2.389053 2.172437 2.605669
```

```
##        fit       lwr      upr
## 1 2.389053 0.9646584 3.813447
```

```
## [1] 63.86598
```

b)

```
##       fit      lwr      upr
## 1 3.17454 2.270398 4.078682
```

```
##       fit      lwr      upr
## 1 3.17454 1.501384 4.847695
```

c)

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72964 -0.45764  0.02812  0.46403  1.57013
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.26809    0.54350  -0.493  0.62298
## lcavol       0.55164    0.07467   7.388  6.3e-11 ***
## lweight      0.50854    0.15017   3.386  0.00104 **
## svi          0.66616    0.20978   3.176  0.00203 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7168 on 93 degrees of freedom
## Multiple R-squared:  0.6264, Adjusted R-squared:  0.6144
## F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

```
##        fit      lwr      upr
## 1 2.372534 2.197274 2.547794
```

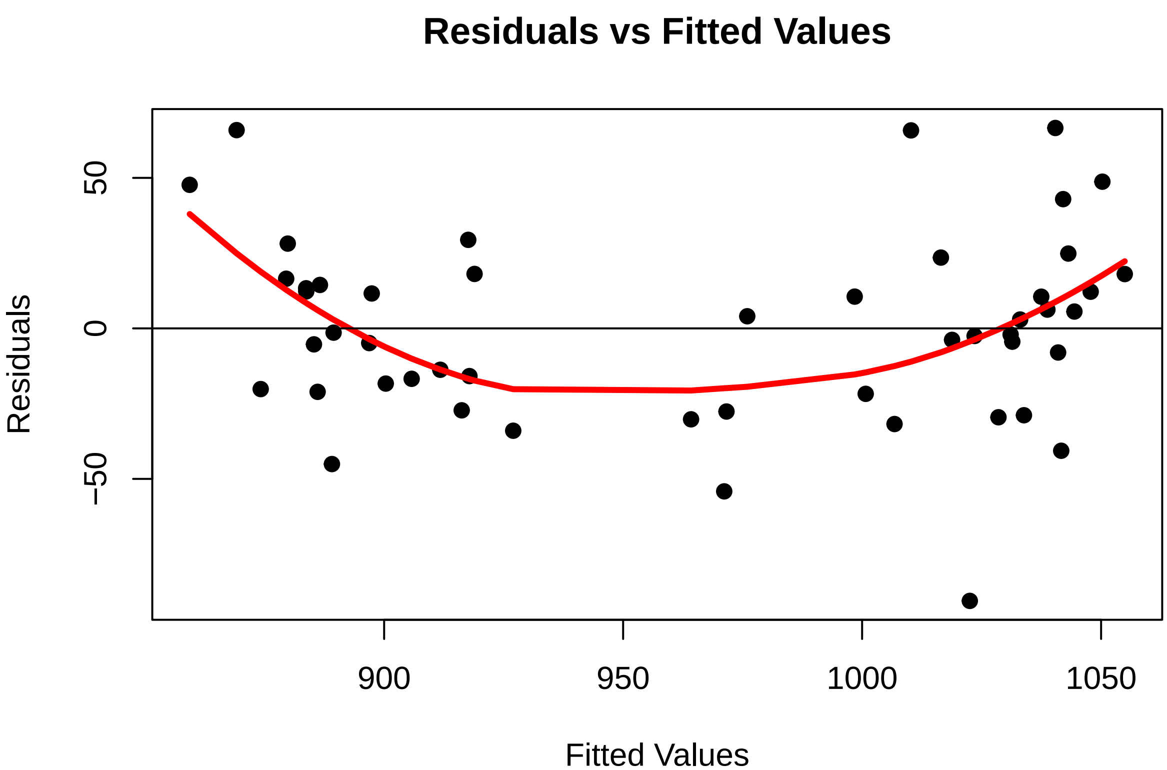
```
##        fit       lwr      upr
## 1 2.372534 0.9383436 3.806724
```

```
## Analysis of Variance Table
##
## Model 1: lpsa ~ lcavol + lweight + svi
## Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     93 47.785
## 2     88 44.163  5    3.6218 1.4434 0.2167
```

LMR 6.1

```
##
## Call:
## lm(formula = total ~ expend + salary + ratio + takers, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.531 -20.855  -1.746  15.979  66.571
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
## expend         4.4626    10.5465   0.423    0.674
## salary         1.6379     2.3872   0.686    0.496
## ratio         -3.6242     3.2154  -1.127    0.266
## takers        -2.9045     0.2313 -12.559 2.61e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.7 on 45 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.809
## F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

a)



Results of the Breusch - Pagan Test

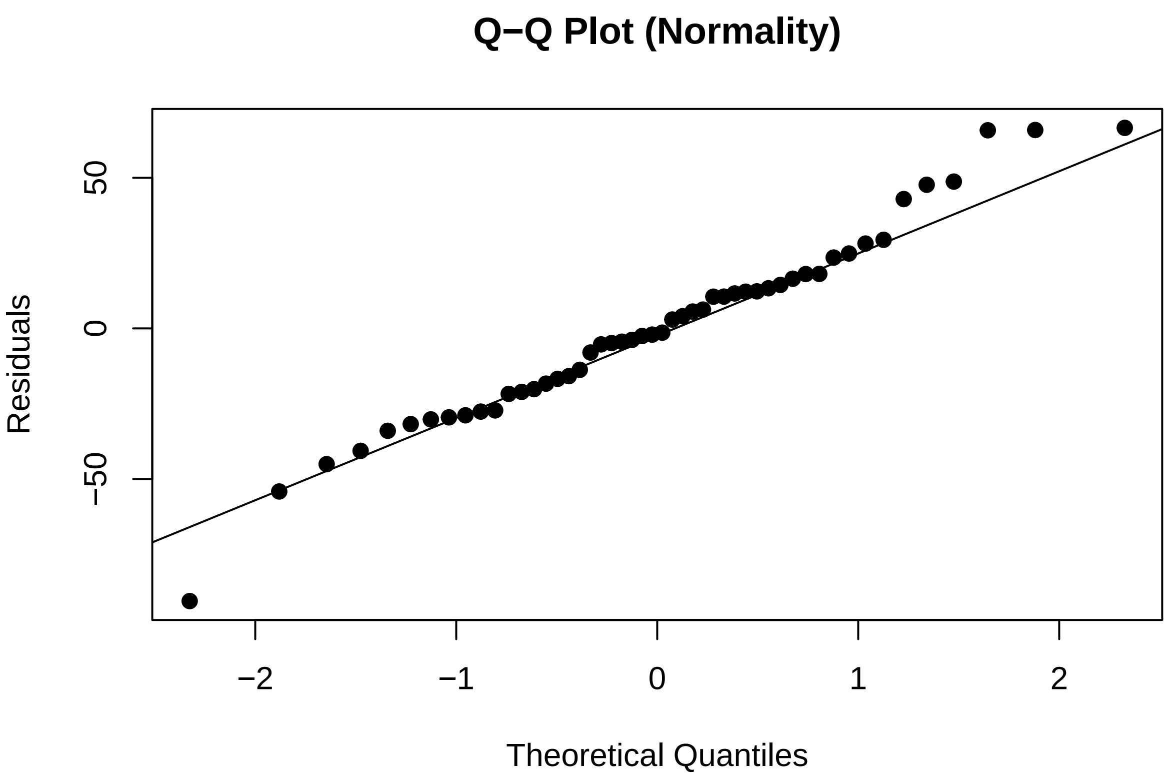
```
##
```

```
##  Breusch-Pagan test
##
## data:  sat_model1
## BP = 2.7234, df = 4, p-value = 0.6051
```

The results of the Non Constant Variance Test

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.6972119, Df = 1, p = 0.40372
```

b)

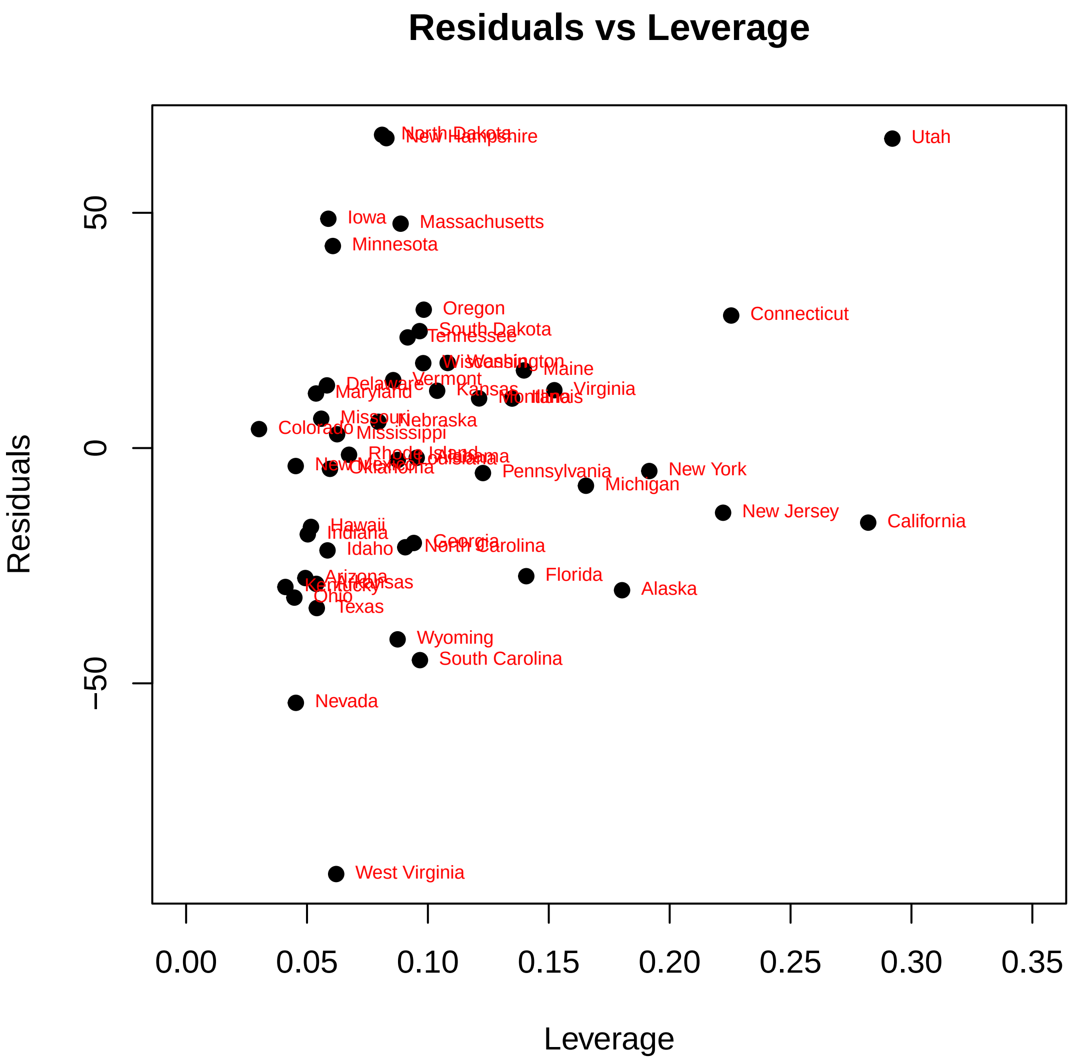


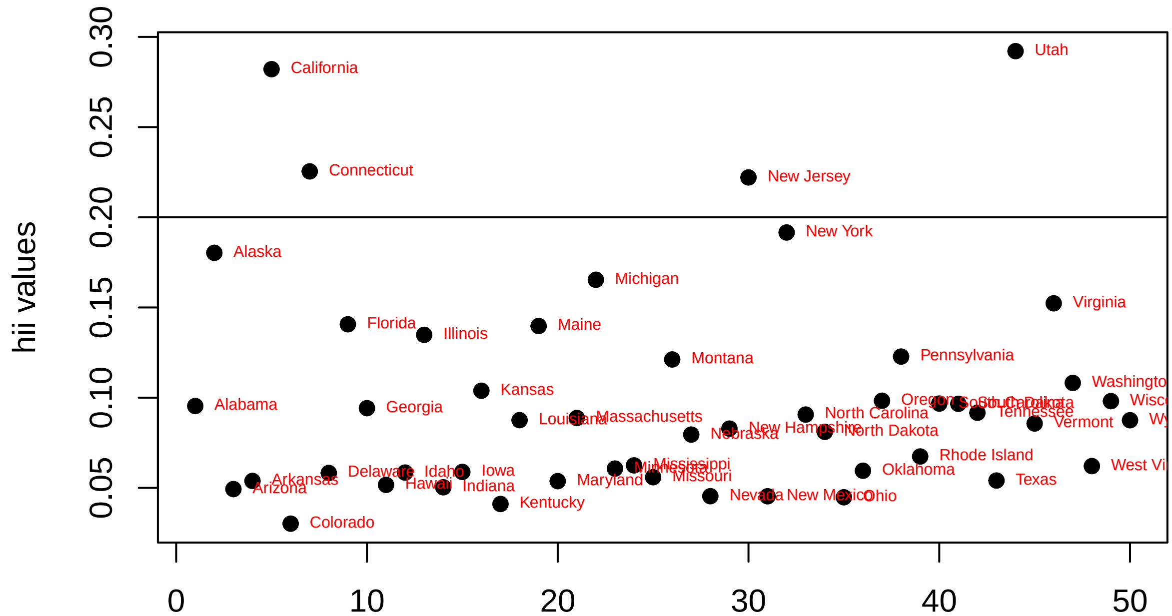
Results of the Shapiro and the Durbin-Watson Tests

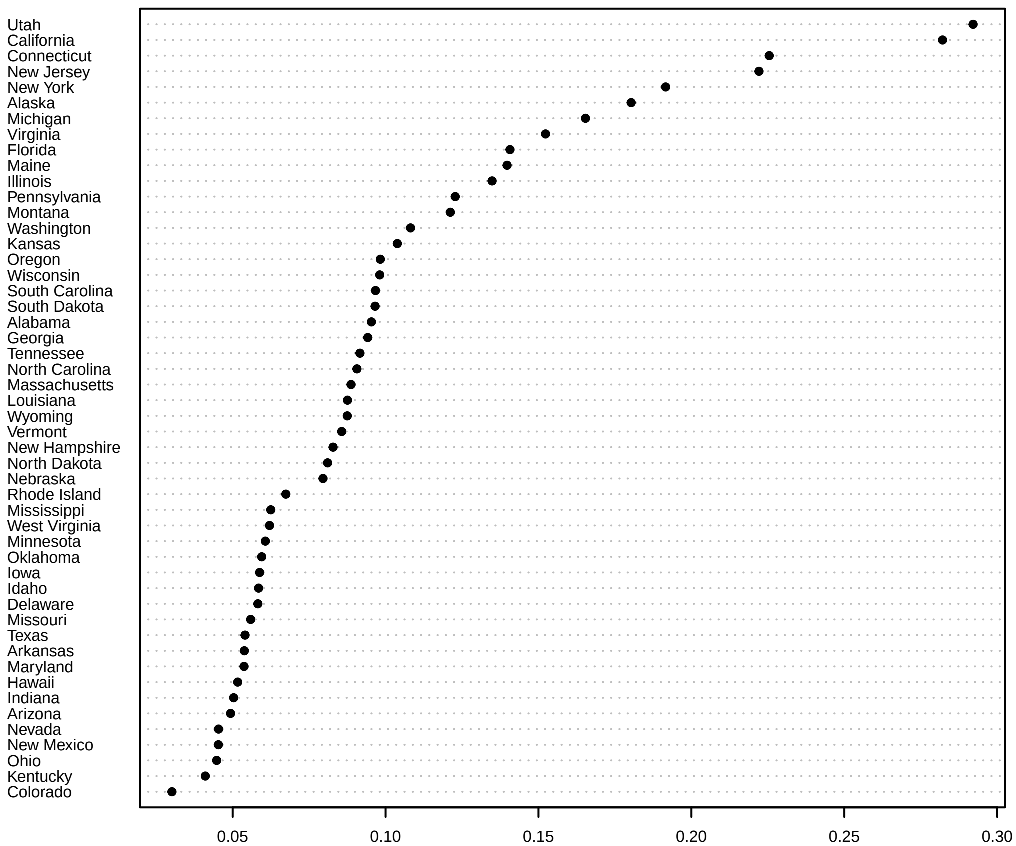
```
##
##  Shapiro-Wilk normality test
##
## data:  sat_model1$residuals
## W = 0.97691, p-value = 0.4304
```

```
##
##  Durbin-Watson test
##
## data:  sat_model1
## DW = 2.4525, p-value = 0.9459
## alternative hypothesis: true autocorrelation is greater than 0
```

c)

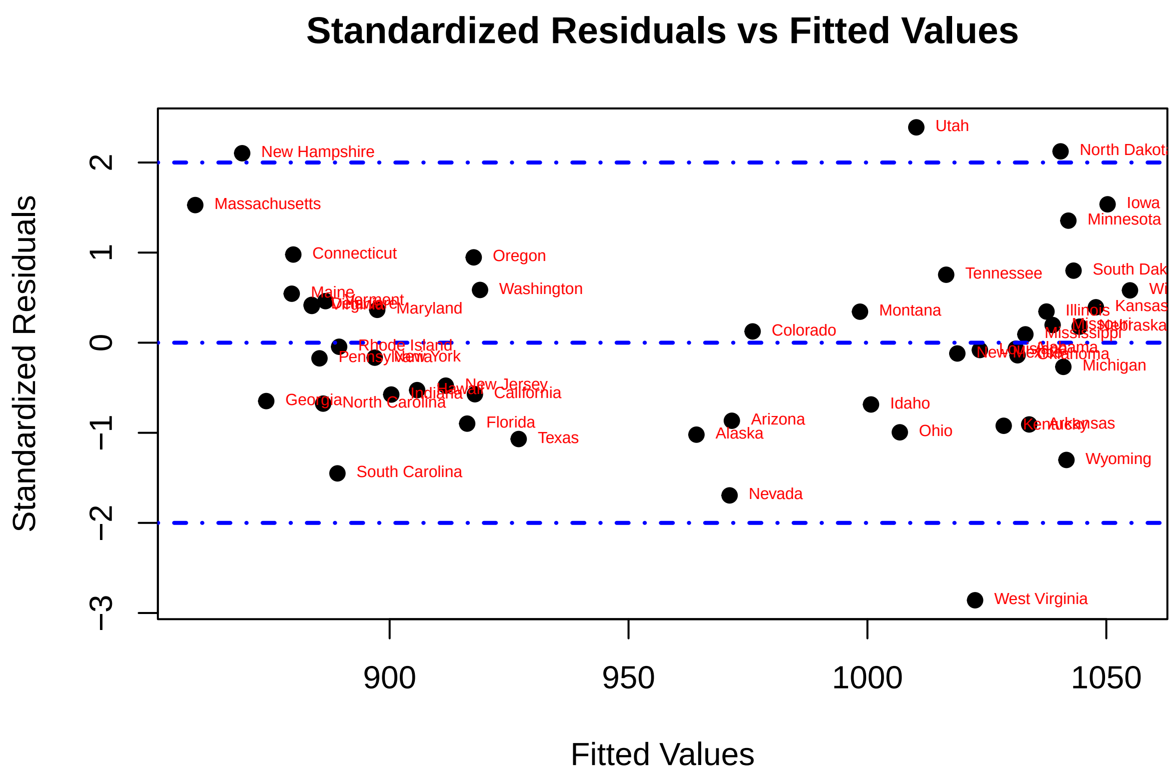

Residuals vs Leverage
Residuals
Leverage
50
0
−50
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Utah
Iowa
Massachusetts
Minnesota
Oregon
Connecticut
South Dakota
Tennessee
Maine
Virginia
Kansas
Maryland
Vermont
Missouri
Nebraska
Colorado
Mississippi
Rhode Island
Oklahoma
Pennsylvania
New York
Michigan
New Jersey
California
Hawaii
Indiana
Idaho
North Carolina
Florida
Alaska
Arizona
Ohio
Texas
Wyoming
South Carolina
Nevada
West Virginia

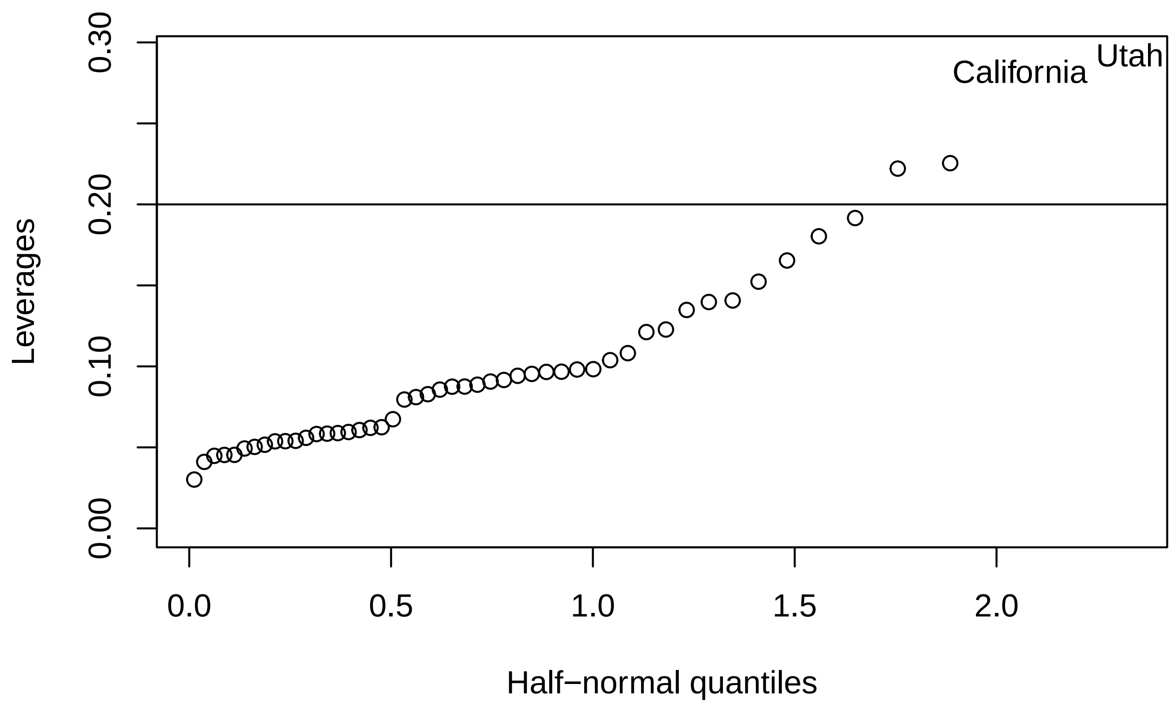




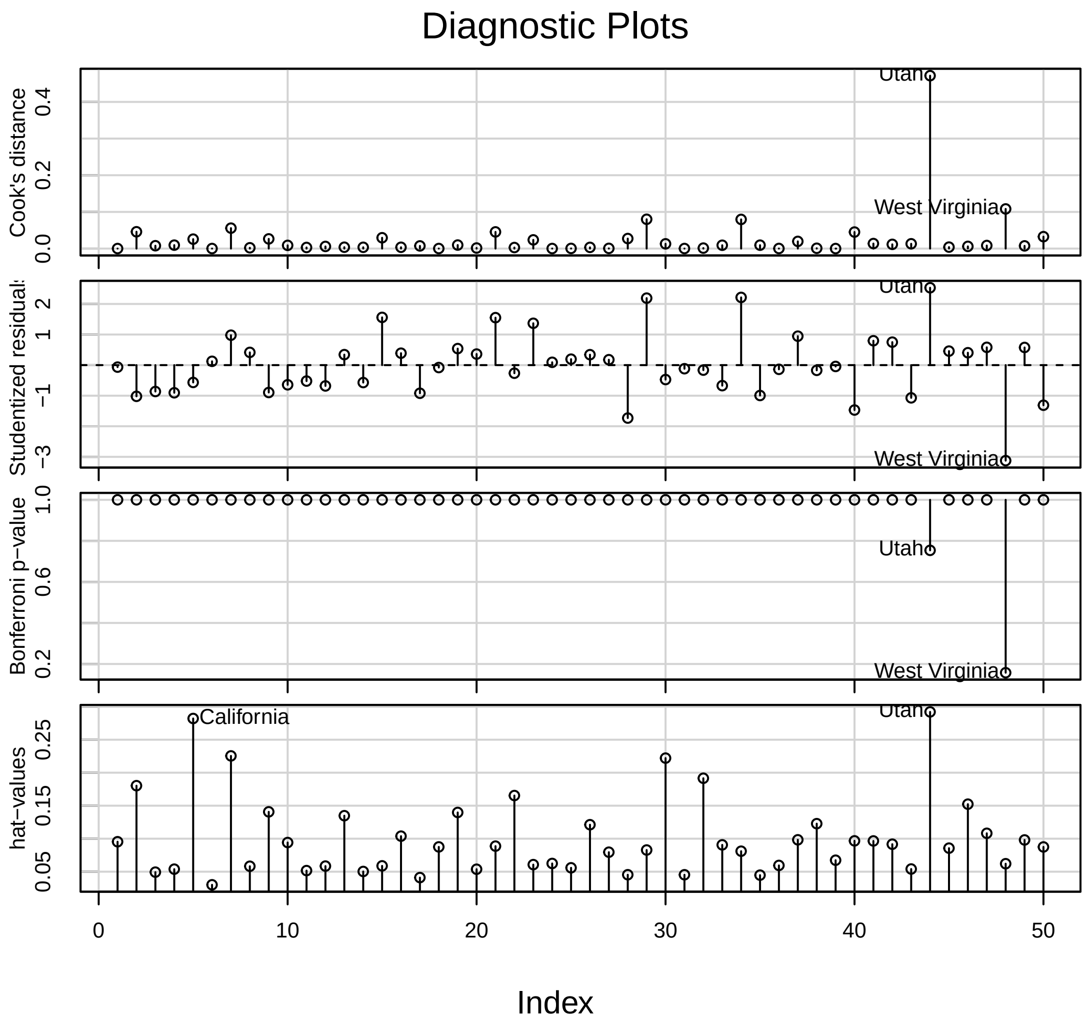
d)

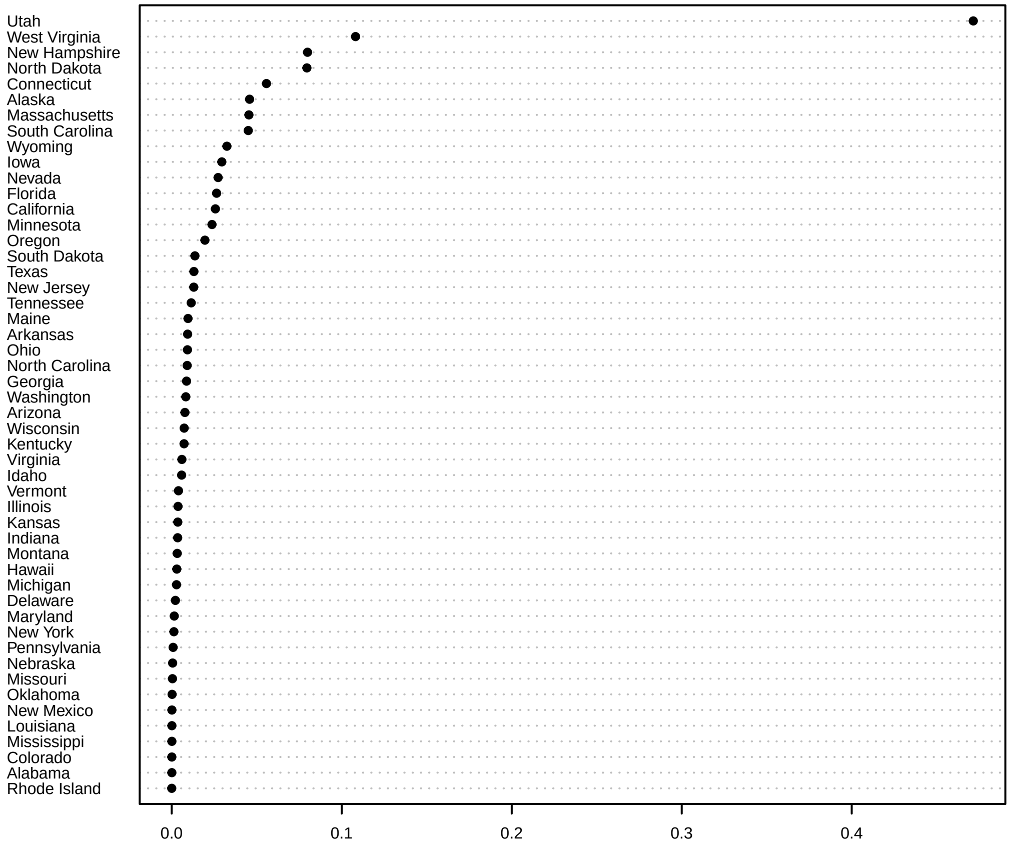
```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##                  rstudent unadjusted p-value Bonferroni p
## West Virginia -3.124428          0.0031496      0.15748
```


Standardized Residuals vs Fitted Values
Standardized Residuals
Fitted Values
Utah
New Hampshire
North Dakot
Massachusetts
Iowa
Minnesota
Connecticut
Oregon
Tennessee
South Dak
Washington
Maryland
Colorado
Montana
Kansas
Michigan
New Jersey
California
Florida
Texas
Arizona
Alaska
Idaho
Ohio
Wyoming
South Carolina
Nevada
West Virginia


California
Utah
Leverages
Half−normal quantiles

e)


Diagnostic Plots
Cook's distance
Studentized residuals
Bonferroni p−value
hat−values
Utah
West Virginia
California
Index
0
10
20
30
40
50


Utah
West Virginia
New Hampshire
North Dakota
Connecticut
Alaska
Massachusetts
South Carolina
Wyoming
Iowa
Nevada
Florida
California
Minnesota
Oregon
South Dakota
Texas
New Jersey
Tennessee
Maine
Arkansas
Ohio
North Carolina
Georgia
Washington
Arizona
Wisconsin
Kentucky
Virginia
Idaho
Vermont
Illinois
Kansas
Indiana
Montana
Hawaii
Michigan
Delaware
Maryland
New York
Pennsylvania
Nebraska
Missouri
Oklahoma
New Mexico
Louisiana
Mississippi
Colorado
Alabama
Rhode Island
0.0
0.1
0.2
0.3
0.4

LMR 6.8

a)

```
##  [1]   1.00000  17.47144  25.30482  58.60610  83.59121 100.63222 137.89717
##  [8] 175.28623 192.61449 213.00748 228.15747 268.20620 555.67072
```

```
##                 vif_x
##  [1,] "age"     "2.25045023586224"
##  [2,] "weight"  "33.50931979125"
##  [3,] "height"  "1.67459083959811"
##  [4,] "neck"    "4.32446326424943"
##  [5,] "chest"   "9.46087732137424"
##  [6,] "abdom"   "11.7670733753212"
##  [7,] "hip"     "14.796519836923"
##  [8,] "thigh"   "7.77786469219344"
##  [9,] "knee"    "4.61214673600487"
## [10,] "ankle"   "1.90796099507021"
## [11,] "biceps"  "3.61974357536549"
## [12,] "forearm" "2.1924921166644"
## [13,] "wrist"   "3.37751489619545"
```

b)

```
##  [1]   1.00000  18.39787  26.21547  61.53224  91.07633 114.44792 148.72518
##  [8] 178.80871 202.08708 211.78359 240.69468 276.35018 554.79777
```

```
##                 vif_x2
##  [1,] "age"     "2.27819073931863"
##  [2,] "weight"  "45.2988434683429"
##  [3,] "height"  "3.43958730850223"
##  [4,] "neck"    "3.97889794019821"
##  [5,] "chest"   "10.7125052438853"
##  [6,] "abdom"   "11.9675796797964"
##  [7,] "hip"     "12.1462491415389"
##  [8,] "thigh"   "7.15371104845086"
##  [9,] "knee"    "4.44175243874575"
## [10,] "ankle"   "1.81025271253279"
## [11,] "biceps"  "3.40952410870031"
## [12,] "forearm" "2.42287790935087"
## [13,] "wrist"   "3.26367710797482"
```

c)

```
##
## Call:
## lm(formula = brozek ~ age + weight + height, data = new_fat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0260  -3.6537   0.0569   3.7588  11.9011
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 54.31985    9.63347   5.639 4.69e-08 ***
## age          0.12575    0.02599   4.838 2.31e-06 ***
## weight       0.23519    0.01373  17.124  < 2e-16 ***
## height      -1.18089    0.14638  -8.067 3.17e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.986 on 246 degrees of freedom
## Multiple R-squared:  0.5838, Adjusted R-squared:  0.5787
## F-statistic:   115 on 3 and 246 DF,  p-value: < 2.2e-16
```

```
## [1]  1.00000 13.87911 25.03771
```

```
##               vif_x3
## [1,] "age"    "1.08330491018921"
## [2,] "weight" "1.3811645842945"
## [3,] "height" "1.46966479387396"
```

d)

```
##        fit      lwr      upr
## 1 18.48834 8.647863 28.32882
```

e)

```
##        fit      lwr      upr
## 1 20.18367 10.32046 30.04688
```

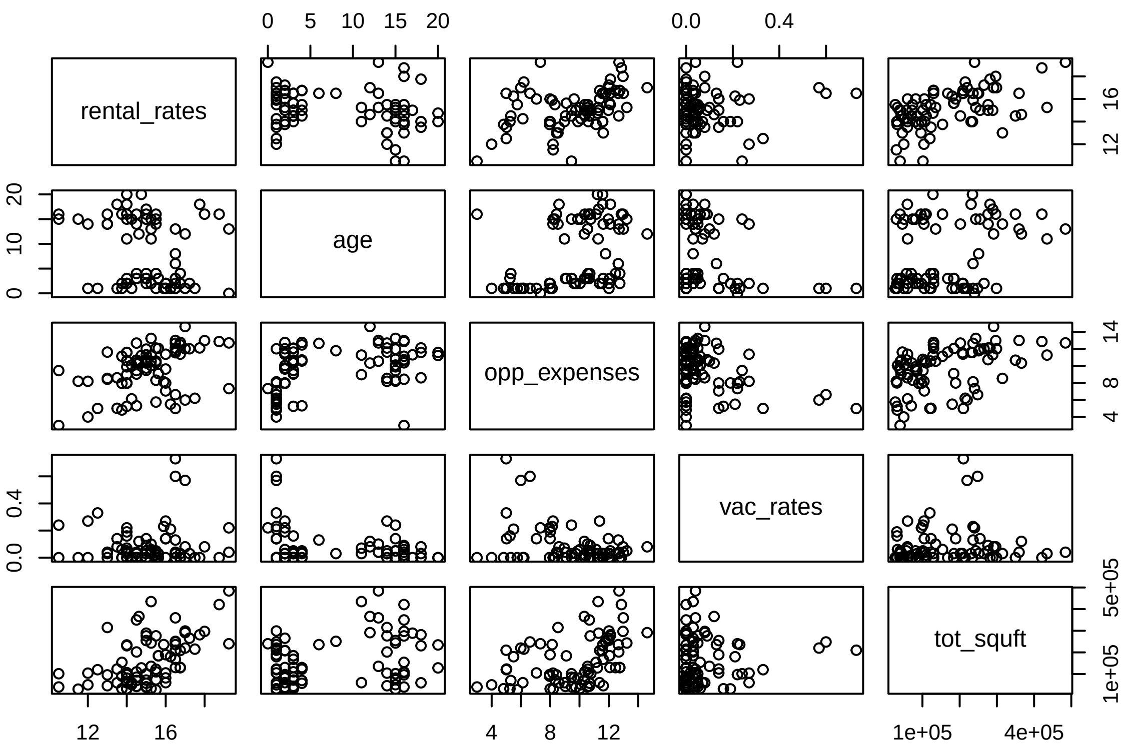
f)

```
##        fit      lwr      upr
## 1 3.720148 -6.28208 13.72238
```

Question 4

Part A

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   rental_rates = col_double(),
##   age = col_double(),
##   opp_expenses = col_double(),
##   vac_rates = col_double(),
##   tot_squft = col_double()
## )
```



The only pair of variables that show a relationship is between tot_sqft and rental_rates which displays a slight linear relationship. The other pairs do not show strong relationship, linear or otherwise.

Part B

```
##              rental_rates        age opp_expenses   vac_rates  tot_squft
## rental_rates   1.00000000 -0.2502846    0.4137872  0.06652647 0.53526237
## age           -0.25028456  1.0000000    0.3888264 -0.25266347 0.28858350
## opp_expenses   0.41378716  0.3888264    1.0000000 -0.37976174 0.44069713
## vac_rates      0.06652647 -0.2526635   -0.3797617  1.00000000 0.08061073
## tot_squft      0.53526237  0.2885835    0.4406971  0.08061073 1.00000000
```

The strongest positive correlation is the one between tot_sqft and rental_rates. The strongest negative is the one between opp_expenses and vac_rates. In both of those cases, the relationship was moderate. Age showed a moderate negative ralation with rental_rates and vac_rates. It also had a moderate positive relationship with tot_squft and opp_expenses. Overall, the only two pairs that did not indicate a significant relationship is vac_rates/rental_rates and vac_rates/tot_squft.

Part C

```
##
## Call:
## lm(formula = rental_rates ~ ., data = df_commercial)
##
## 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 ***
## age          -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
## opp_expenses  2.820e-01  6.317e-02   4.464 2.75e-05 ***
## vac_rates     6.193e-01  1.087e+00   0.570     0.57
## tot_squft     7.924e-06  1.385e-06   5.722 1.98e-07 ***
## ---
## 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
```

```
##   (Intercept)           age  opp_expenses     vac_rates     tot_squft
##  1.220059e+01 -1.420336e-01  2.820165e-01  6.193435e-01  7.924302e-06
```

c. part a) through c)
y = 12.2 - 0.142 x_1 + 0.282 x_2 + 0.620 x_3 + 0.000000792 x_4
where:
x_1 -> age of dwelling (age)
x_2 -> operating expenses and taxes (opp_expenses)
x_3 -> vacancy rates (vac_rates)
x_4 -> total square footage (tot_squft)

The predictor with seemingly the smallest effect on rental rates is total square footage as can be seen by the near zero value, but it is still statistically significant. The small coefficient value is offset by the large values of the predictor itself. Vacancy rates were the least statistically significant predictor as can be seen with the low p value. If this predictor is dropped from the model, the model prediction will still produce a reasonably accurate result. Age and operating expenses were both statistically significant to the model and produced a large effect on the resulting prediction.

```
## [1] "The R squared values is 0.585"
```

```
## [1] "The adjusted R squared values is 0.563"
```

c. part d) The null hypothesis, H_0, states that all 4 coefficients are equal to zero. The alternate hypothesis, H_A states that at least one of the coefficients is not zero. The test statistic is given as a the F-statistic in the model summary which is equal to 26.76. This results in a p value of nearly zero using an f distribution of degrees of freedom of 4 and 76. Since the p value is less than the 0.05 significance level, we have enough evidence to reject the null hypothesis in favor of the alternative hypothesis.

c. part e)

```
## Generalized least squares fit by REML
##   Model: rental_rates ~ .
##   Data: df_commercial
##        AIC     BIC    logLik
##   293.2916 307.276 -140.6458
##
## Coefficients:
##                  Value Std.Error   t-value p-value
## (Intercept)  12.200586 0.5779562 21.109881  0.0000
## age          -0.142034 0.0213426 -6.654933  0.0000
## opp_expenses  0.282017 0.0631723  4.464240  0.0000
## vac_rates     0.619344 1.0868128  0.569871  0.5704
## tot_squft     0.000008 0.0000014  5.722446  0.0000
##
##  Correlation:
##              (Intr) age    opp_xp vc_rts
## age          -0.032
## opp_expenses -0.889 -0.201
## vac_rates    -0.516  0.175  0.412
## tot_squft     0.200 -0.189 -0.454 -0.322
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.8034364 -0.5198884 -0.0800025  0.4907637  2.5896591
##
## Residual standard error: 1.136885
## Degrees of freedom: 81 total; 76 residual
```

c. part f)

The test statistic for β_3 is 0.570 which gives a p value of 0.5704. If H_0 states that $\beta_3 = 0$ and H_A states that β_3 is not equal to zero, then in this case we do no have enough evidence to reject the null hypothesis. This means that there is not enough evidence to reject the possibility that β_3 is zero. We can then say that β is not statistically significant to the model.

c. part h)

```
##
## Call:
## lm(formula = rental_rates ~ . - vac_rates, data = df_commercial)
##
## 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 ***
## age          -1.442e-01  2.092e-02  -6.891 1.33e-09 ***
## opp_expenses  2.672e-01  5.729e-02   4.663 1.29e-05 ***
```

```
## tot_squft     8.178e-06  1.305e-06   6.265 1.97e-08 ***
## ---
## 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
```

Part D

d. part I)

Estimate	Error	t value	Pr(>|t|)	CI90neg	CI90pos
12.3705818	0.4928469	25.100251	0.00e+00	11.5500486	13.1911150
-0.1441646	0.0209201	-6.891195	0.00e+00	-0.1789942	-0.1093351
0.2671670	0.0572949	4.663018	1.29e-05	0.1717777	0.3625564
0.0000082	0.0000013	6.265018	0.00e+00	0.0000060	0.0000104

d. part II

```
##   (Intercept)           age  opp_expenses     tot_squft
##  1.237058e+01 -1.441646e-01  2.671670e-01  8.178210e-06
```

d. part III

```
## # A tibble: 4 x 5
##     age opp_expenses vac_rates tot_squft pred_rrates
##   <dbl>        <dbl>     <dbl>     <dbl>       <dbl>
## 1     5         8.25         0    250000        15.9
## 2     6         8.5          0    270000        16.0
## 3    14        11.5          0    300000        15.9
## 4    12        10.2          0    310000        15.9
```

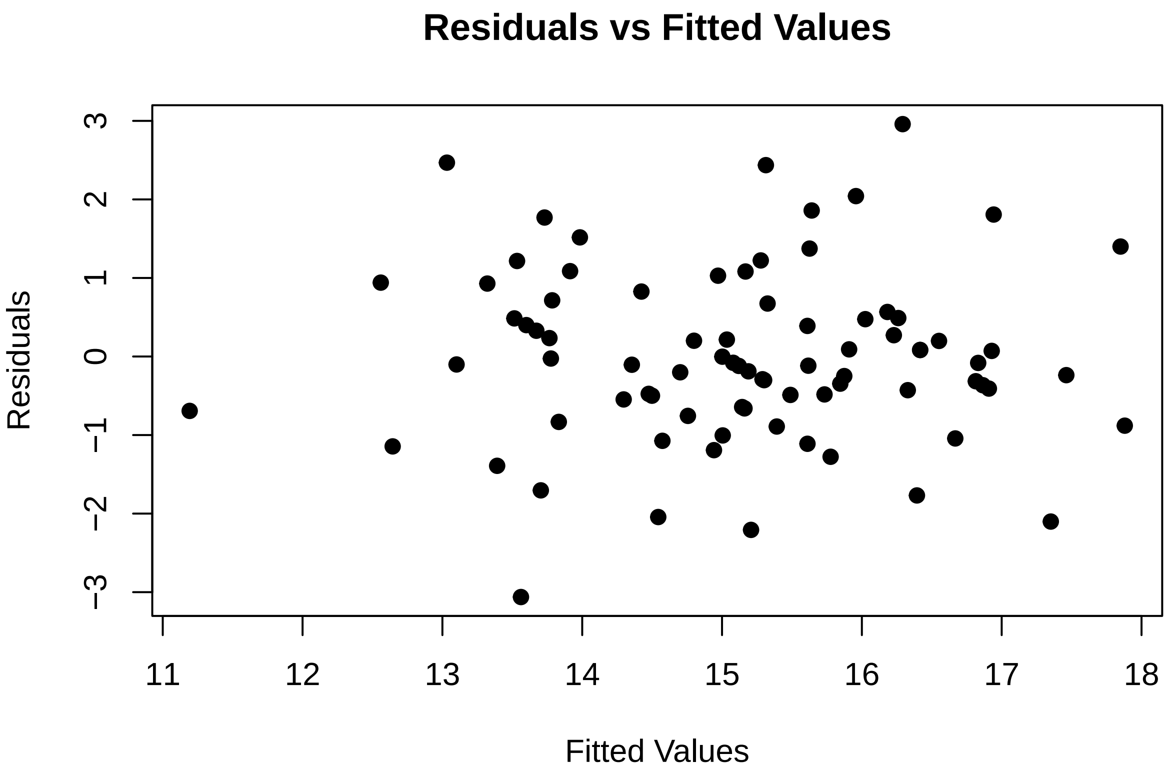
d. part IV

```
## # A tibble: 3 x 5
##     age opp_expenses vac_rates tot_squft pred_rrates
##   <dbl>        <dbl>     <dbl>     <dbl>       <dbl>
## 1     4           10         0     80000        15.1
## 2     6           11.5       0    120000        15.6
## 3    12           12.5       0    340000        16.8
```

```
##        fit      lwr      upr
## 1 15.11985 12.83659 17.40311
## 2 15.55940 13.27329 17.84551
## 3 16.76079 14.45322 19.06835
```

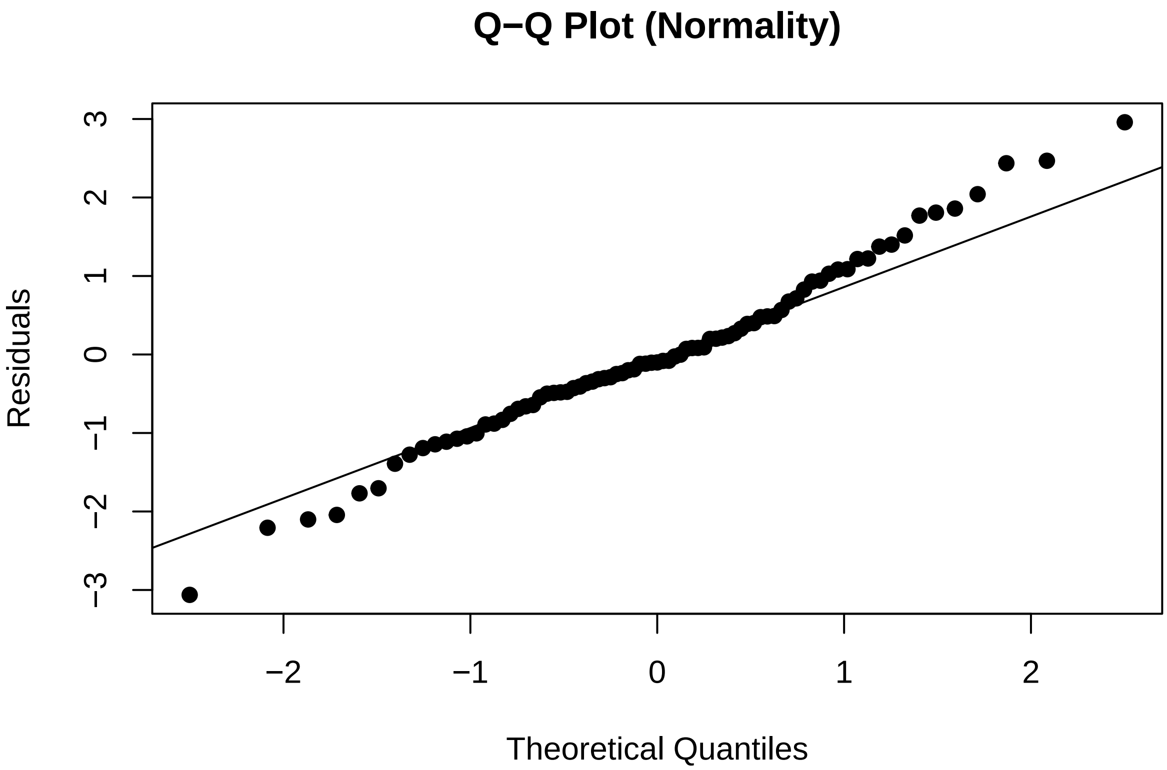
Part E

e. part a)

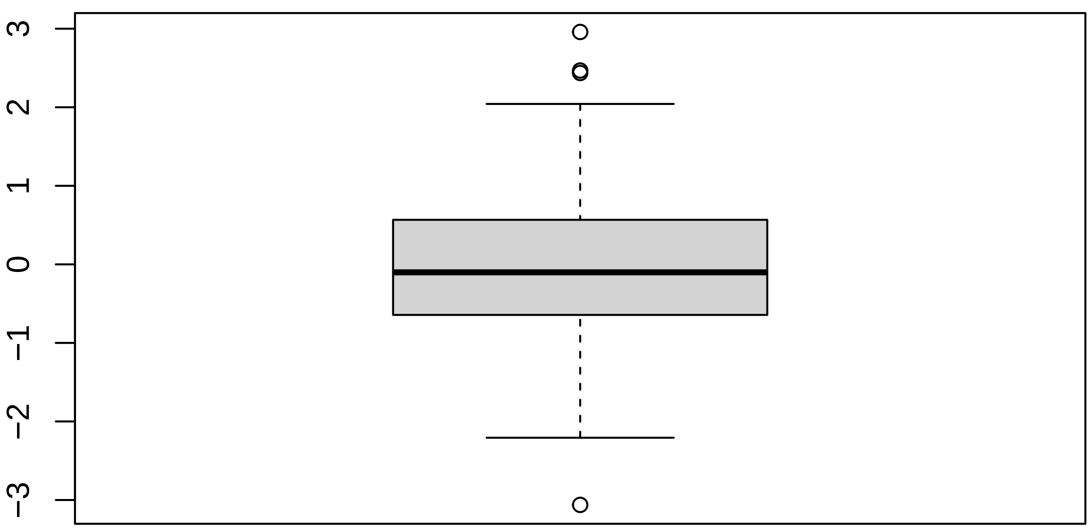


The plot indicates that variance is constant and that the data is normal. A deeper investigation is needed in order to confirm the strength of these assessments.

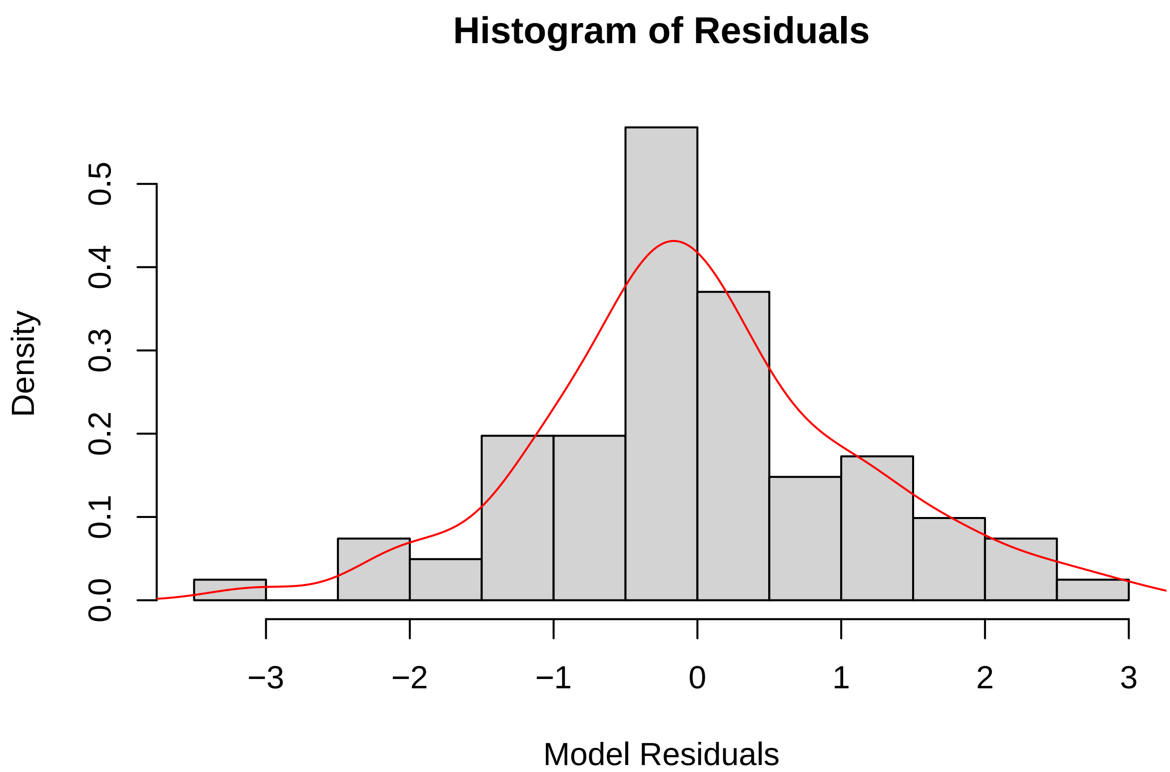
e. part b)



The QQ plot shows some deviation from normality. This means that the data was not strictly following a normal distribution.



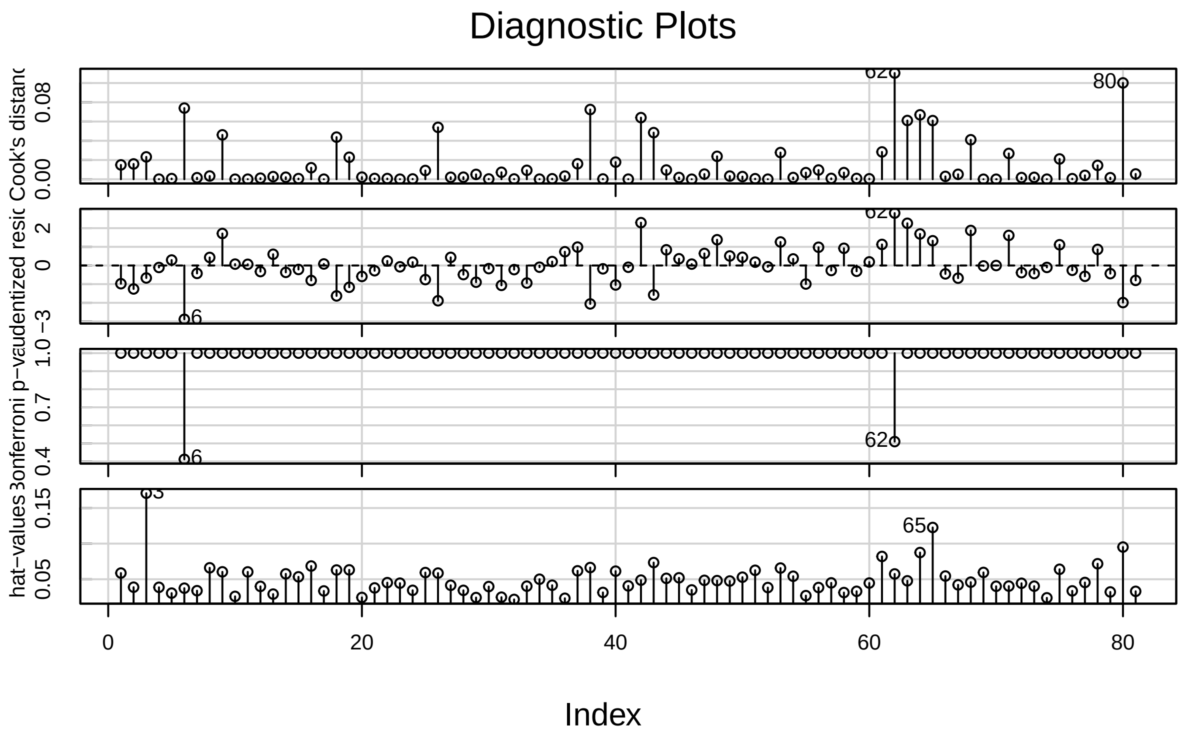
The boxplot shows a number of outliers at both ends of the distribution. These should be investigated more using specific outlier testing methods.



The histogram shows more evidence for outliers and even some skew in the data.

e. part c)

```
##
##  Breusch-Pagan test
##
## data:  comm_model2
## BP = 17.281, df = 3, p-value = 0.0006187
```



e part d and g)

The diagnostic plots do reveal a few outliers, high leverage, and influential points. Data points 6 and 62 are outliers, while 3, 62, 65, and 80 indicate high leverage/influence. 3, 6, 62, and 80 can be dropped from the model. 65 can be left in the model since it is not too much higher in leverage than the other points.

e part e and f)

```
##
##  Shapiro-Wilk normality test
##
## data:  comm_model2$residuals
## W = 0.98776, p-value = 0.6406
```

```
##
##  Durbin-Watson test
##
## data:  comm_model2
## DW = 1.5867, p-value = 0.02463
## alternative hypothesis: true autocorrelation is greater than 0
```

The high p value calculated from the Shapiro-Wilk test, means that we do not have enough evidence to reject the null hypothesis (H_0 : the data is normal). This means our assumption of the normality of the data is held true. The Durbin-Watson test indicates that the model errors are correlated (H_0: errors are not correlated, H_A: They are correlated). The small p value (< 0.05) means that we have enough evidence to reject the null hypothesis.

Part F

```
##
## Call:
## lm(formula = rental_rates ~ . - vac_rates, data = df_commercial[-c(3,
```

```
##     6, 62, 80), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.35002 -0.62221 -0.04052  0.62566  2.24922
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.239e+01  4.716e-01  26.261  < 2e-16 ***
## age          -1.239e-01  1.987e-02  -6.236 2.62e-08 ***
## opp_expenses  2.571e-01  5.487e-02   4.685 1.27e-05 ***
## tot_squft     7.957e-06  1.236e-06   6.438 1.12e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.01 on 73 degrees of freedom
## Multiple R-squared:  0.5778, Adjusted R-squared:  0.5605
## F-statistic: 33.31 on 3 and 73 DF,  p-value: 1.13e-13
```

By dropping outliers, R^2 values did not improve in the new model. They actually dropped slightly, indicating that outliers might not have had a strong effect on the model after all.

```
##
## Call:
## lm(formula = rental_rates ~ . - vac_rates, data = df_commercial[-c(6,
##     62, 80), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.29323 -0.62754 -0.04509  0.58356  2.32591
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.224e+01  4.432e-01  27.626  < 2e-16 ***
## age          -1.291e-01  1.899e-02  -6.796 2.36e-09 ***
## opp_expenses  2.739e-01  5.148e-02   5.320 1.06e-06 ***
## tot_squft     8.003e-06  1.233e-06   6.491 8.62e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.008 on 74 degrees of freedom
## Multiple R-squared:  0.6203, Adjusted R-squared:  0.6049
## F-statistic:  40.3 on 3 and 74 DF,  p-value: 1.515e-15
```

Leaving data point 3 in the model improved the R^2 values a little.

```
##
## Call:
## lm(formula = rental_rates ~ . - vac_rates, data = df_commercial[-c(6,
##     62), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -2.15356 -0.66614 -0.08314  0.63095  2.34999
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.226e+01  4.521e-01  27.121  < 2e-16 ***
## age          -1.283e-01  1.937e-02  -6.625 4.65e-09 ***
## opp_expenses  2.804e-01  5.241e-02   5.350 9.22e-07 ***
## tot_squft     7.296e-06  1.206e-06   6.050 5.26e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.029 on 75 degrees of freedom
## Multiple R-squared:  0.5995, Adjusted R-squared:  0.5834
## F-statistic: 37.42 on 3 and 75 DF,  p-value: 6.844e-15
```

Removing outliers and ignoring high leverage also improved the model slightly as can be seen by the higher R^2 values.

Part G

```
##        fit      lwr      upr
## 1 15.75576 14.01053 17.50099
```

5A

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Rep = col_double(),
##   Software = col_double(),
##   SalesLastQuarter = col_double(),
##   SalesThisQuarter = col_double()
## )
```

```
##
## Call:
## lm(formula = SalesThisQuarter ~ . - SalesLastQuarter - Rep, data = df_software)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.583  -6.833   1.417   7.583  32.417
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   81.583      3.281  24.866   <2e-16 ***
## Software2     -2.000      4.640  -0.431    0.669
## Software3     -7.667      4.640  -1.652    0.108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.37 on 33 degrees of freedom
## Multiple R-squared:  0.08176,    Adjusted R-squared:  0.02611
## F-statistic: 1.469 on 2 and 33 DF,  p-value: 0.2448
```

Part I

y = 81.6 - 2.00 x_1 - 7.67 x_2
where:
x_1 -> Whether Software 2 was used
x_2 -> Whether Software 3 was used

$E(y | x_1 = 1, x_2 = 0) = (81.6 - 2.00) - 7.67 * 0 = 79.42 $
$E(y | x_1 = 0, x_2 = 1) = (81.6 - 7.67) - 2.00 * 0 = 73.93 $

Part II

```
## Analysis of Variance Table
##
## Response: SalesThisQuarter
##           Df Sum Sq Mean Sq F value Pr(>F)
## Software   2  379.6  189.78  1.4692 0.2448
## Residuals 33 4262.8  129.17
```

```
## [1] "Variance explained by the software is 0.082"
```

Part III

H_0 states that software has no effect on sales.
H_A states that software does have an effect on sales.

The *f statistic* given from the anova table, 1.47, and the corresponding *p value*, 0.25, indicate that we do not have enough evidence to reject the null hypothesis at the 0.05 significance level.

5B

```
##
## Call:
## lm(formula = SalesThisQuarter ~ . - Rep, data = df_software)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.974  -4.194   0.554   3.536  16.049
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -36.4423    16.4223  -2.219   0.0337 *
## Software2          0.7535     2.9249   0.258   0.7984
## Software3         -1.2835     3.0311  -0.423   0.6748
## SalesLastQuarter   1.5019     0.2073   7.244 3.14e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.104 on 32 degrees of freedom
## Multiple R-squared:  0.6521, Adjusted R-squared:  0.6195
## F-statistic:    20 on 3 and 32 DF,  p-value: 1.741e-07
```

Part I

y = -36.44 + 0.754 x_1 - 1.284 x_2 + 1.502 x_3
where:
x_1 -> Whether Software 2 was used
x_2 -> Whether Software 3 was used
x_3 -> Sales from last quarter

$E(y|x_1 = 1, x_2 = 0) = (-36.44 + 0.754) - 1.284 * 0 + 1.502x_3 = -35.69 + 1.502x_3$
$E(y|x_1 = 0, x_2 = 1) = (-36.44 - 1.284) + 0.754 * 0 + 1.502x_3 = -37.72 + 1.502x_3$

Part II - V

H_0 states that software has no effect on sales.
H_A states that software does have an effect on sales.

The *f statistic* given from the anova table, 3.76, and the corresponding *p value*, 0.0341, indicates that we do have enough evidence to reject the null hypothesis at the 0.05 significance level.

Part VI

```
## Analysis of Variance Table
```

```
##
## Response: SalesThisQuarter
##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## Software           2  379.56  189.78  3.7606   0.03413 *
## SalesLastQuarter  1 2647.88 2647.88 52.4699 3.141e-08 ***
## Residuals        32 1614.87   50.46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "Variance explained by the software is 0.082"
```

5C

```
##
## Call:
## lm(formula = SalesThisQuarter ~ SalesLastQuarter:Software + SalesLastQuarter,
##     data = df_software)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.5909  -4.2560   0.7545   3.1806  16.4523
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -3.622e+01  1.596e+01  -2.270   0.0301 *
## SalesLastQuarter           1.503e+00  2.034e-01   7.390 2.09e-08 ***
## SalesLastQuarter:Software2  1.088e-04  3.767e-02   0.003   0.9977
## SalesLastQuarter:Software3 -2.128e-02  3.966e-02  -0.536   0.5954
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.115 on 32 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6183
## F-statistic:  19.9 on 3 and 32 DF,  p-value: 1.829e-07
```

Part I - III

y = -36.22 + 1.504 x_1 + 1.504 x_2 + 1.482 x_3
where:
x_1 -> Interaction term between sales last quarter and software 1
x_2 -> Interaction term between sales last quarter and software 2
x_3 -> Interaction term between sales last quarter and software 4

Part i)

$E(y | x_1 = 1, x_2 = 0, x_3 = 0) = (-36.22 + 1.504) - 7.67 * 0 = -34.72 $
$E(y | x_1 = 0, x_2 = 1, x_3 = 0) = (-36.22 + 1.504) - 2.00 * 0 = -34.72 $
$E(y|x_1 = 0, x_2 = 0, x_3 = 1) = (-36.22 + 1.482) - 2.00 * 0 = -34.74$

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
35	4642.306	NA	NA	NA	NA
33	4262.750	2	379.5556	1.469161	0.2447802

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Software	2	379.5556	189.7778	1.469161	0.2447802
Residuals	33	4262.7500	129.1742	NA	NA

df_vector	SS	MS	F_stat	p_value
2	379.5556	189.7778	1.469161	0.2447802
33	4262.7500	129.1742	NA	NA
35	4642.3056	NA	NA	NA

6

part a

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Drug = col_double(),
##   Momweight = col_double(),
##   Dadweight = col_double(),
##   Pigweight = col_double()
## )
```

```
##
## Call:
## lm(formula = Pigweight ~ ., data = df_pigs)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.905 -1.174  0.187  1.351  3.657
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.48163    9.14917   0.818  0.41628
## Drug2       -1.60557    0.52788  -3.042  0.00331 **
## Drug3       -0.70480    0.52871  -1.333  0.18684
## Momweight    0.26363    0.04727   5.578 4.28e-07 ***
## Dadweight    0.17442    0.03465   5.034 3.58e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.855 on 70 degrees of freedom
## Multiple R-squared:  0.4561, Adjusted R-squared:  0.425
## F-statistic: 14.67 on 4 and 70 DF,  p-value: 9.393e-09
```

```
## (Intercept)
##    5.876066
```

```
## (Intercept)
##    6.776831
```

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

x_1 -> Whether drug 2 was administered
x_2 -> Whether drug 3 was administered
x_3 -> Mom weight
x_4 -> Dad weight

part b

Drug	$E(y\|x)$
1	7.48 - 0.264 x_3 + 0.174 x_4
2	5.88 - 0.264 x_3 + 0.174 x_4
3	6.78 - 0.264 x_3 + 0.174 x_4

part c

```
##        fit      lwr      upr
## 1 75.05263 71.15725 78.94801
```

part d

The 95% confidence interval is

$$\Delta_{drug2,drug3} \pm t_{criticalvalue} * SE_{drug2+drug3}$$

which calculates to be -2.6627036, 0.8611729.

part e. i)

```
## Analysis of Variance Table
##
## Model 1: Pigweight ~ Momweight + Dadweight
## Model 2: Pigweight ~ Drug + Momweight + Dadweight
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     72 272.70
## 2     70 240.81  2    31.894 4.6356 0.01286 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0 : \beta_0 = \beta_1 = \beta_2 = 0$
$H_A : \beta_0 \neq 0$ or $\beta_1 \neq 0$ or $\beta_2 \neq 0$

In this case, the null hypothesis can be rejected at the 0.05 significance level according to the ANOVA table. The type of drug does have an effect on the final pig weight.

part e. ii)

part e. iii)

part e. iv)

Code

```
library(faraway)
library(car)
library(MASS)
library(lmtest)
library(tidyverse)
library(nlme)
pros_model1 <- lm(lpsa ~ ., data = prostate)
summary(pros_model1)
new_patient <- tibble(lcavol = 1.44692,  lweight = 3.62301,  age = 65.00000,
                      lbph = 0.30010,  svi = 0.00000,  lcp = -0.79851,
                      gleason = 7.0000,   pgg45 = 15.0000)

predict(pros_model1, newdata = new_patient, interval = "confidence")
predict(pros_model1, newdata = new_patient, interval = "prediction")
mean(prostate$age)
new_patient1 <- tibble(lcavol = 1.44692,  lweight = 3.62301,  age = 25.00000,
                      lbph = 0.30010,  svi = 0.00000,  lcp = -0.79851,
                      gleason = 7.0000,   pgg45 = 15.0000)

predict(pros_model1, newdata = new_patient1, interval = "confidence")
predict(pros_model1, newdata = new_patient1, interval = "prediction")
pros_model2 <- lm(lpsa ~ lcavol + lweight + svi, data = prostate)
summary(pros_model2)
predict(pros_model2, newdata = new_patient1, interval = "confidence")
predict(pros_model2, newdata = new_patient1, interval = "prediction")
print(anova(pros_model2, pros_model1))
sat_model1 <- lm(total ~ expend + salary + ratio + takers, data = sat)
summary(sat_model1)
plot(x = sat_model1$fitted.values, y = sat_model1$residuals, pch = 19,
     main = "Residuals vs Fitted Values", xlab = "Fitted Values", ylab = "Residuals")
res_line <- loess(sat_model1$residuals ~ sat_model1$fitted.values)
j <- order(sat_model1$fitted.values)
lines(sat_model1$fitted.values[j], res_line$fitted[j], col = "red", lwd = 3)
abline(h = 0)
bptest(sat_model1, studentize = FALSE)
ncvTest(sat_model1)
qqnorm(sat_model1$residuals, ylab = "Residuals", main = "Q-Q Plot (Normality)", pch = 19)
qqline(sat_model1$residuals)
shapiro.test(sat_model1$residuals)
dwtest(sat_model1)
X <- sat %>% select(expend, salary, ratio, takers)
X <- cbind(rep(1, nrow(X)), X)

X <- data.matrix(X)

#print(X)
Xt <- t(X)

XtX_inv <- solve(Xt %*% X)

XtY <- Xt %*% sat$total
```

```
beta_hat <- XtX_inv %*% XtY

P <- X %*% XtX_inv %*% Xt

res_mean <- mean(sat_model1$residuals)
res_sd <- sd(sat_model1$residuals)
stan_res <- (sat_model1$residuals - res_mean) / res_sd

plot(x = diag(P),
     y = sat_model1$residuals,
     pch = 19,
     main = "Residuals vs Leverage",
     xlab = "Leverage",
     ylab = "Residuals",
     xlim = c(0, 0.35))
text(diag(P), sat_model1$residuals, rownames(P),
     cex = 0.6, pos = 4, col = "red")
hat_data <- cbind(1:50, diag(P))

plot(hat_data, pch = 19, xlab = "", ylab = "hii values")
abline(h = 2 * 5 / 50)
text(hat_data[,1], hat_data[,2], names(diag(P)),
     cex = 0.5, pos = 4, col = "red")
dotchart(sort(diag(P)), pch = 19, cex = 0.5)
outlierTest(sat_model1)

plot(x = sat_model1$fitted.values, y = rstandard(sat_model1), pch = 19,
     main = "Standardized Residuals vs Fitted Values", xlab = "Fitted Values", ylab = "Standardized Resi
abline(h = c(0, 2, -2), col = "blue", lty = 4, lwd = 2)
text(x = sat_model1$fitted.values, y = rstandard(sat_model1), rownames(sat),
     cex = 0.5, pos = 4, col = "red")
#cooks.distance(sat_model1)

halfnorm(diag(P), labs = rownames(P), ylab = "Leverages")
abline(h = 2 * sat_model1$rank / nrow(sat))

SST <- sum((sat$total - mean(sat$total))^2)
SSR <- sum((sat_model1$fitted.values - mean(sat$total))^2)
SSE <- sum((sat$total - sat_model1$fitted.values)^2)
MSR <- SSR/(sat_model1$rank - 1)
MSE <- SSE/(nrow(sat) - sat_model1$rank)
influenceIndexPlot(sat_model1)
cooks_dis_calc <- sat_model1$residuals^2 /
  (sat_model1$rank * MSE) * (diag(P)/(1-diag(P))^2)

dotchart(sort(cooks_dis_calc), cex = 0.5, pch = 19)
fat_model1 <- lm(brozek ~ age + weight + height + neck + chest +
                   abdom + hip + thigh + knee + ankle + biceps +
                   forearm + wrist, data = fat)

X <- model.matrix(fat_model1)[,-1]
eigen_X <- eigen(t(X) %*% X)
```

```r
sqrt(eigen_X$values[1]/eigen_X$values)

vif_x <- rep(0, ncol(X))


for(i in 1:ncol(X)){
    vif_x[i] <- 1 / (1 - summary(lm(X[,i] ~ X[,-i]))$r.squared)
}

vif_x <- cbind(colnames(X), vif_x)
print(vif_x)
new_fat <- fat[-c(39,42),]

fat_model2 <- lm(brozek ~ age + weight + height + neck + chest +
                   abdom + hip + thigh + knee + ankle + biceps +
                   forearm + wrist, data = new_fat)

X2 <- model.matrix(fat_model2)[,-1]
eigen_X2 <- eigen(t(X2) %*% X2)

sqrt(eigen_X2$values[1]/eigen_X2$values)

vif_x2 <- rep(0, ncol(X2))


for(i in 1:ncol(X2)){
    vif_x2[i] <- 1 / (1 - summary(lm(X2[,i] ~ X2[,-i]))$r.squared)
}

vif_x2 <- cbind(colnames(X2), vif_x2)
print(vif_x2)

fat_model3 <- lm(brozek ~ age + weight + height, data = new_fat)
summary(fat_model3)

X3 <- model.matrix(fat_model3)[,-1]
eigen_X3 <- eigen(t(X3) %*% X3)

sqrt(eigen_X3$values[1]/eigen_X3$values)

vif_x3 <- rep(0, ncol(X3))


for(i in 1:ncol(X3)){
    vif_x3[i] <- 1 / (1 - summary(lm(X3[,i] ~ X3[,-i]))$r.squared)
}

vif_x3 <- cbind(colnames(X3), vif_x3)
print(vif_x3)


fat_1 <- as.data.frame(t(apply(X3, 2, median)))
```

```r
predict(fat_model3, newdata = fat_1, interval = "prediction")


fat_2 <- as.data.frame(cbind(age = 40, weight = 200, height = 73))

predict(fat_model3, newdata = fat_2, interval = "prediction")

fat_3 <- as.data.frame(cbind(age = 40, weight = 130, height = 73))

predict(fat_model3, newdata = fat_3, interval = "prediction")
df_commercial <- read_table("commercial_property.txt")
#a)
options(repr.plot.width = 8, repr.plot.height = 6, repr.plot.res = 150)
df_commercial %>% pairs()
#b)
cor(df_commercial)
#c)
comm_model1 <- lm(rental_rates ~ ., data = df_commercial)
summary(comm_model1)
comm_model1$coefficients
sprintf("The R squared values is %.3f", summary(comm_model1)$r.squared)
sprintf("The adjusted R squared values is %.3f", summary(comm_model1)$adj.r.squared)
#c. part e)

summary(gls(rental_rates ~ ., data = df_commercial))
# c. part h)

comm_model2 <- lm(rental_rates ~ .-vac_rates, data = df_commercial)
summary(comm_model2)
#d. part I)
model_coeff <- summary(comm_model2)$coefficients

model_coeff <- as_tibble(model_coeff)

colnames(model_coeff)[2] = 'Error'

model_coeff <- model_coeff %>% mutate(CI90neg = Estimate - qt(0.95, 77) * Error, CI90pos = Estimate + q

knitr::kable(model_coeff)
#d. part II)

pvals <- summary(comm_model2)$coef[,4]
padj <- p.adjust(pvals, method="bonferroni")
print(coef(comm_model2)[padj < 0.1])
#d. part III)
new_dwellings <- tibble(age = c(5.0, 6.0, 14.0, 12.0),
                        opp_expenses = c(8.25, 8.50, 11.50, 10.25),
                        vac_rates = rep(0, 4),
                        tot_squft = c(250000, 270000, 300000, 310000))

new_dwellings <- new_dwellings %>% mutate(pred_rrates = predict(comm_model2, newdata = new_dwellings))
print(new_dwellings)
#d part IV)
```

```r
new_dwellings2 <- tibble(age = c(4.0, 6.0, 12.0),
                         opp_expenses = c(10.0, 11.50, 12.5),
                         vac_rates = rep(0, 3),
                         tot_squft = c(80000, 120000, 340000))
new_dwellings2 <- new_dwellings2 %>% mutate(pred_rrates = predict(comm_model2, newdata = new_dwellings2
print(new_dwellings2)
predict(comm_model2, newdata = new_dwellings2, interval = "prediction", level = 0.95)

plot(x = comm_model2$fitted.values, y = comm_model2$residuals, pch = 19,
     main = "Residuals vs Fitted Values", xlab = "Fitted Values", ylab = "Residuals")

# e part b)
qqnorm(comm_model2$residuals, ylab = "Residuals", main = "Q-Q Plot (Normality)", pch = 19)
qqline(comm_model2$residuals)
boxplot(comm_model2$residuals)
hist(comm_model2$residuals, breaks = 20, xlab = "Model Residuals", main = "Histogram of Residuals", prob
lines(density(comm_model2$residuals), col = "red")
#e part c)
bptest(comm_model2, studentize = FALSE)
#e part d and g
influenceIndexPlot(comm_model2)
#e part e and f)

shapiro.test(comm_model2$residuals)
dwtest(comm_model2)
# Part F
comm_model3 <- lm(rental_rates ~ . - vac_rates, df_commercial[-c(3, 6, 62, 80), ])

summary(comm_model3)
comm_model4 <- lm(rental_rates ~ . - vac_rates, df_commercial[-c(6, 62, 80), ])

summary(comm_model4)
comm_model5 <- lm(rental_rates ~ . - vac_rates, df_commercial[-c(6, 62), ])

summary(comm_model5)
# G
single_unit <- tibble(age = 5, opp_expenses = 8.25, vac_rates = 0, tot_squft = 250000)

predict(comm_model5, newdata = single_unit, level = 0.90, interval = "prediction")
df_software <- read_table("software_sales.txt")
df_software$Software <- as.factor(df_software$Software)

#5A
sw_model1 <- lm(SalesThisQuarter ~ . - SalesLastQuarter - Rep, data = df_software)
summary(sw_model1)
# Part ii)
anova(sw_model1)

sprintf("Variance explained by the software is %0.3f", (anova(sw_model1)[1,2] / sum(anova(sw_model1)[,2
#5B
sw_model2 <- lm(SalesThisQuarter ~ . -Rep, data = df_software)
summary(sw_model2)
anova(sw_model2)
```

```r
sprintf("Variance explained by the software is %0.3f", (anova(sw_model2)[1,2] / sum(anova(sw_model2)[,2
sw_model3 <- lm(SalesThisQuarter ~ SalesLastQuarter:Software + SalesLastQuarter, data = df_software)
summary(sw_model3)
null_model1 <- lm(SalesThisQuarter ~ NULL, data = df_software)
knitr::kable(anova(null_model1, sw_model1))
knitr::kable(anova(sw_model1))

SST <- sum((df_software$SalesThisQuarter - mean(df_software$SalesThisQuarter))^2)
SSR <- sum((sw_model1$fitted.values - mean(df_software$SalesThisQuarter))^2)
SSE <- sum((df_software$SalesThisQuarter - sw_model1$fitted.values)^2)
MSR <- SSR/(sw_model1$rank - 1)
MSE <- SSE/(nrow(df_software) - sw_model1$rank)
df_vector <- c(sw_model1$rank - 1, nrow(df_software) - sw_model1$rank, nrow(df_software) - 1)
SS <- c(SSR, SSE, SST)

anova_table <- as_tibble(cbind(df_vector, SS))

anova_table <- anova_table %>% mutate(MS = SS/df_vector)

anova_table[3,3] <- NA

F_stat <- c(as.numeric(anova_table[1,3] / anova_table[2,3]), NA, NA)

p_value <- c(1 - (pf(F_stat[1], (sw_model1$rank - 1) , (nrow(df_software) - sw_model1$rank))), NA, NA)

anova_table <- cbind(anova_table, F_stat, p_value)

knitr::kable(anova_table)
df_pigs <- read_table("pig_weight.txt")
df_pigs$Drug <- as.factor(df_pigs$Drug)
pig_model1 <- lm(Pigweight ~ ., data = df_pigs)
summary(pig_model1)
pig_model1$coefficients[1] + pig_model1$coefficients[2]
pig_model1$coefficients[1] + pig_model1$coefficients[3]
# Part c
new_pig <- tibble(Drug = as.factor(2), Momweight = 140, Dadweight = 185)

predict(pig_model1, newdata = new_pig, interval = "prediction", level = 0.95)
# Part d
delta_2_3 <- summary(pig_model1)$coefficient[2,1] - summary(pig_model1)$coefficient[3,1]

SE_delta <- summary(pig_model1)$coefficient[2,2] + summary(pig_model1)$coefficient[3,2]

t_cv <- qt(0.95, 68)

d2d3_CI <- delta_2_3 + c(-1, 1) * SE_delta * t_cv
#Part e
pig_model_nodrug <- lm(Pigweight ~ Momweight + Dadweight, data = df_pigs)
pig_model_noparent <- lm(Pigweight ~ Drug, data = df_pigs)

anova(pig_model_nodrug, pig_model1)
#anova(pig_model_noparent, pig_model1)
```

```r
#anova(pig_model1)
```