HW#3

Halid Kopanski

2022-06-06

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
```

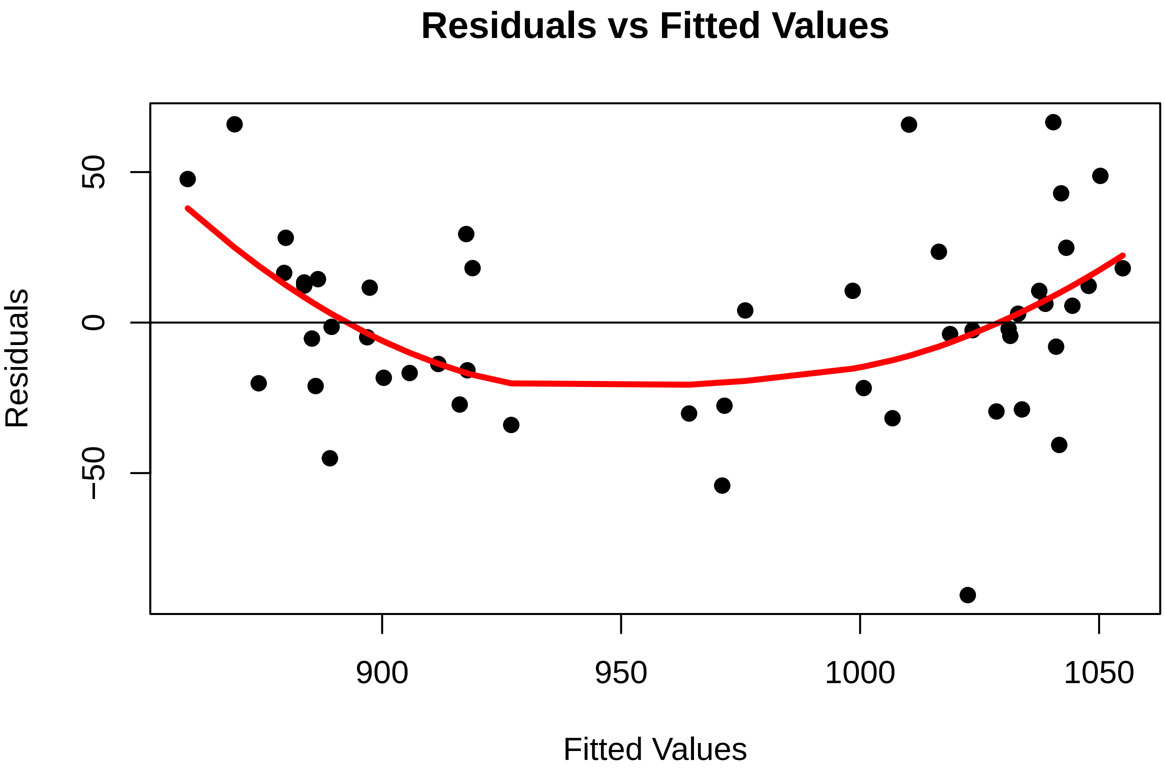
```
##        fit       lwr      upr
## 1 2.372534 0.9383436 3.806724
```

```
## 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

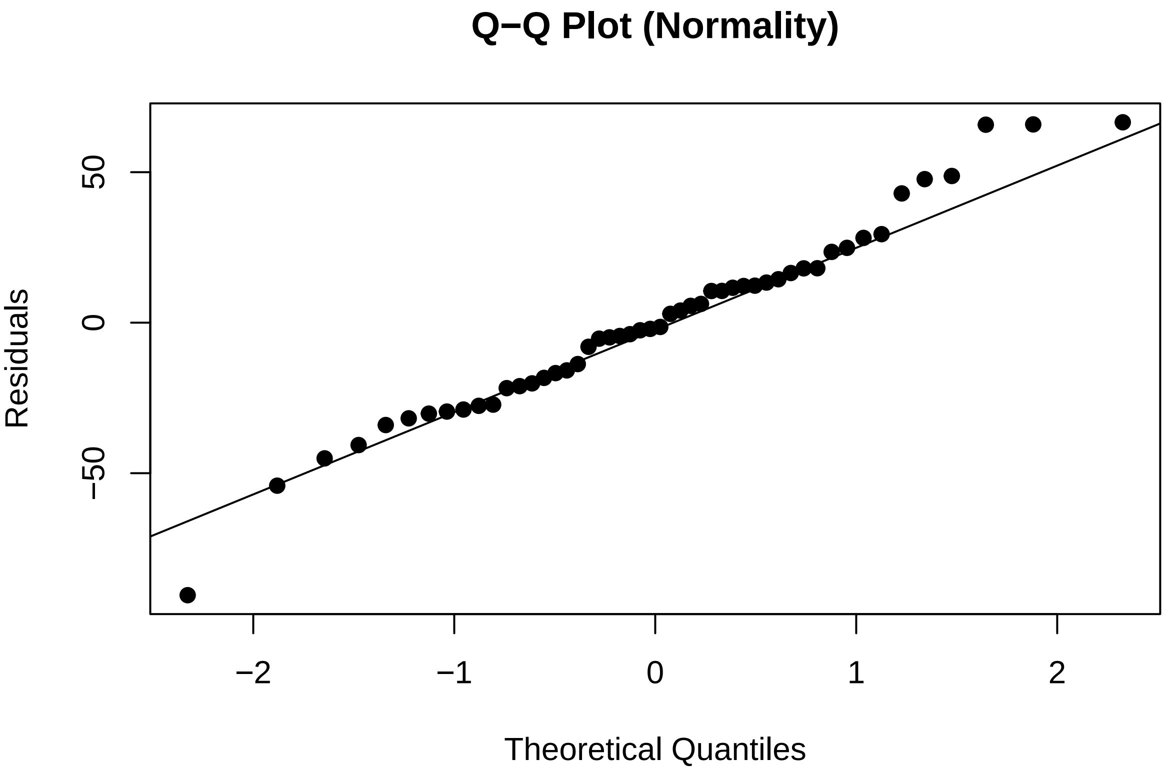
```
##
```

```
##  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)

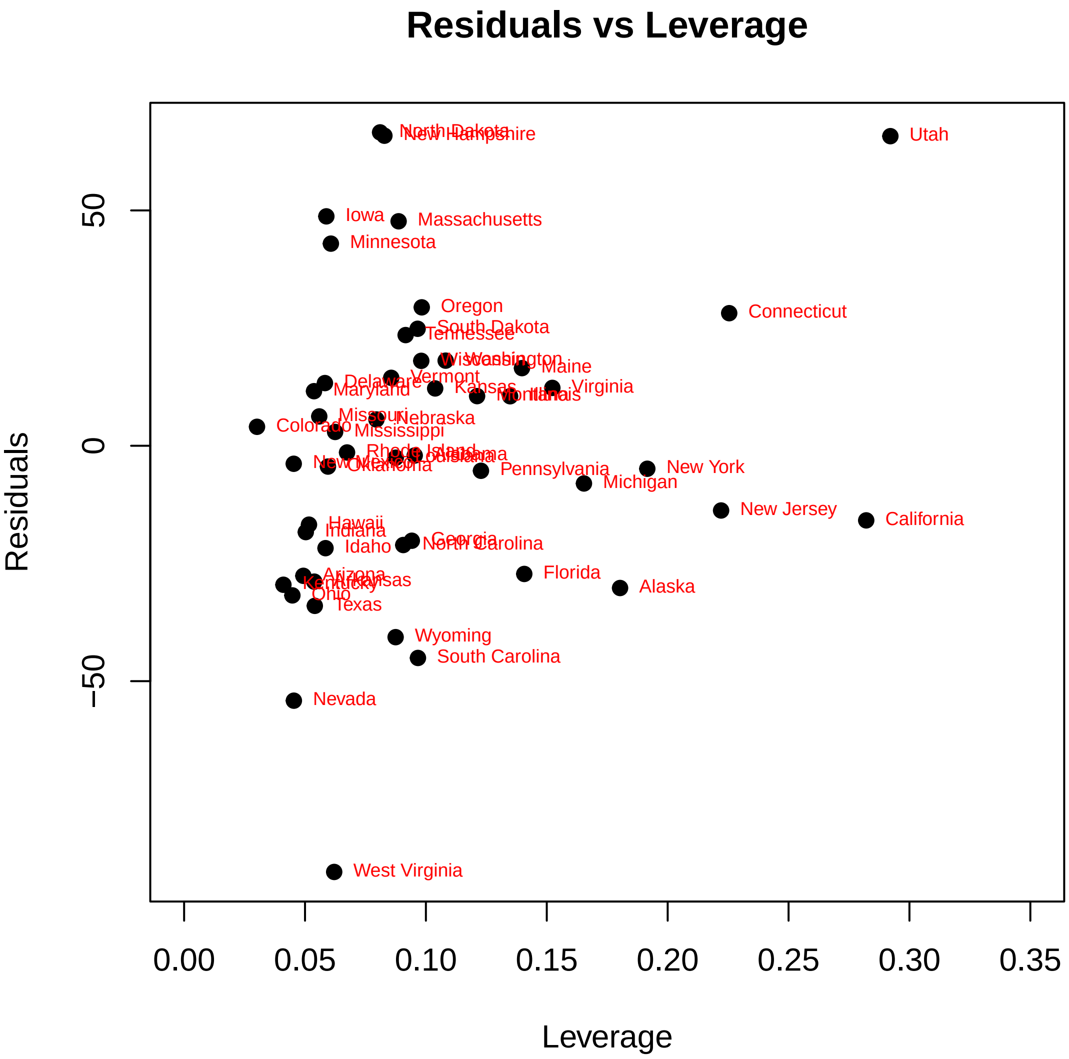


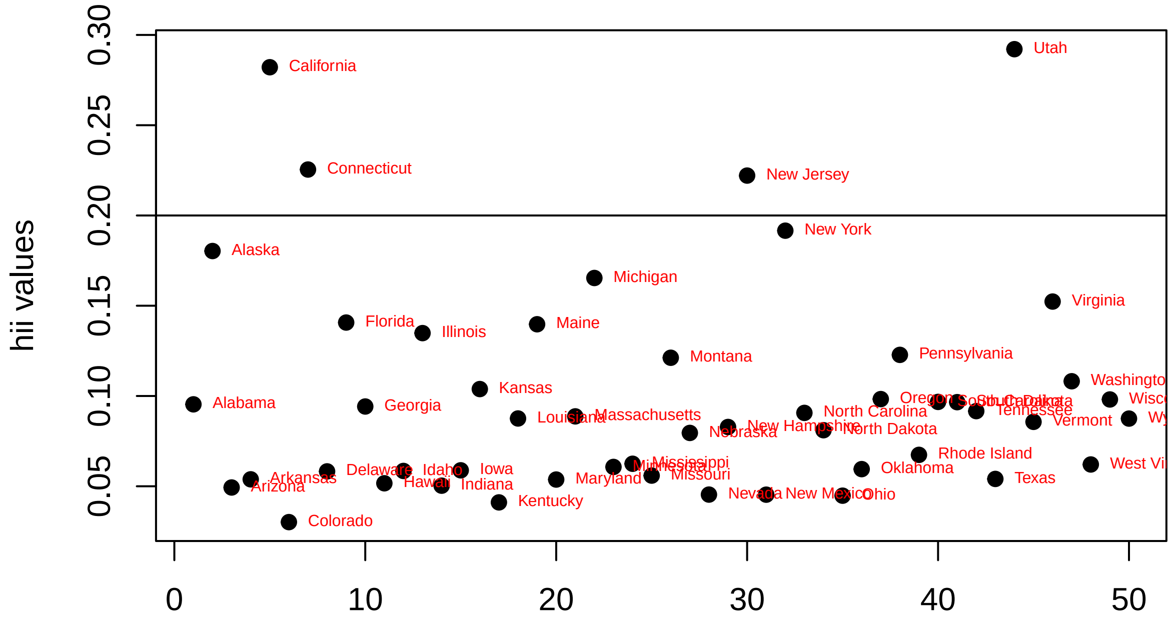
Results of the Shapiro and the Durbin-Watson Tests

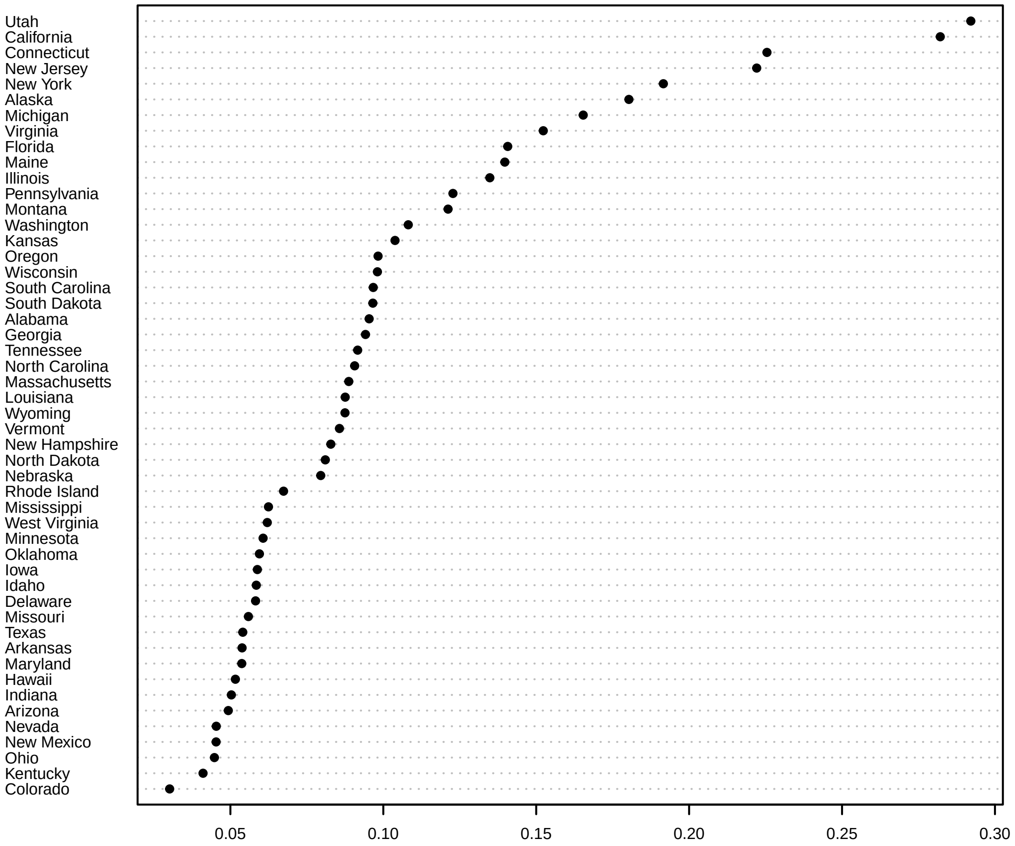
```
##
##  Shapiro-Wilk normality test
##
## data:  sat_model1$residuals
## W = 0.97691, p-value = 0.4304
```

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

c)

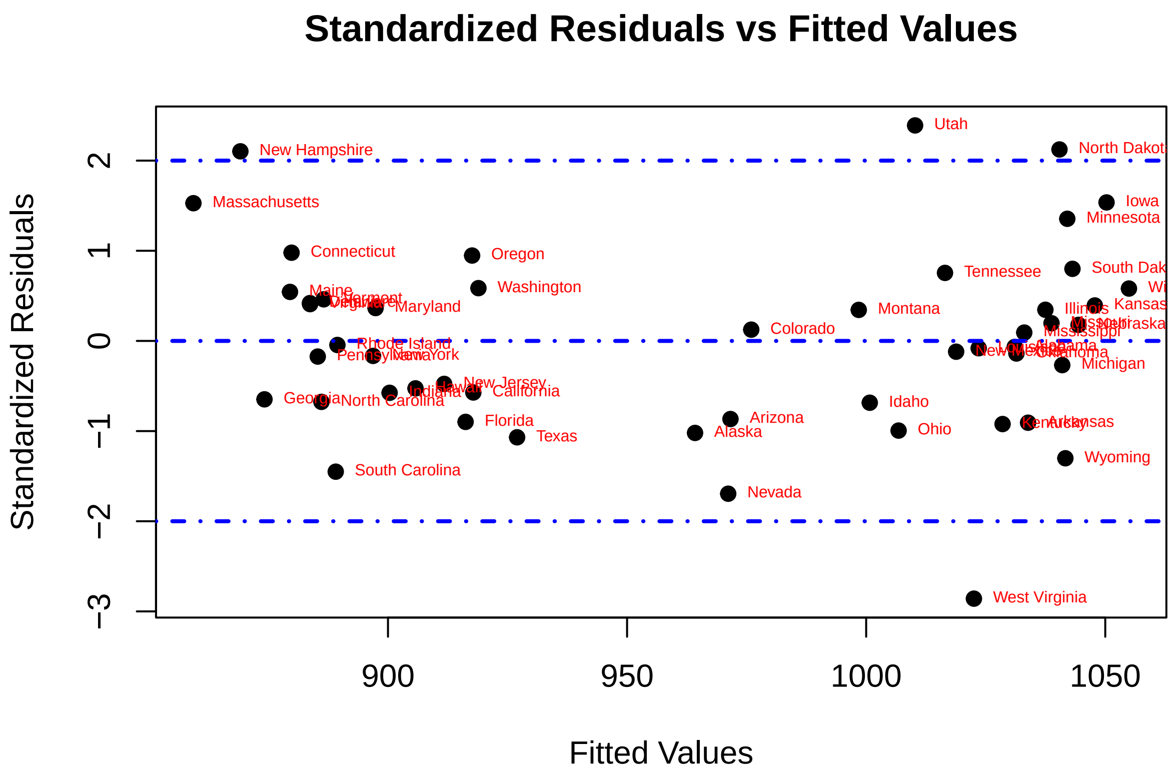

Residuals vs Leverage
Residuals
Leverage
50
0
−50
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
North Dakota
New Hampshire
Utah
Iowa
Massachusetts
Minnesota
Oregon
Connecticut
South Dakota
Tennessee
Wisconsin
Washington
Maine
Delaware
Vermont
Maryland
Kansas
Virginia
Montana
Illinois
Missouri
Nebraska
Colorado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

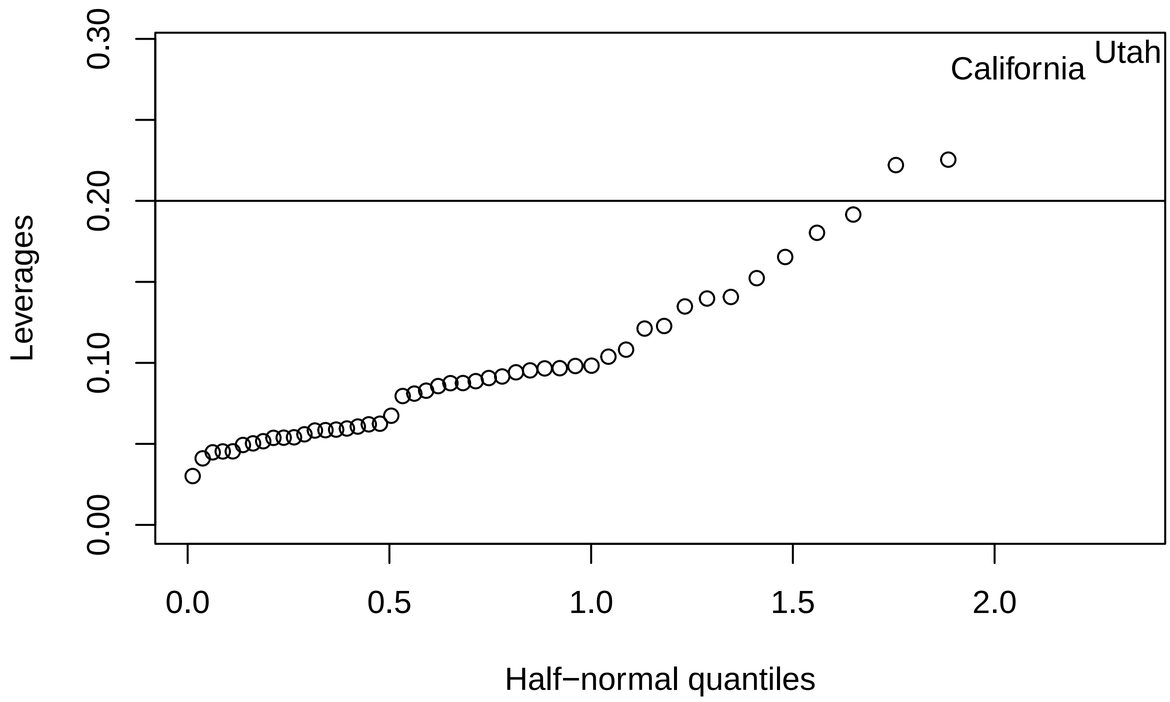




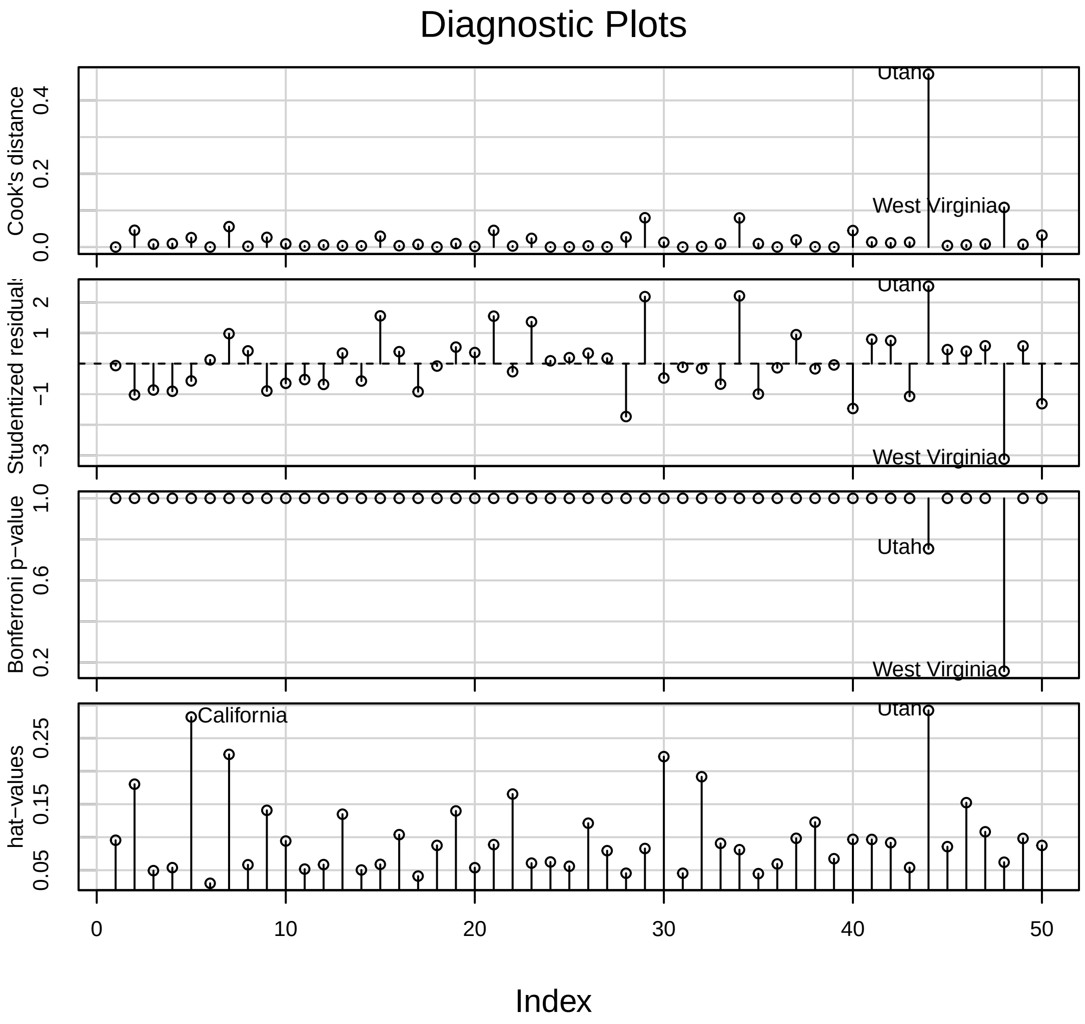
d)

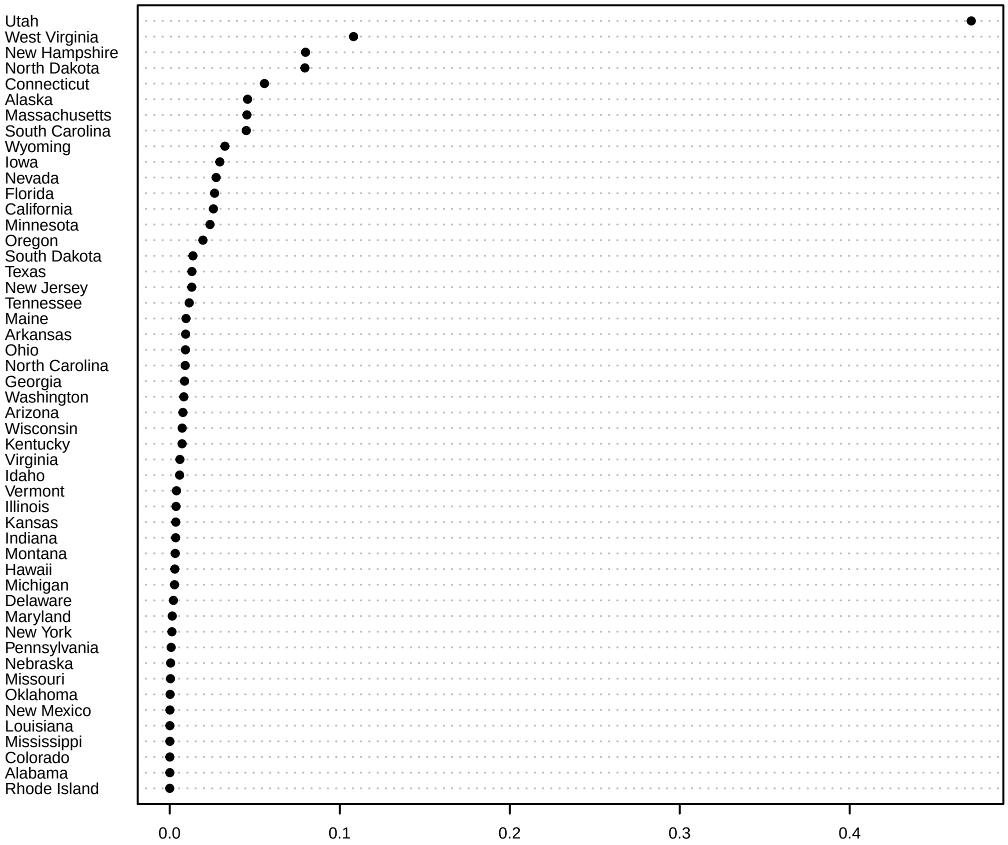
```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##                rstudent unadjusted p-value Bonferroni p
## West Virginia -3.124428          0.0031496      0.15748
```


Standardized Residuals vs Fitted Values
Standardized Residuals
Fitted Values
2
1
0
−1
−2
−3
900
950
1000
1050
Utah
North Dakota
New Hampshire
Massachusetts
Iowa
Minnesota
Connecticut
Oregon
Tennessee
South Dak
Washington
Wi
Maine
Vermont
Delaware
Virginia
Maryland
Illinois
Kansas
Colorado
Montana
Missouri
Nebraska
Mississippi
Rhode Island
Pennsylvania
New York
Louisiana
Alabama
New Mexico
Oklahoma
Michigan
New Jersey
Hawaii
California
Indiana
Georgia
North Carolina
Idaho
Florida
Arizona
Texas
Alaska
Ohio
Kentucky
Arkansas
Wyoming
South Carolina
Nevada
West Virginia


Leverages
Half−normal quantiles
0.00
0.10
0.20
0.30
0.0
0.5
1.0
1.5
2.0
California
Utah

e)


Diagnostic Plots
Cook's distance
0.0
0.2
0.4
Utah
West Virginia
Studentized residuals
2
1
−1
−3
Utah
West Virginia
Bonferroni p−value
1.0
0.6
0.2
Utah
West Virginia
hat−values
0.05
0.15
0.25
California
Utah
0
10
20
30
40
50
Index


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
```

Code

```
library(faraway)
library(car)
library(MASS)
library(lmtest)
library(tidyverse)
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$fitted.values, y = rstandard(sat_model1), pch = 19,
     main = "Standardized Residuals vs Fitted Values", xlab = "Fitted Values", ylab = "Standardized Res
abline(h = c(0, 2, -2), col = "blue", lty = 4, lwd = 2)
text(x = sat_model1$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)
```

```
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")
```