Assignment 4 Solutions

Question 1

(a)

$$\text{tr}(H) = \text{tr}\left(X(X^TX)^{-1}X^T\right) = \text{tr}\left((X^TX)^{-1}X^TX\right) = \text{tr}(I_{p+1}) = p+1$$

where the second equivalence is due to the hint and I_{p+1} is the $(p+1)\times(p+1)$ identity matrix.

(b)

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{\text{tr}(H)}{n} = \frac{p+1}{n}$$

(c) The i^{th} diagonal element of HH is found by multiplying the i^{th} row of H by the i^{th} column of H:

$$\begin{bmatrix} h_{i1} & h_{i2} & \cdots & h_{in} \end{bmatrix} \begin{bmatrix} h_{1i} \\ h_{2i} \\ \vdots \\ h_{ni} \end{bmatrix} = \sum_{j=1}^{n} h_{ij}h_{ji} = h_{ii}^2 + \sum_{j\neq i} h_{ij}h_{ji} = h_{ii}^2 + \sum_{j\neq i} h_{ij}^2.$$

Note that this last equivalence is due to the fact that H is a symmetric matrix. And now, because H is also an idempotent matrix, the i^{th} diagonal element of H is equivalent to the i^{th} diagonal element of HH:

$$h_{ii} = h_{ii}^2 + \sum_{j\neq i} h_{ij}^2.$$

Because the right hand side is a sum of squares, it is clear that $h_{ii} \geq 0$. We also see that $h_{ii} \geq h_{ii}^2$, implying that $h_{ii} \leq 1$.

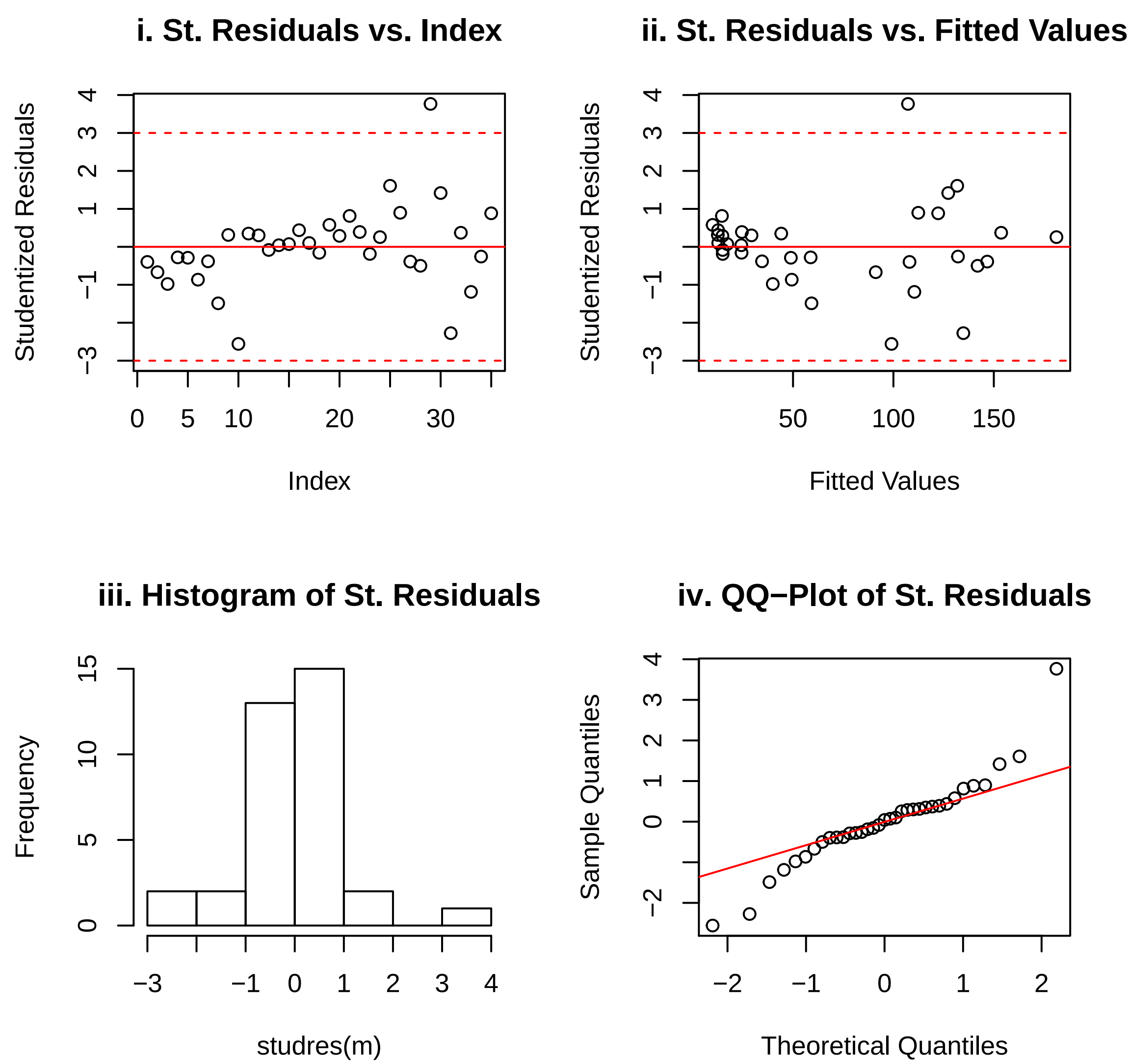
Question 2

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{(i)} &= \left(X_{(i)}^TX_{(i)}\right)^{-1} X_{(i)}^T\boldsymbol{y}_{(i)} \\
&= \left(X^TX - \boldsymbol{x}_i\boldsymbol{x}_i^T\right)\left(X^T\boldsymbol{y} - \boldsymbol{x}_iy_i\right) \\
&= \left[\left(X^TX\right)^{-1} + \frac{\left(X^TX\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T\left(X^TX\right)^{-1}}{1-\boldsymbol{x}_i^T\left(X^TX\right)^{-1}\boldsymbol{x}_i}\right]\left(X^T\boldsymbol{y} - \boldsymbol{x}_iy_i\right) \\
&= \left(X^TX\right)^{-1}X^T\boldsymbol{y} - \left(X^TX\right)^{-1}\boldsymbol{x}_iy_i + \frac{\left(X^TX\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T\left(X^TX\right)^{-1}\left(X^T\boldsymbol{y} - \boldsymbol{x}_iy_i\right)}{1-\boldsymbol{x}_i^T\left(X^TX\right)^{-1}\boldsymbol{x}_i} \\
&= \hat{\boldsymbol{\beta}} - \left(X^TX\right)^{-1}\boldsymbol{x}_iy_i + \frac{\left(X^TX\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} - \left(X^TX\right)^{-1}\boldsymbol{x}_ih_{ii}y_i}{1-h_{ii}} \\
&= \hat{\boldsymbol{\beta}} - \frac{\left(X^TX\right)^{-1}\boldsymbol{x}_i}{1-h_{ii}}\left[y_i(1-h_{ii}) - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} + h_{ii}y_i\right] \\
&= \hat{\boldsymbol{\beta}} - \frac{\left(X^TX\right)^{-1}\boldsymbol{x}_i}{1-h_{ii}}\left[y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right] \\
&= \hat{\boldsymbol{\beta}} - \frac{\left(X^TX\right)^{-1}\boldsymbol{x}_ie_i}{1-h_{ii}}
\end{aligned}$$

Question 3

(a)

```r
library(MASS)
m <- lm(area ~ height + caliper, data = forest)
h <- hatvalues(m)
d <- studres(m)
par(mfrow = c(2,2))
plot(d, ylim = c(min(-3, min(d)),max(3, max(d))),
     main = "i. St. Residuals vs. Index", xlab = "Index",
     ylab = "Studentized Residuals")
abline(h=c(0,3,-3), lty = c(1,2,2), col = "red")
plot(m$fitted.values, d, ylim = c(min(-3, min(d)),max(3, max(d))),
     main = "ii. St. Residuals vs. Fitted Values",
     xlab = "Fitted Values", ylab = "Studentized Residuals")
abline(h=c(0,3,-3), lty = c(1,2,2), col = "red")
hist(studres(m), main = "iii. Histogram of St. Residuals")
qqnorm(studres(m), main = "iv. QQ-Plot of St. Residuals")
qqline(studres(m), col = "red")
```



(b)

- i. No – the residuals vs. index plot does not indicate a random scattering of points; instead we see somewhat of a ‘bow-tie’ pattern.
- ii. No – the residuals vs. fitted values plots indicates increased variability in the residuals for larger fitted values.
- iii. No – There appears to be one unusually large residual, and the rest of the residuals appear to be left skewed.
- iv. Yes – on each of the four plots there is evidence of a single residual that appears to be substantially different from all of the rest.

(c)

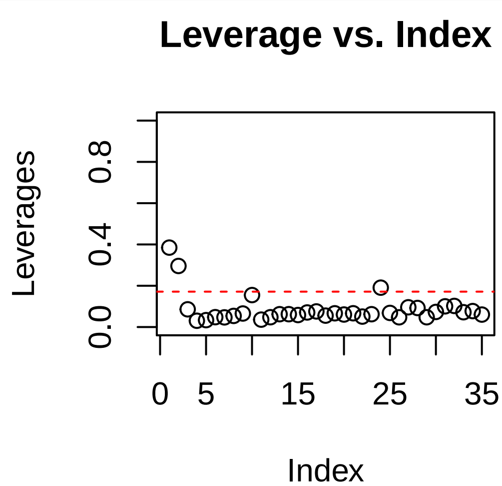
```
which(d == max(d))
```

```
## 29
## 29
```

The output above indicates the observation 29 has the largest studentized residual.

(d)

```
plot(h, ylim = c(0,1), main = "Leverage vs. Index", ylab = "Leverages")
abline(h = 2*mean(h), lty = 2, col = "red")
```



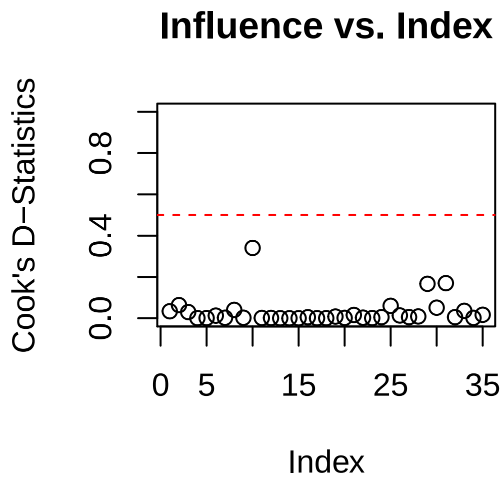

```
which(h > 2*mean(h))
```

```
##  1  2 24
##  1  2 24
```

From the output above, we see that there are three points with “high” leverage (i.e., leverage larged than twice the average leverage: $2\bar{h}$. These are observations 1, 2, and 24.

(e)

```
cook_d <- cooks.distance(m)
plot(cook_d, ylim = c(0,1), main = "Influence vs. Index",
     ylab = "Cook's D-Statistics")
abline(h = 0.5, lty = 2, col = "red")
```



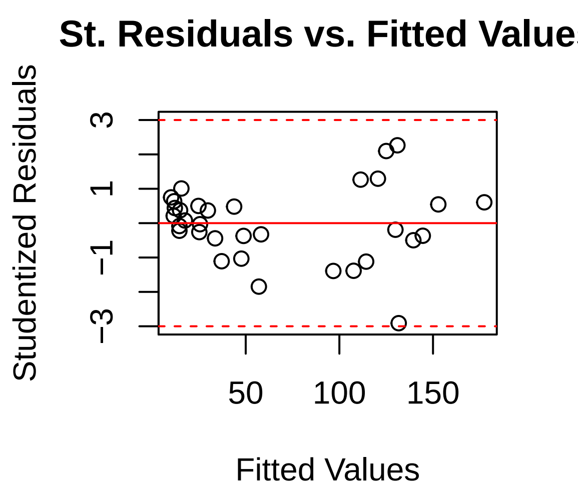

```
which(cook_d > 0.1)
```

```
## 10 29 31
## 10 29 31
```

Thus the top three most infuential observations are 10, 29, and 31.

(f)

```
m_new <- lm(area ~ height + caliper, data = forest[-29,][-10,])
d_new <- studres(m_new)
plot(m_new$fitted.values, d_new, ylim = c(min(-3, min(d_new)),max(3, max(d_new))),
     main = "St. Residuals vs. Fitted Values", xlab = "Fitted Values",
     ylab = "Studentized Residuals")
abline(h=c(0,3,-3), lty = c(1,2,2), col = "red")
```

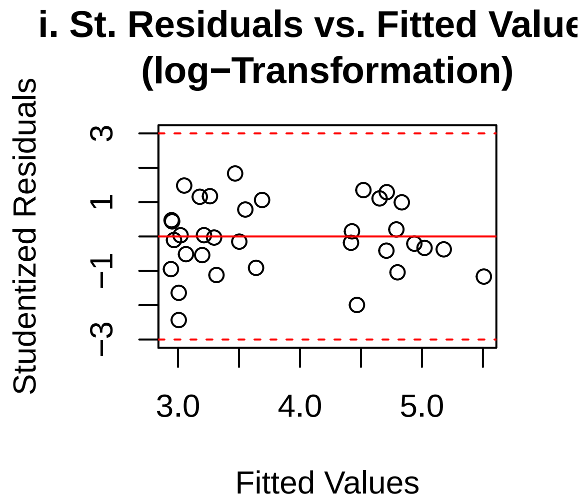


(g) No – we see that the outlier has now been removed, but we still see larger residual variation for large fitted values (i.e., the fan/ funnel shape).

(h) i.

```
m_l <- lm(log(area) ~ height + caliper, data = forest[-29,][-10])
d_l <- studres(m_l)
plot(m_l$fitted.values, d_l, ylim = c(min(-3, min(d_new)),max(3, max(d_new))),
     main = "i. St. Residuals vs. Fitted Values \n(log-Transformation)",
     xlab = "Fitted Values",
```

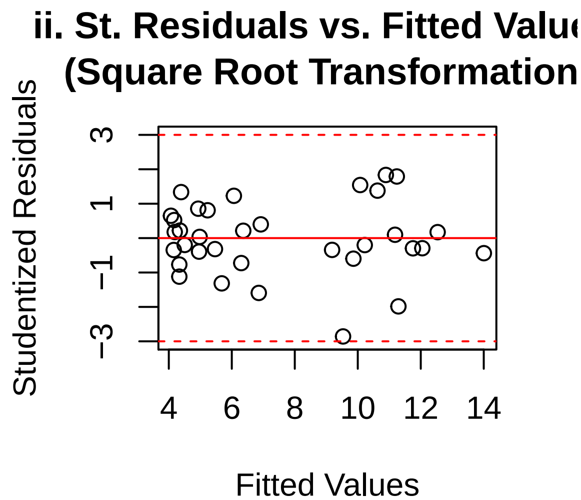
```
     ylab = "Studentized Residuals")
abline(h=c(0,3,-3), lty = c(1,2,2), col = "red")
```



Yes – the plot displays a random scattering of points, suggesting that the log-transformation has stabilized the variance.

(h) ii.

```
m_sq <- lm(sqrt(area) ~ height + caliper, data = forest[-29,][-10])
d_sq <- studres(m_sq)
plot(m_sq$fitted.values, d_sq, ylim = c(min(-3, min(d_new)),max(3, max(d_new))),
     main = "ii. St. Residuals vs. Fitted Values \n(Square Root Transformation)",
     xlab = "Fitted Values",
     ylab = "Studentized Residuals")
abline(h=c(0,3,-3), lty = c(1,2,2), col = "red")
```



Yes – the plot displays a random scattering of points, suggesting that the square root transformation has stabilized the variance.

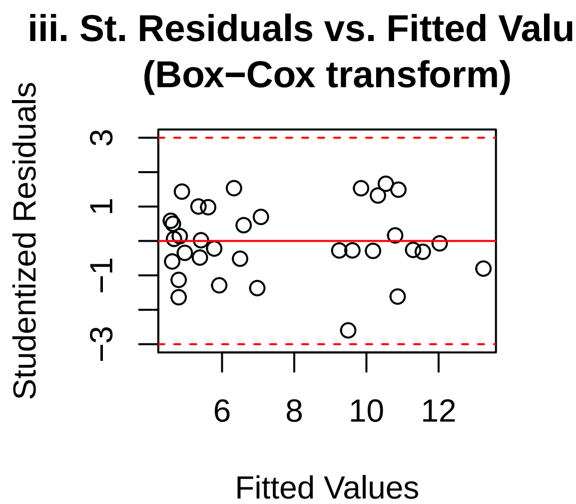
(h) iii.

```
bc <- boxcox(m_new, plotit = FALSE)
lambda <- bc$x[which(bc$y == max(bc$y))]
```

```
lambda
```

```
## [1] 0.3
```

```
m_bc <- lm((area^lambda - 1)/lambda ~ height + caliper, data = forest[-29,][-10])
d_bc <- studres(m_bc)
plot(m_bc$fitted.values, d_bc, ylim = c(min(-3, min(d_new)),max(3, max(d_new))),
     main = "iii. St. Residuals vs. Fitted Values \n(Box-Cox transform)",
     xlab = "Fitted Values",
     ylab = "Studentized Residuals")
abline(h=c(0,3,-3), lty = c(1,2,2), col = "red")
```



Yes – the plot displays a random scattering of points, suggesting that the Box-Cox transformation with $\lambda = 0.3$ has stabilized the variance.

(i)

```
coef(m_l)
```

```
## (Intercept)      height     caliper
##  1.65404534  0.06876379  0.49506397
```

```
exp(coef(m_l))
```

```
## (Intercept)      height     caliper
##    5.228086    1.071183    1.640603
```

- The output above indicaes that $\hat{\beta}_0 = 0.0688$ and that $e^{\hat{\beta}_0} = 1.0712$. This indicates that for every unit-increase in height (for some fixed caliper) we expect the tree's needle area to increase by a factor of 1.0712 (i.e., the need area will be 1.0712 times bigger).
- The output above also indicaes that $\hat{\beta}_1 = 0.4951$ and that $e^{\hat{\beta}_1} = 1.6406$. This indicates that for every unit-increase in caliper (for some fixed height) we expect the tree's needle area to increase by a factor of 1.6406 (i.e., the need area will be 1.6406 times bigger).

Question 4

(a)

```
library(car, quietly = TRUE)
m_full <- lm(Salary ~ ., data = hitters)
sort(vif(m_full), decreasing = TRUE)
```

```
##      CHits     CAtBat      CRuns       CRBI     CHmRun       Hits
## 502.954289 251.561160 162.520810 131.965858  46.488462  30.281255
##      AtBat     CWalks       Runs        RBI      Years      HmRun
##  22.944366  19.744105  15.246418  11.921715   9.313280   7.758668
##      Walks     League  NewLeague    Assists     Errors    PutOuts
##   4.148712   4.134115   4.099063   2.709341   2.214543   1.236317
##   Division
##   1.075398
```

As can be seen from the VIFs calculated above, multicollinearity does appear to be an issue. In particular, the explanatory variables CRuns, CAtBat and CHits have the three largest (and they are extremely large) VIFs. This is unsurprising since the number of runs scored and RBIs in a player's career are both going to be highy correlated with the number of hits the player gets in their career and these will all by highly correlated with the number of times a player gets to bat in their career.

(b)

```r
library(leaps, quietly = TRUE)
all_poss <- regsubsets(Salary ~ ., data = hitters,
                       nvmax = 19, nbest = 2^19, really.big = TRUE)
all_poss_summ <- summary(all_poss)
min_indx <- which.min(all_poss_summ$bic)
all_poss_summ$which[min_indx,]
```

Explanatory Variable	In the Model
AtBat	TRUE
Hits	TRUE
HmRun	FALSE
Runs	FALSE
RBI	FALSE
Walks	TRUE
Years	FALSE
CAtBat	FALSE
CHits	FALSE
CHmRun	FALSE
CRuns	FALSE
CRBI	TRUE
CWalks	FALSE
League	FALSE
Division	TRUE
PutOuts	TRUE
Assists	FALSE
Errors	FALSE
NewLeague	FALSE

The *all possible regressions* approach fits all 2^q possible models (where q is the number of potential explanatory variables) and identifies the optimal model based on some goodness-of-fit criteria. Here we used BIC as the decision criteria, and among all $2^{19} = 524,288$ models, the one with the minimum BIC value is the one that contains the explanatory these variables: AtBat, Hits, Walks, CRBI, Division, PutOuts.

(c)

```r
library(MASS, quietly = TRUE)
# Preliminary stuff that will be required by all three stepwise selection techniques:
n <- dim(hitters)[1]
```

```r
sml <- lm(Salary ~ 1, data = hitters)
lrg <- lm(Salary ~ ., data = hitters)

# Forward Selection
m_f <- stepAIC(object = sml, scope = list(upper = lrg, lower = sml),
               direction = "forward", trace = 0, k = log(n))
variable.names(m_f)[2:length(variable.names(m_f))]
```

```
## [1] "CRBI"      "Hits"      "PutOuts"   "DivisionW" "AtBat"     "Walks"
```

Thus the final model that is chosen by *Forward Selection* contains the following explanatory variables: CRBI, Hits, PutOuts, Division, AtBat and Walks. The order in which they arrived into the model and the corresponding reductions in BIC are shown in the following table. Note that this is the *optimal model* identified by the all-possible-regressions approach.

Iteration	Variable Added	Change in BIC
1	CRBI	-96.42
2	Hits	-38.08
3	PutOuts	-6.70
4	Division	-6.18
5	AtBat	-2.26
6	Walks	-3.85

(d)

```r
# Backward Elimination
m_b <- stepAIC(object = lrg, scope = list(upper = lrg, lower = sml),
               direction = "backward", trace = 0, k = log(n))
variable.names(m_b)[2:length(variable.names(m_b))]
```

```
## [1] "AtBat"     "Hits"      "Walks"     "CRuns"     "CRBI"      "CWalks"
## [7] "DivisionW" "PutOuts"
```

Thus the final model that is chosen by *Backward Elimination* contains the following explanatory variables: AtBat, Hits, Walks, CRuns, CRBI, CWalks, Division, and PutOuts. The order in which the variables were eliminated from the model (before this final model was selected) and the corresponding reductions in BIC are shown in the following table.

Iteration	Variable Eliminated	Change in BIC
1	CHmRun	-5.59
2	Years	-5.49
3	NewLeague	-5.46
4	RBI	-5.40
5	CHits	-5.43
6	HmRun	-5.13
7	Errors	-5.09
8	Runs	-4.99
9	League	-4.36
10	Assists	-2.22
11	CAtBat	-1.94

(e)

```r
# Hybrid Selection
m_h <- stepAIC(object = sml, scope = list(upper = lrg, lower = sml),
               direction = "both", trace = 0, k = log(n))
variable.names(m_h)[2:length(variable.names(m_h))]
```

```
## [1] "CRBI"      "Hits"      "PutOuts"   "DivisionW" "AtBat"     "Walks"
```

Thus the final model that is chosen by *Hybrid Selection* contains the following explanatory variables: CRBI, Hits, PutOuts, Division, AtBat, and Walks. The order in which the variables were added to/ eliminated from the model (before this final model was selected) and the corresponding reductions in BIC are shown in the following table.

Iteration	Variable Changed	Change in BIC
1	+CRBI	-96.42
2	+Hits	-38.08
3	+PutOuts	-6.70
4	+Division	-6.18
5	+AtBat	-2.26
6	+Walks	-3.85