Assignment 3 Solutions

Question 1

(a) The comparison of two nested models (model 1 is nested within model 2) can be done using the *additional sum of squares principle*. Within this framework model 2 would be considered the *full model* and model 1 would considered the *reduced model*. In class we proved that the SSE in the full model would always be smaller than the SSE in the reduced model. Thus, in the context of this problem, we know

$$SSE_1 > SSE_2.$$

Then, dividing by the total sum of squares yields

$$\frac{SSE_1}{SST} > \frac{SSE_2}{SST}.$$

Note that because model 1 and model 2 are fit to the same dataset there is a common SST (since this depends only on the response observations and not the model):

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Given the inequality above, we have

$$1 - \frac{SSE_1}{SST} < 1 - \frac{SSE_2}{SST}$$

which is equivalent to

$$\frac{SSR_1}{SST} < \frac{SSR_2}{SST}$$

which is the desired result:

$$R_1^2 < R_2^2.$$

(b) In part (a) we proved that by adding explanatory variables into the model, the R^2 value always increases. This means that we can make R^2 arbitrarily close to 1 by adding more and more explanatory variables into the model – even if they aren't significantly related to the response variable. The R^2_{adj} metric protects us from this; it penalizes you (i.e., its value may decrease) when you add unimportant explanatory variables into the model. Thus R^2_{adj} is a better measure of *goodness of fit*.

(c)

$$\begin{aligned}
\lim_{n\to\infty} R^2_{adj} &= \lim_{n\to\infty} 1 - (1 - R^2)\left(\frac{n-1}{n-p-1}\right) \\
&= 1 - (1 - R^2)\left(\lim_{n\to\infty} \frac{n-1}{n-p-1}\right) \\
&= 1 - (1 - R^2)\left(\lim_{n\to\infty} \frac{1 - \frac{1}{n}}{1 - \frac{p}{n} - \frac{1}{n}}\right) \\
&= 1 - (1 - R^2)\left(\lim_{n\to\infty} \frac{1 - 0}{1 - 0 - 0}\right) \\
&= 1 - (1 - R^2)(1) \\
&= R^2
\end{aligned}$$

Question 2

(a)

$$\beta_1 = \mathrm{E}[y_i | x_{i1} = 1]$$

which means that β_1 is the expected response (length of gameplay) in condition 1 (the lollipop hammer condition).

$$\beta_2 = \mathrm{E}[y_i | x_{i2} = 1]$$

which means that β_2 is the expected response (length of gameplay) in condition 2 (the jelly fish condition).

$$\beta_3 = \mathrm{E}[y_i | x_{i3} = 1]$$

which means that β_3 is the expected response (length of gameplay) in condition 3 (the colour bomb condition).

Since μ_1, μ_2, μ_3 are defined as the expected response in each condition we see that $\mu_1 = \beta_1$, $\mu_2 = \beta_2$, $\mu_3 = \beta_3$ and so hypotheses [1] and [2] are equivalent.

(b) In order to calculate $\hat{\boldsymbol{\beta}}$ we must first define X and $\boldsymbol{y}$:

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \boldsymbol{y} = \begin{bmatrix} 12 \\ 11 \\ 10 \\ 10 \\ 7 \\ 13 \\ 15 \\ 14 \\ 16 \end{bmatrix}.$$

Using these we calculate the least squares estimates of the β's as

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$$

$$= \left(\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 12 \\ 11 \\ 10 \\ 10 \\ 7 \\ 13 \\ 15 \\ 14 \\ 16 \end{bmatrix}$$

$$= \left(\begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix} \right)^{-1} \begin{bmatrix} 33 \\ 30 \\ 45 \end{bmatrix}$$

$$= \left(\begin{bmatrix} 1/3 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/3 \end{bmatrix} \right) \begin{bmatrix} 33 \\ 30 \\ 45 \end{bmatrix}$$

$$= \begin{bmatrix} 11 \\ 10 \\ 15 \end{bmatrix}$$

Therefore $\hat{\beta}_1 = 10$, $\hat{\beta}_2 = 11$ and $\hat{\beta}_3 = 15$.

(c)

$$\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 11 \\ 10 \\ 15 \end{bmatrix} = \begin{bmatrix} 11 \\ 11 \\ 11 \\ 10 \\ 10 \\ 10 \\ 15 \\ 15 \\ 15 \end{bmatrix}.$$

Thus the residuals are given by

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{\mu}} = \begin{bmatrix} 12 \\ 11 \\ 10 \\ 10 \\ 7 \\ 13 \\ 15 \\ 14 \\ 16 \end{bmatrix} - \begin{bmatrix} 11 \\ 11 \\ 11 \\ 10 \\ 10 \\ 10 \\ 15 \\ 15 \\ 15 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \\ -3 \\ 3 \\ 0 \\ -1 \\ 1 \end{bmatrix}.$$

The sum of squared error is then given by

$$SSE = \boldsymbol{e}^T\boldsymbol{e} = \sum_{i=1}^{n} e_i^2 = 1^2 + 0^2 + (-1)^2 + 0^2 + (-3)^2 + 3^2 + 0^2 + (-1)^2 + 1^2 = 22$$

(d) Any of the following six possibilities is permissible (order of rows doesn't matter)

$$A = \pm \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$$

$$A = \pm \begin{bmatrix} -1 & 1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$$

$$A = \pm \begin{bmatrix} -1 & 0 & 1 \\ 0 & -1 & 1 \end{bmatrix}$$

The reduced model that is a by-product of the null hypothesis $A\boldsymbol{\beta} = \mathbf{0}$ has just one common β, but because $x_{i1} + x_{i2} + x_{i2} = 1$ for each i, the model is the intercept-only model with no x's:

$$y_i = \beta + \varepsilon_i$$

(e) As discussed in class, the sum of squared error for the intercept-only model is

$$SSE_A = \sum_{i=1}^{n} (y_i - \hat{\beta})^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Since the the least squares estimate of β is $\hat{\beta} = \bar{y} = 12$ the required sum of squares is:

$$SSE_A = (12-12)^2+(11-12)^2+(10-12)^2+(10-12)^2+(7-12)^2+(13-12)^2+(15-12)^2+(14-12)^2+(16-12)^2 = 64$$

(f)

$$t = \frac{(SSE_A - SSE)/l}{SSE/(n-p-1)} = \frac{(64-22)/2}{22/(9-3)} = 5.272727$$

(g) The p-value for this test is $\mathrm{P}(T \geq t) = \mathrm{P}(T \geq 5.272727) = 0.0406$ where $T \sim F_{(2,6)}$. The `R` code used for this calculation is the following:

```r
t <- ((64 - 22)/2)/(22/6)
pval <- pf(q = t, df1 = 2, df2 = 6, lower.tail = FALSE)
print(pval)
```

```
## [1] 0.0406189
```

Since $0.0406 < \alpha = 0.05$ we reject H_0. Therefore, NO, the expected length of game play is not the same in the three booster conditions.

Question 3

(a)

```r
m <- lm(Salary ~ ., data = hitters)
summary(m)
```

```
##
## Call:
## lm(formula = Salary ~ ., data = hitters)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -907.62 -178.35  -31.11  139.09 1877.04
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  163.10359   90.77854   1.797 0.073622 .
## AtBat         -1.97987    0.63398  -3.123 0.002008 **
## Hits           7.50077    2.37753   3.155 0.001808 **
## HmRun          4.33088    6.20145   0.698 0.485616
## Runs          -2.37621    2.98076  -0.797 0.426122
## RBI           -1.04496    2.60088  -0.402 0.688204
## Walks          6.23129    1.82850   3.408 0.000766 ***
## Years         -3.48905   12.41219  -0.281 0.778874
## CAtBat        -0.17134    0.13524  -1.267 0.206380
## CHits          0.13399    0.67455   0.199 0.842713
## CHmRun        -0.17286    1.61724  -0.107 0.914967
## CRuns          1.45430    0.75046   1.938 0.053795 .
## CRBI           0.80771    0.69262   1.166 0.244691
## CWalks        -0.81157    0.32808  -2.474 0.014057 *
## LeagueN       62.59942   79.26140   0.790 0.430424
## DivisionW   -116.84925   40.36695  -2.895 0.004141 **
## PutOuts        0.28189    0.07744   3.640 0.000333 ***
## Assists        0.37107    0.22120   1.678 0.094723 .
## Errors        -3.36076    4.39163  -0.765 0.444857
## NewLeagueN   -24.76233   79.00263  -0.313 0.754218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 315.6 on 243 degrees of freedom
## Multiple R-squared:  0.5461, Adjusted R-squared:  0.5106
## F-statistic: 15.39 on 19 and 243 DF,  p-value: < 2.2e-16
```

From the output above we see that $R^2 = 0.5461$ meaning that this model explains 54.61% of the variation in salaries.

(b) The ANOVA table required here can be easily calculated using the output of the `anova()` function applied to the full model from part (a). This output is shown below

```
## Analysis of Variance Table
##
## Response: Salary
##            Df   Sum Sq Mean Sq F value    Pr(>F)
## AtBat       1  8309469 8309469 83.4356 < 2.2e-16 ***
## Hits        1  2545894 2545894 25.5634 8.431e-07 ***
## HmRun       1  1254597 1254597 12.5974 0.0004636 ***
## Runs        1     7331    7331  0.0736 0.7863812
## RBI         1   896118  896118  8.9980 0.0029839 **
## Walks       1  3335249 3335249 33.4893 2.199e-08 ***
## Years       1  5434238 5434238 54.5654 2.401e-12 ***
## CAtBat      1  2472329 2472329 24.8247 1.193e-06 ***
## CHits       1   865572  865572  8.6912 0.0035090 **
## CHmRun      1   894204  894204  8.9787 0.0030144 **
## CRuns       1    17771   17771  0.1784 0.6730889
## CRBI        1    61684   61684  0.6194 0.4320490
## CWalks      1   457229  457229  4.5911 0.0331331 *
## League      1   178992  178992  1.7973 0.1812950
## Division    1   927646  927646  9.3145 0.0025259 **
## PutOuts     1  1152884 1152884 11.5761 0.0007811 ***
## Assists     1   242089  242089  2.4308 0.1202715
## Errors      1    55332   55332  0.5556 0.4567610
## NewLeague   1     9784    9784  0.0982 0.7542178
## Residuals 243 24200700   99591
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The required table is shown below. Notice that the regression sum of squares and degrees of freedom are calculated by summing all of the sums of squares and degrees of freedom associated with the individual explanatory variables in the table above. MSR and the F statistic are calculated as usual ($MSR = SSR/p$ and $t = MSR/MSE$). Note that a p-value is not required for full points here.

Source	DF	SS	MS	F
Regression	19	29118413	1532548	15.38836
Error	243	24200700	99591.36	
Total	262	53319113		

(c)

i.

$$\mathrm{H}_0 : \beta_1 = \beta_2 = \cdots \beta_{19} = 0 \text{ vs. } \mathrm{H}_\mathrm{A} : \beta_j \neq 0 \text{ for some } j = 1, 2, \ldots, 19$$

ii.

$$t = 15.38836$$

iii.

$$F_{(19,243)}$$

iv.

$$\text{p-value} = \mathrm{P}(T \geq 15.38836) = 7.84 \times 10^{-32}$$

where $T \sim F_{(19,243)}$

(d)

```r
mr <- lm(Salary ~ AtBat + Hits + Walks + CRuns + CWalks + Division + PutOuts +
    Assists, data = hitters)
summary(mr)
```

```
##
## Call:
## lm(formula = Salary ~ AtBat + Hits + Walks + CRuns + CWalks +
##     Division + PutOuts + Assists, data = hitters)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -835.35 -166.39  -29.07  125.09 2008.27
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  107.3087    65.9664   1.627 0.105037
## AtBat         -2.0489     0.5470  -3.746 0.000223 ***
## Hits           6.8459     1.6841   4.065 6.41e-05 ***
## Walks          6.1822     1.5560   3.973 9.25e-05 ***
## CRuns          1.1429     0.2014   5.676 3.76e-08 ***
## CWalks        -0.7305     0.2671  -2.735 0.006685 **
## DivisionW   -115.7654    39.7760  -2.910 0.003930 **
## PutOuts        0.3094     0.0757   4.086 5.88e-05 ***
## Assists        0.0810     0.1507   0.538 0.591285
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 319.3 on 254 degrees of freedom
## Multiple R-squared:  0.5142, Adjusted R-squared:  0.4989
## F-statistic:  33.6 on 8 and 254 DF,  p-value: < 2.2e-16
```

From the output above we see that $R^2 = 0.5142$ which means that this reduced model explains 51.42% of the variation in salaries. This is less than the R^2 value from the full model, but it's not *a lot* less. This suggest that the 11 explanatory variables removed do not explain very much variation in the response. Furthermore, the p-values associated with a test of $\mathrm{H}_0 : \beta = 0$ for each of the eliminated variables are all significantly larger than 0.05, suggesting that these variables do not individually significantly influence the response.

(e)

$$\mathrm{H}_0 : \beta_3 = \beta_4 = \beta_5 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{12} = \beta_{14} = \beta_{18} = \beta_{19} = 0$$

(f) The required comparison can be made using the `anova()` function and passing in both the full and reduced models as follows:

```r
anova(mr, m)
```

```
## Analysis of Variance Table
##
## Model 1: Salary ~ AtBat + Hits + Walks + CRuns + CWalks + Division + PutOuts +
##     Assists
## Model 2: Salary ~ AtBat + Hits + HmRun + Runs + RBI + Walks + Years +
##     CAtBat + CHits + CHmRun + CRuns + CRBI + CWalks + League +
##     Division + PutOuts + Assists + Errors + NewLeague
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1    254 25904006
## 2    243 24200700 11   1703307 1.5548  0.113
```

From the output above we see that the test statistic is $t = 1.5548$ and the p-value is $\mathrm{P}(T \geq 1.5548) = 0.113$ where $T \sim F_{(11,243)}$. Since 0.113>0.05, at a 5% level of significance we fail to reject H_0 meaning that the 11 explanatory variables removed from the model did not significantly influence the response.

(g)

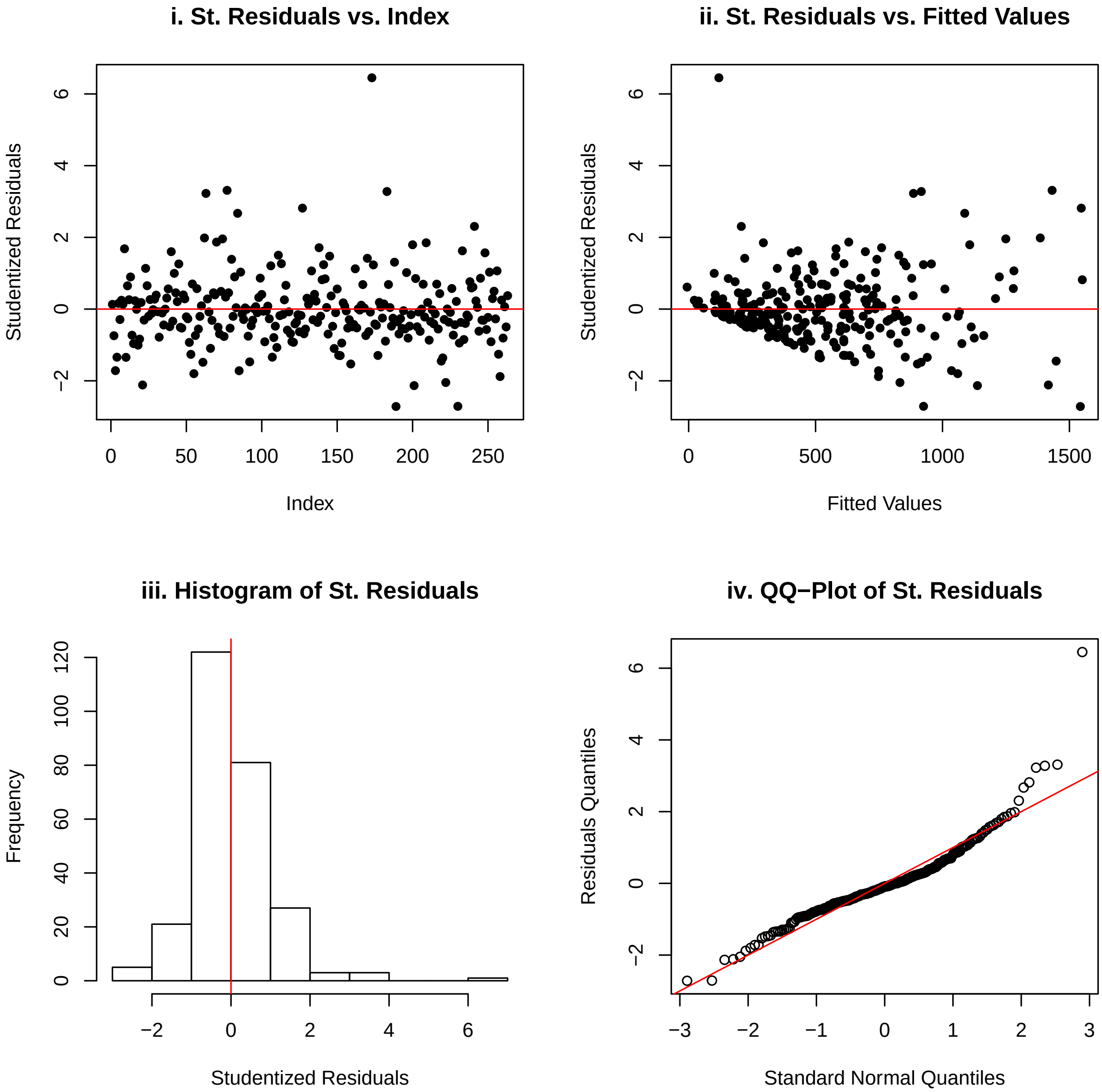
```r
# Calculate the studentized residuals
e <- mr$residuals
sigmahat <- summary(mr)$sigma
h <- hatvalues(mr)
d <- e/(sigmahat * sqrt(1 - h))

par(mfrow = c(2, 2))
# Studentized Residuals vs. Index
n <- dim(hitters)[1]
plot(x = 1:n, y = d, main = "i. St. Residuals vs. Index", xlab = "Index",
    ylab = "Studentized Residuals", pch = 16)
abline(h = 0, col = "red")

# Studentized Residuals vs. Fitted Values
plot(x = mr$fitted.values, y = d, main = "ii. St. Residuals vs. Fitted Values",
    xlab = "Fitted Values", ylab = "Studentized Residuals", pch = 16)
abline(h = 0, col = "red")

# Histogram of Studentized Residuals
hist(x = d, main = "iii. Histogram of St. Residuals", xlab = "Studentized Residuals")
abline(v = 0, col = "red")

# QQ-Plot of Studentized Residuals
qqnorm(y = d, main = "iv. QQ-Plot of St. Residuals", xlab = "Standard Normal Quantiles",
    ylab = "Residuals Quantiles")
abline(a = 0, b = 1, col = "red")
```



Note that the red reference lines are not required for full marks.

(h)

i. YES: the residuals vs. index plot does not suggest the existence of any patterns or relationships.

ii. NO: the residuals vs. fitted values plot indicates an increase in variation as the fitted values increase.

iii. This one is a bit ambiguous. I will accept both answers (YES and NO) as long as the corresponding justification is sound. I would personally say yes, but I also think someone would be justfied in worrying. Both versions of a correct answer are shown below.

- YES: aside from a very small number of extreme residuals the histogram exhibits a bell-shaped and symmetric distribution and the points almost all fall along the line of equality on the QQ-plot.
- NO: both the histogram and QQ-plot suggest that the residuals are slightly right skewed as opposed to being symmetric.

iv. YES: on all four plots we see one residual that is very different from the others (it's value is larger than

6 and all of the others roughly vary betwen ± 3).

(i) Hypothesis tests, confidence intervals, prediction intervals.