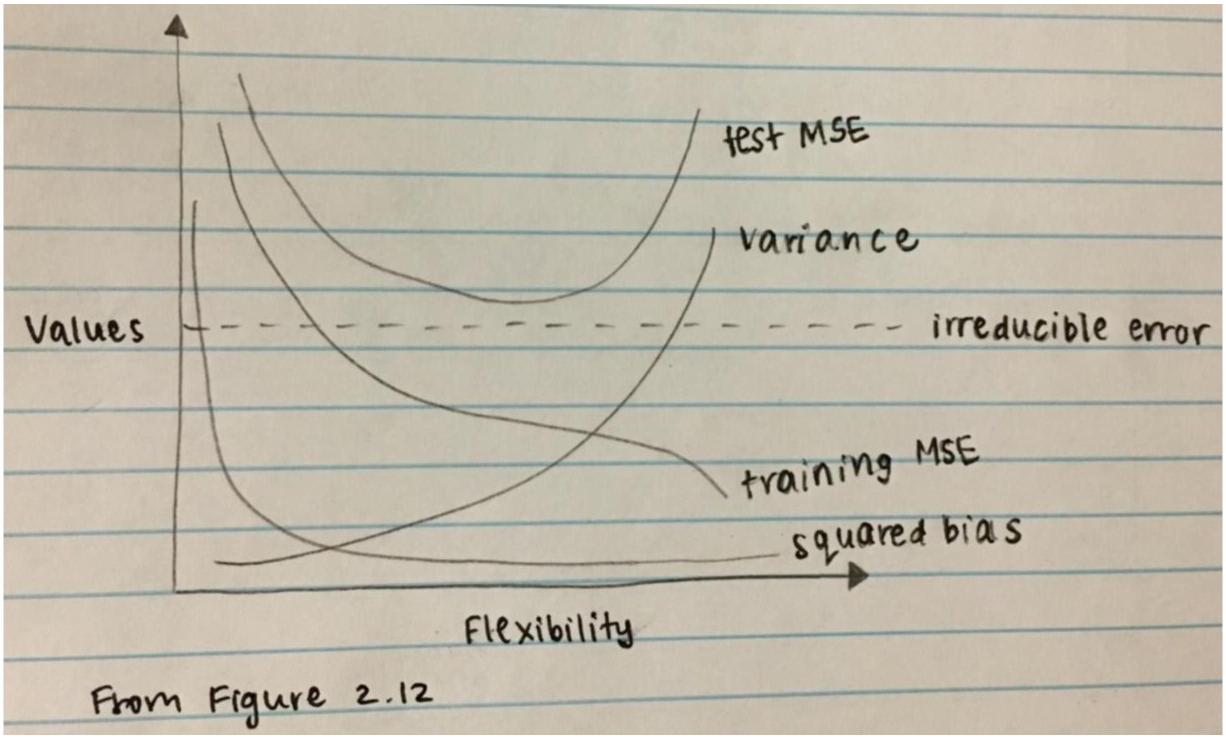
MGSC 310 Problem Set #2

Ananya Vittal

Ch 2

3.

a)



b) The squared bias decreases as the flexibility increases. The variance increases as the flexibility increases. The training error decreases as flexibility increases because the curve fits the observed data more closely. The test error decreases as flexiility increases, then starts to increase again because if there is too much flexibility then overfitting can occur. The irreducible error is constant and the parallel line is below the test error because the expected test MSE will be greater than the irreducible var(e).

Ch 3

1.

The null hypothesis is that there is no relationship between number of units sold and radio, TV, and newspaper advertising budgets. Based on Table 3.4, the p-values for TV and radio are less than 0.05 which means that they are statistically significant. We reject the null hypothesis for radio and TV and can conclude that there is a relationship between number of units sold vs. TV and radio advertising budgets. The p-value for newspaper is greater than 0.05, so we do not reject the null hypothesis and conclude that there is no relationship between number of units sold and newspaper advertising budgets.

3.

a) yhat = 50 + 20(GPA) + 0.07(IQ) + 35(Gender) + 0.01(GPA*IQ) - 10(GPA*Gender)

yhat.f = 50 + 20(GPA) + 0.07(IQ) + 35(1) + 0.01(GPA*IQ) - 10(GPA*1)
yhat.f = 85 + 10(GPA) + 0.07(IQ) + 0.01(GPA*IQ)

yhat.m = 50 + 20(GPA) + 0.07(IQ) + 35(0) + 0.01(GPA*IQ) - 10(GPA*0)
yhat.m = 50 + 20(GPA) + 0.07(IQ) + 0.01(GPA*IQ)

yhat.m >=? yhat.f

50 + 20(GPA) + 0.07(IQ) + 0.01(GPA*IQ) >= 85 + 10(GPA) + 0.07(IQ) + 0.01(GPA*IQ)

10*GPA >= 35

GPA >= 3.5

Answer iii is correct because for a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is greater than 3.5.

b)

```
yhat.f = 85 + 10*4 + 0.07*110 + 0.01*4*110
```

The predicted salary of a female with an IQ of 110 and a GPA of 4.0 is $137,100.

c) False - to determine if there is an interaction effect for GPA and IQ you would have to determine their correlation to each other or look at the p-values/test statistic.

8.

a)

```
#install.packages('ISLR')
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 3.4.2
```

```
data(Auto)
regression<-lm(mpg~horsepower, data = Auto)
summary(regression)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

i.. Yes, there is a relationship between the predictor and response because the p-value is statistically significant.

ii. The relationship between the predictor and the response is pretty strong because the R^2 value is 0.6049.

iii.The relationship between mpg and horsepower is negative because the coefficient for horsepower is negative.

iv.

```
pred <- predict(regression, data.frame(horsepower = 98), interval = "confiden
ce")

conf <- predict(regression, data.frame(horsepower = 98), interval = "predicti
on")
```

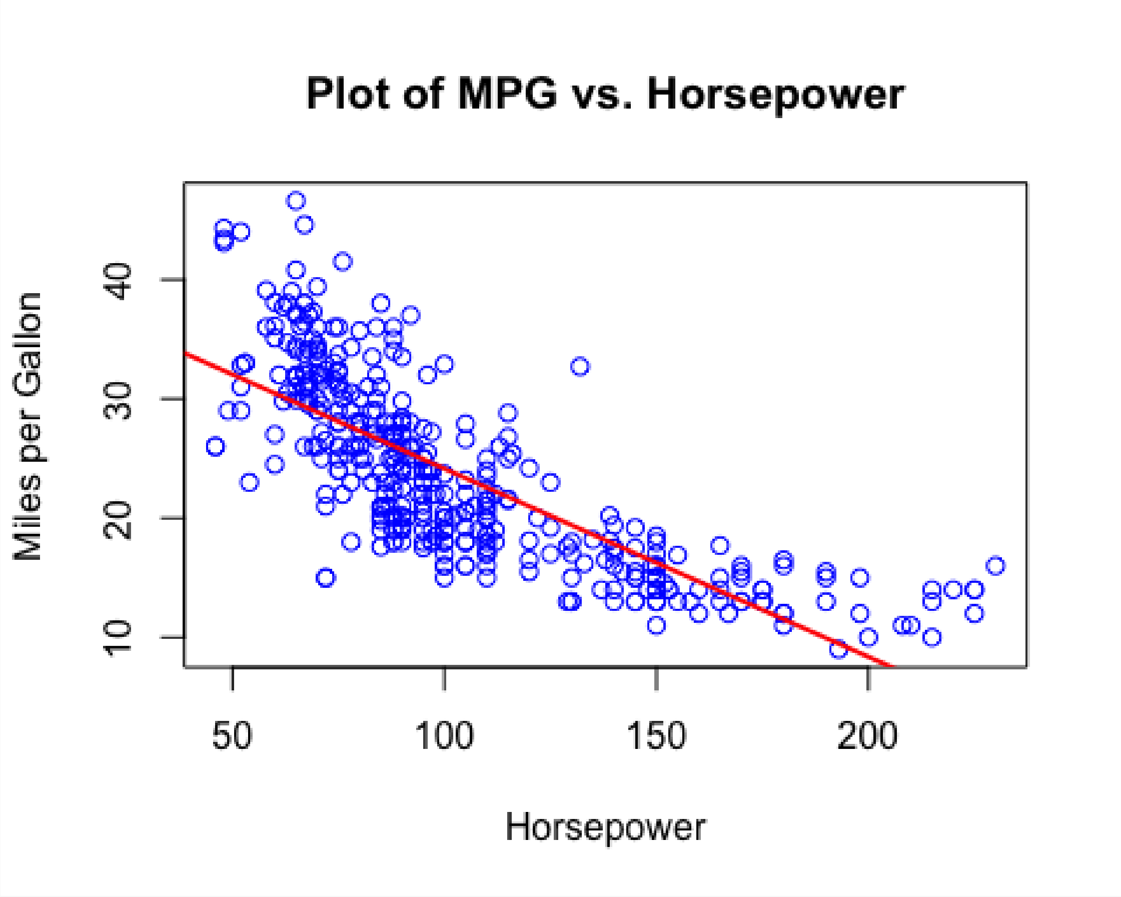
The predicted mpg associated with a horsepower of 98 is 24.46708. The associated confidence interval is (23.97308, 24.96108) and prediction interval is (14.8094, 34.12476).

b)

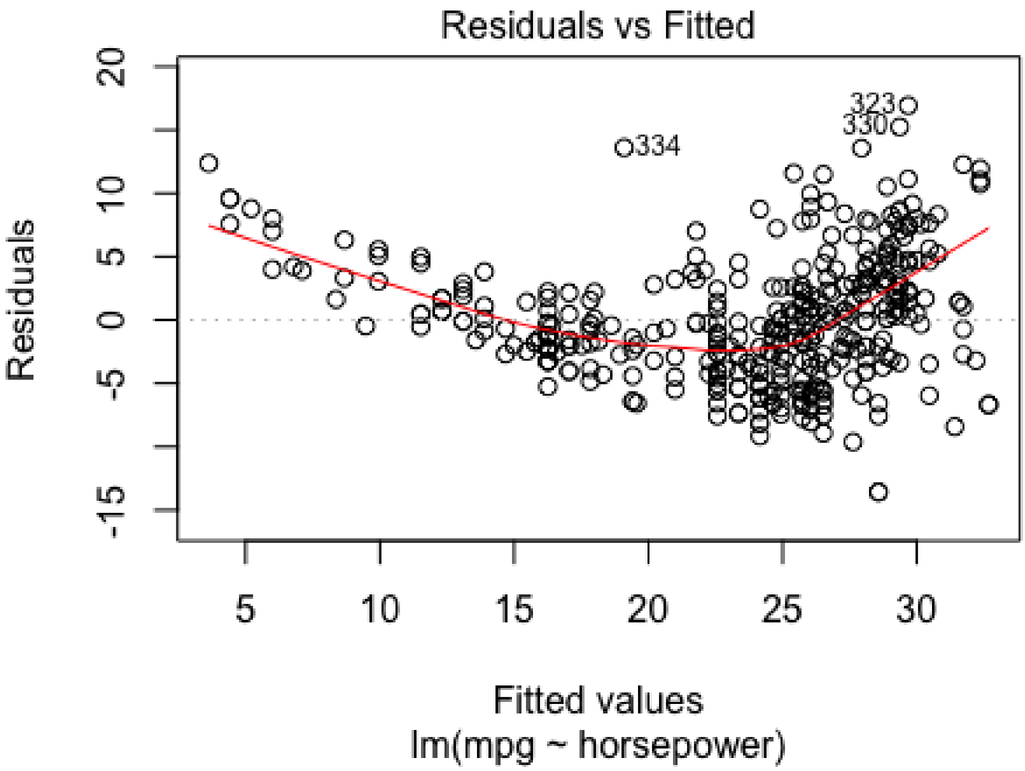
```
plot(Auto$horsepower, Auto$mpg, xlab = "Horsepower", ylab = "Miles per Gallon
", main = "Plot of MPG vs. Horsepower", col = "blue")
abline(regression, col = "red", lwd = 2)
```

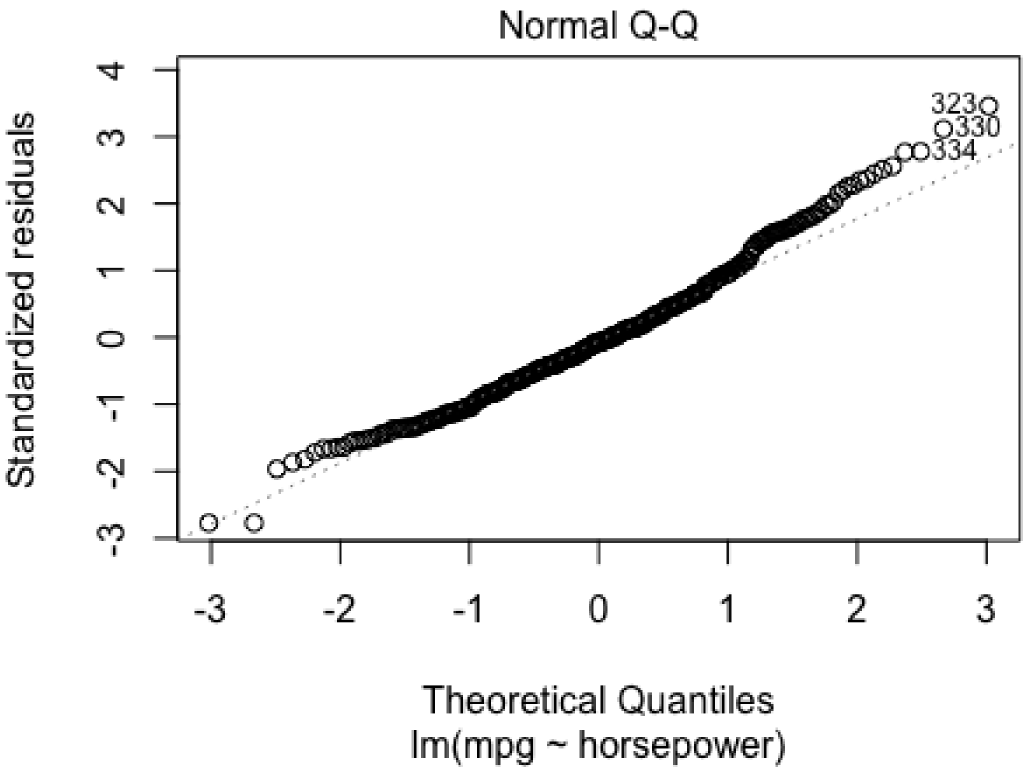
c)

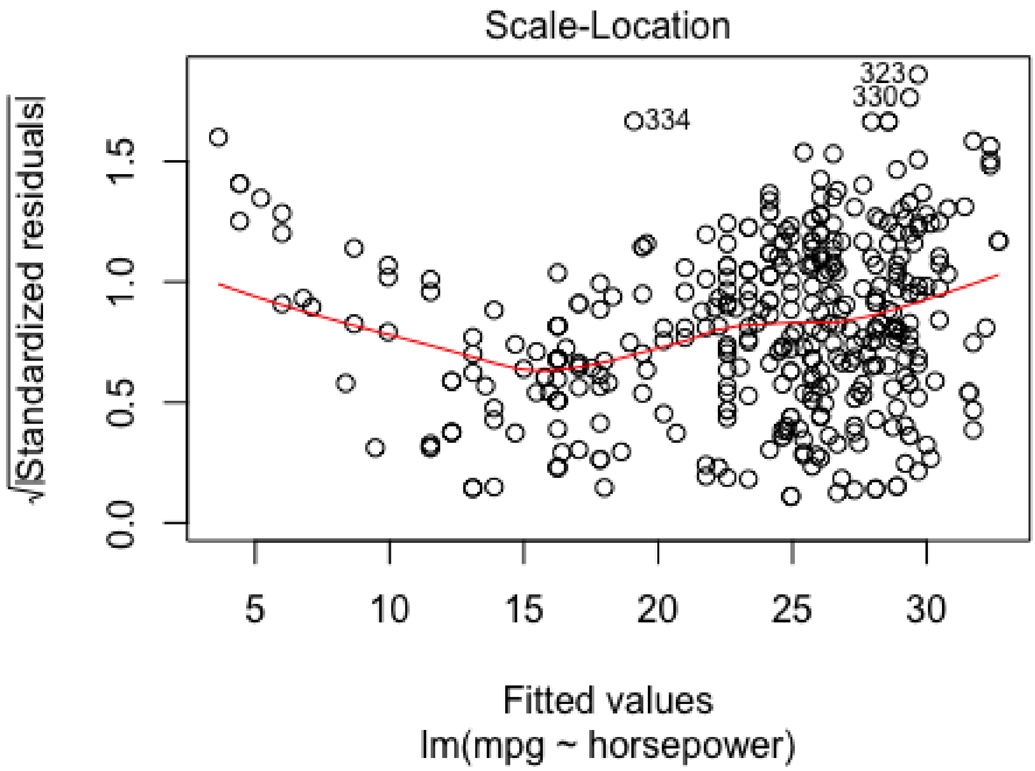
```
plot(regression
```

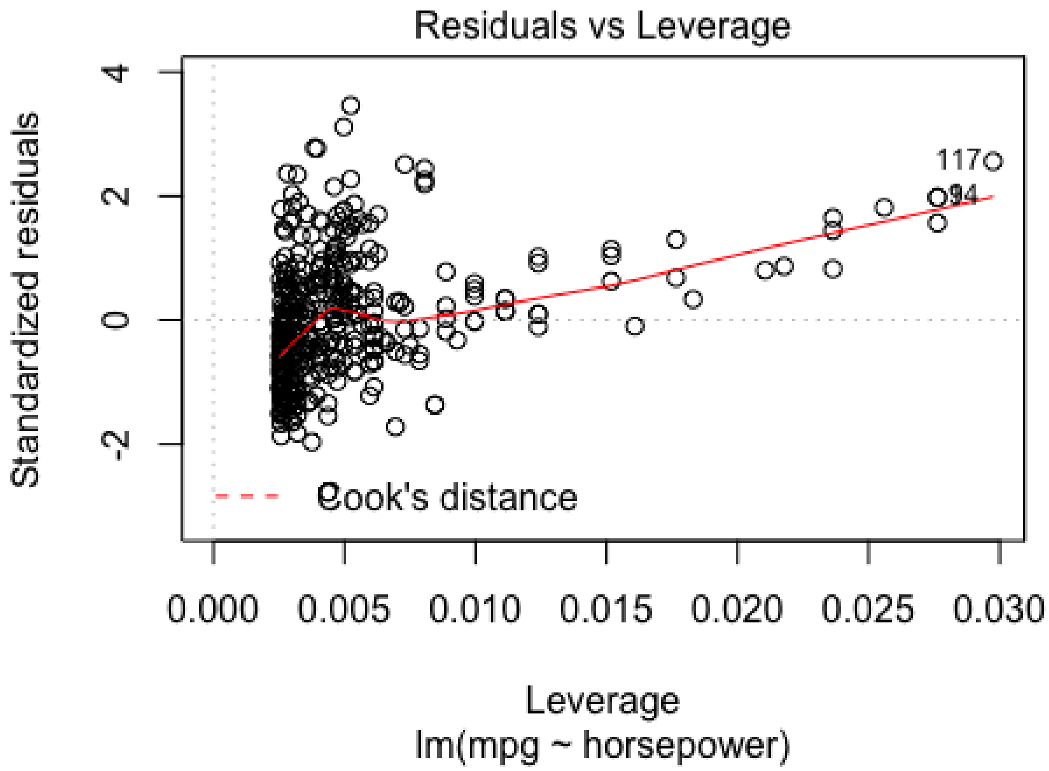


```
)
```


Residuals vs Fitted
Residuals
20
10
5
0
-5
-15
5
10
15
20
25
30
Fitted values
lm(mpg ~ horsepower)
334
323
330


Normal Q-Q
Standardized residuals
4
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
323
330
334
Theoretical Quantiles
lm(mpg ~ horsepower)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
5
10
15
20
25
30
Fitted values
lm(mpg ~ horsepower)
323
330
334



There could be some problems with the fit. For example, in the residuals vs. fitted plot, there is some non-linearity with the data. In the residuals vs. leverage plot, there are also some outliers and high leverage points that deviate from the line.

14.

a)

```
set.seed(2)
x1 <- runif(100)
x2 <- 0.5 * x1 + rnorm(100)/10 #true regression coefficients
y <- 2 + 2 * x1 + 0.3 *x2 + rnorm(100) #rnorm(100) to simulate error term
```

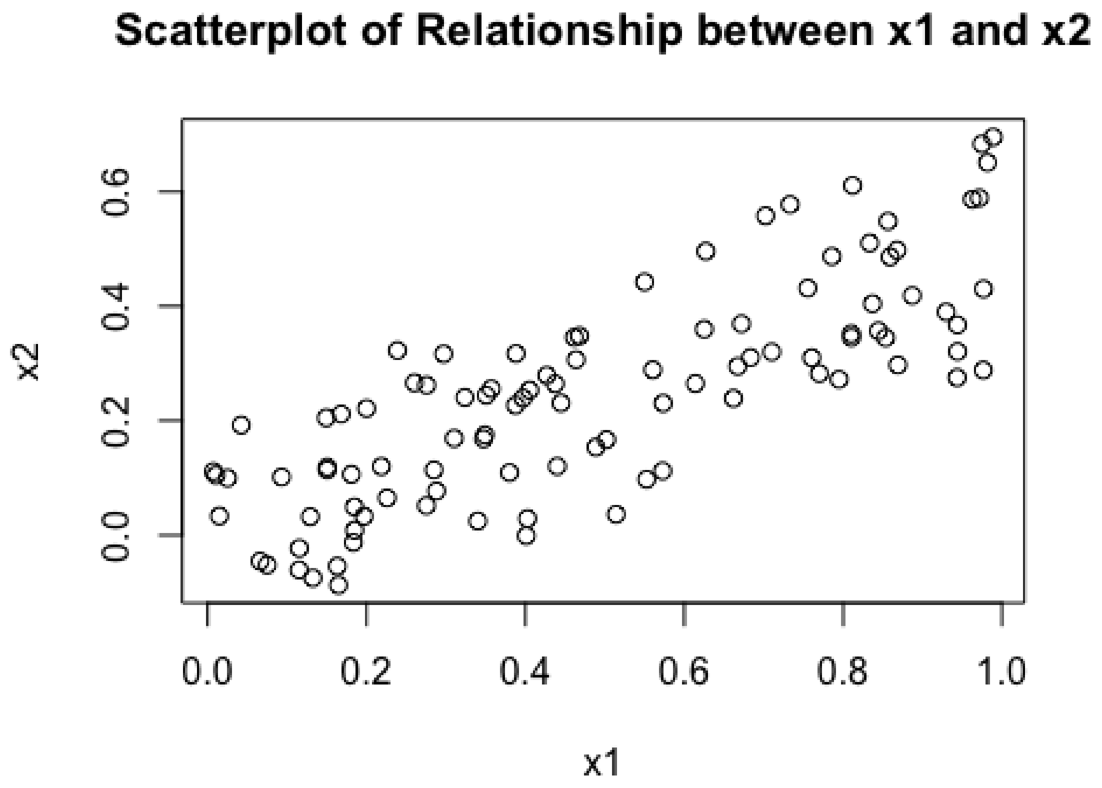
The equation of the linear model is y = 2 + 2X1 + 0.3X2 + e. The regression coefficients are 2, 2, and 0.3.

b)

```
cor(x1, x2)

## [1] 0.7974115

plot(x1, x2, main = "Scatterplot of Relationship between x1 and x2")
```



The correlation between x1 and x2 is 0.7974 which is a pretty strong positive correlation.

c)

```
ls.regression <- lm(y ~ x1 + x2)
summary(ls.regression)

##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05360 -0.75897 -0.02638  0.61691  2.92944
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.8136     0.1963   9.241 5.84e-15 ***
## x1            2.8577     0.5661   5.048 2.09e-06 ***
## x2           -0.5150     0.9122  -0.565    0.574
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.015 on 97 degrees of freedom
## Multiple R-squared:  0.3757, Adjusted R-squared:  0.3628
## F-statistic: 29.19 on 2 and 97 DF,  p-value: 1.192e-10
```

B0 = 1.8136, B1 = 2.8577, B2 = -0.5150. The p-value for B1 is less than 0.05, so we reject the null hypothesis. The p-value for B2 is greater that 0.05 so we do not reject the null hypothesis.

d)

```
ls.regression <- lm(y ~ x1)
summary(ls.regression)

##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0598 -0.7629 -0.0402  0.6129  2.9098
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.8111     0.1955   9.263 4.82e-15 ***
## x1            2.6028     0.3404   7.646 1.43e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.011 on 98 degrees of freedom
## Multiple R-squared:  0.3737, Adjusted R-squared:  0.3673
## F-statistic: 58.46 on 1 and 98 DF,  p-value: 1.431e-11
```

B1 = 2.6028 and the p-value is less than 0.05, so we can reject the null hypothesis.

e)

```
ls.regression <- lm(y ~ x2)
summary(ls.regression)

##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.26089 -0.70897 -0.04271  0.67001  3.04416
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3071     0.1903   12.13  < 2e-16 ***
## x2            3.1570     0.6154    5.13 1.46e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.134 on 98 degrees of freedom
## Multiple R-squared:  0.2117, Adjusted R-squared:  0.2037
## F-statistic: 26.32 on 1 and 98 DF,  p-value: 1.463e-06
```

B2 = 3.1570 and the p-value is also less than 0.05, so we can reject the null hypothesis.

f) The results appear like they contradict each other since we did not reject the null hypothesis when doing the least squares regression with both x1 and x2, but we rejected the hull hypothesis when only using x2. However this actually makes sense because x1 and x2 are highly correlated. When there is collinearity, the regression is less accurate because it is difficult to determine how each predictor individually affects the response.