STAT 425 Assignment 3

Due Sunday, March 7, 11:59pm. Submit through Moodle.

Name: (insert your name here)

Netid: (insert)

Submit your computational work both as an R markdown (*.Rmd) document and as a pdf, along with any files needed to run the code. Embed your answers to each problem in the document below after the question statement. If you have hand-written work, please scan or take pictures of it and include in a pdf file, ideally combined with your pdf output file from R Markdown.

Problem 1

The `fat` data in the `faraway` library contains age, weight, height, and various body circumference measurements for 252 men. The variables `brozek` and `siri` correspond to two different density related equations for percent body fat.

a) Fit a linear model to predict the `siri` percentage body fat from `age`, `weight`, `height`, `neck`, `chest`, `abdom`, `hip`, `thigh`, `knee`, `ankle`, `biceps`, `forearm`, and `wrist`. Display a model summary and state which of the variable coefficients are statistically significant at the 0.05 level.

Answer: We first fit the model as below.

```r
# load package
require('faraway')
```

```
## Loading required package: faraway
```

```r
lmod <- lm(siri ~ age + weight + height + neck + chest + abdom +
             hip + thigh + knee + ankle + biceps + forearm +
             wrist, data = fat)
summary(lmod)
```

```
##
## Call:
## lm(formula = siri ~ age + weight + height + neck + chest + abdom +
##     hip + thigh + knee + ankle + biceps + forearm + wrist, data = fat)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -11.1687  -2.8639  -0.1014   3.2085  10.0068 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -18.18849   17.34857  -1.048  0.29551    
## age           0.06208    0.03235   1.919  0.05618 .  
## weight       -0.08844    0.05353  -1.652  0.09978 .  
## height       -0.06959    0.09601  -0.725  0.46925    
## neck         -0.47060    0.23247  -2.024  0.04405 *  
## chest        -0.02386    0.09915  -0.241  0.81000    
## abdom         0.95477    0.08645  11.044  < 2e-16 ***
## hip          -0.20754    0.14591  -1.422  0.15622    
## thigh         0.23610    0.14436   1.636  0.10326    
## knee          0.01528    0.24198   0.063  0.94970    
## ankle         0.17400    0.22147   0.786  0.43285    
## biceps        0.18160    0.17113   1.061  0.28966    
## forearm       0.45202    0.19913   2.270  0.02410 *  
## wrist        -1.62064    0.53495  -3.030  0.00272 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.305 on 238 degrees of freedom
## Multiple R-squared:  0.749,  Adjusted R-squared:  0.7353 
## F-statistic: 54.65 on 13 and 238 DF,  p-value: < 2.2e-16
```

As we can see from the summary table, under a significance level of 5%, the coefficients of the circumferences of neck, abdom, forearm, and wrist are statistically significant.

b) Perform a model comparison F test to determine whether a reduced model using only the variables `age`, `weight`, `height`, `abdom`, `forearm`, and `wrist` to predict `siri` is appropriate. Is the null hypothesis rejected at level 0.05?

Answer: We can fit the reduced model, and perform the ANOVA F-test as below.

```r
submod <- lm(siri ~ age + weight + height + abdom + forearm +
               wrist, data = fat)
anova(submod, lmod)
```

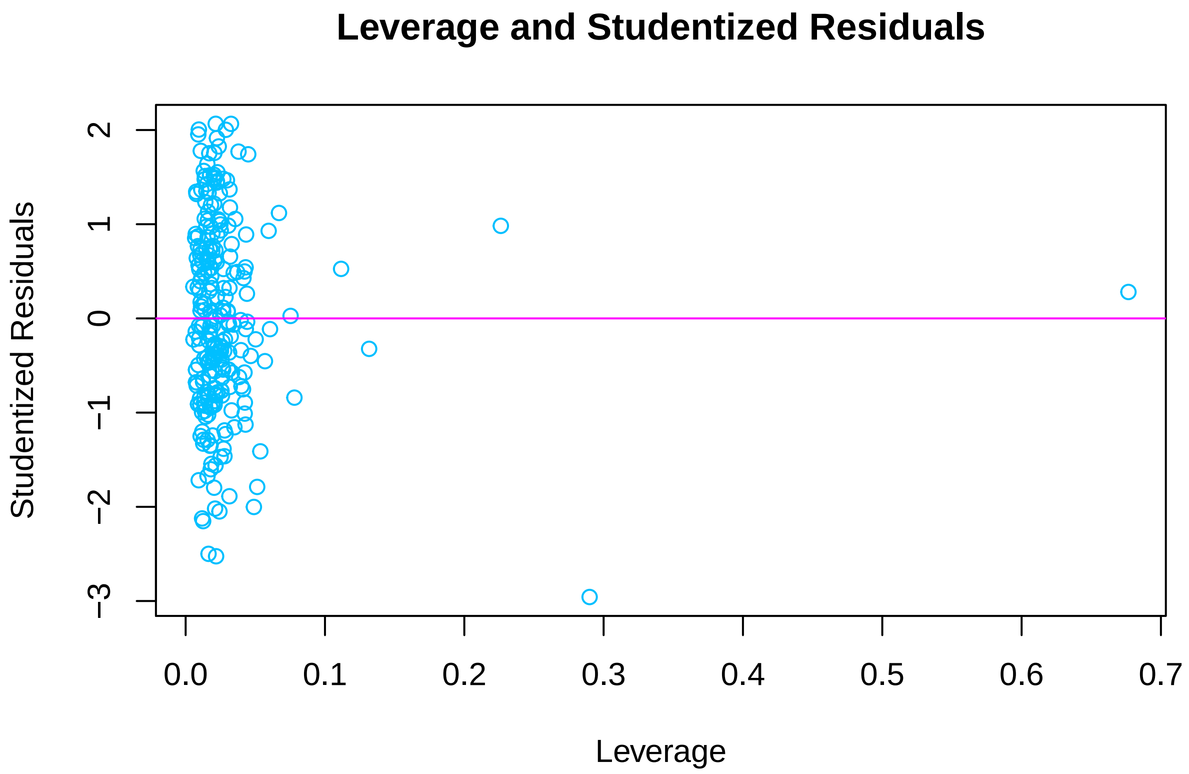
```
## Analysis of Variance Table
## 
## Model 1: siri ~ age + weight + height + abdom + forearm + wrist
## Model 2: siri ~ age + weight + height + neck + chest + abdom + hip + thigh + 
##     knee + ankle + biceps + forearm + wrist
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    245 4601.8
## 2    238 4411.4  7    190.34 1.467 0.1798
```

Under a significance level of 5%, with p-value 0.1798, we fail to reject the null hypothesis.

c) Now consider the reduced model from Part b). Plot studentized residuals on the y-axis versus leverage (diagonal of hat matrix) on the x-axis for this model. Include a horizontal line at height 0 for reference. Recall from our notes that the function `influence` provides diagonals of the hat matrix as influence(ModelObject)$hat. Are there any high leverage observations, and if so do they appear to be response outliers, or are they well fit by the model?

Answer: We first plot

```r
plot(x = influence(submod)$hat, y = rstudent(submod),
     xlab = 'Leverage', ylab = 'Studentized Residuals',
     col = 'deepskyblue',
     main = 'Leverage and Studentized Residuals')
abline(h = 0, col = 'magenta')
```



Using a rule of thumb of $2p/n$, where $n = 252$ and $p = 7$, we can find observations with high leverage as below.

```r
p <- ncol(model.matrix(submod))
n <- nrow(fat)
(hlobs <- which(influence(submod)$hat > 2 * p / n))
```

```
##  36  39  41  42 159 175 205 206 216 226 252
##  36  39  41  42 159 175 205 206 216 226 252
```

We can check whether they are outliers using Bonferroni correction as

```
m <- n
cv <- qt(.05 / (2 * m), df = df.residual(submod))
which(abs(rstudent(submod)[hlobs]) > abs(cv))
```

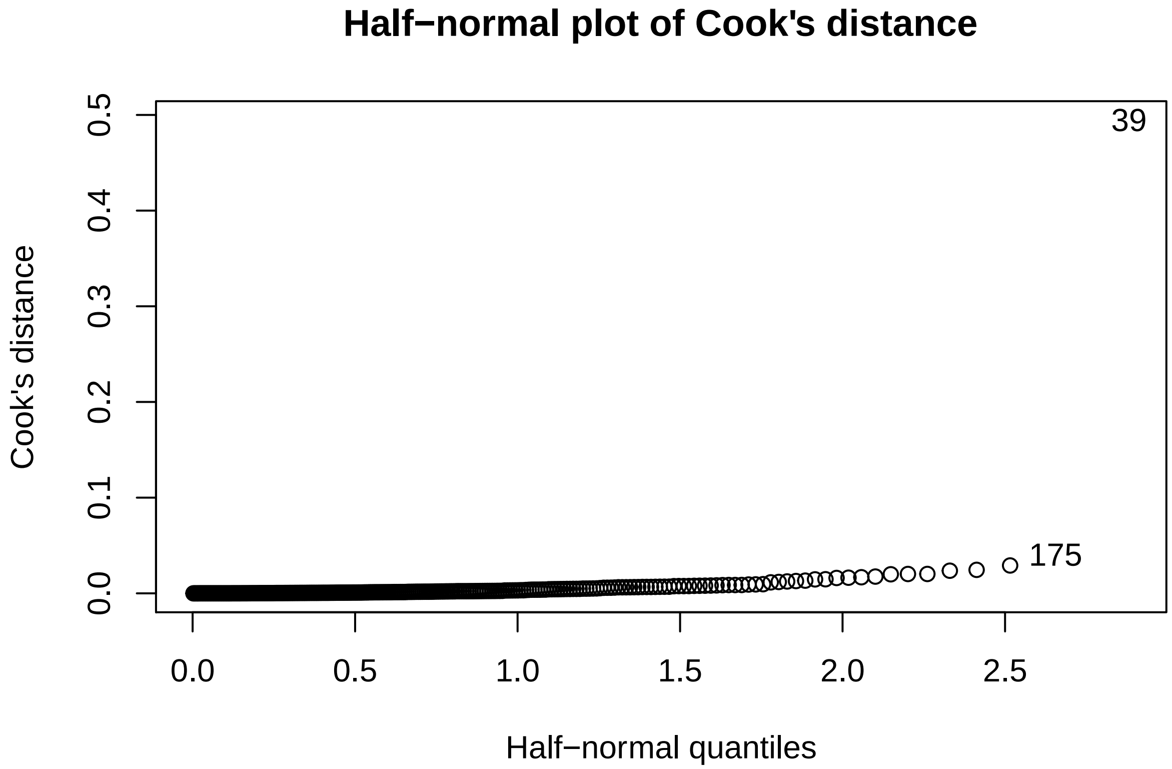
```
## named integer(0)
```

None of the observations is rejected as an outlier after Bonferroni adjustment for the sample size.

d) For the reduced model in parts b) and c), use a half-normal plot or sort function to display which observation has the largest Cook's Distance.

Answer: We can find the observation that has the largest Cook's distance as below.

```
cook <- cooks.distance(submod)
halfnorm(cook,  ylab = "Cook's distance",
         main = "Half-normal plot of Cook's distance")
```



Thus, the 39th observation has the largest Cook's distance.

e) For the same reduced model as in Part d), show the model summaries with and without the observation identified by Cook's Distance in Part d). Indicate which, if any, of the variables have significant coefficient t-tests in one but not the other of these two fitted models, at a significance level of 0.05.

Answer: We can check this result with

```
submod2 <- lm(siri ~ age + weight + height + abdom + forearm +
                wrist, data = fat[-39, ])
tab <- round(cbind(coef(summary(submod))[, 4],
      coef(summary(submod2))[, 4], NA), 4)
tab[, 3] <- ifelse((tab[, 1] >= 0.05 & tab[, 2] >= 0.05 |
                     tab[, 1] < 0.05 & tab[, 2] < 0.05) == TRUE,
                   yes = 0, no = 1)
colnames(tab) <- c('With', 'Without', 'Inconsistent')
tab
```

```
##               With Without Inconsistent
## (Intercept) 0.0094  0.0575            1
## age         0.1541  0.0927            0
## weight      0.0020  0.0704            1
## height      0.3275  0.1530            0
## abdom       0.0000  0.0000            0
## forearm     0.0069  0.1051            1
## wrist       0.0005  0.0002            0
```

As we can see, `weight` and `forearm` have significant coefficient t-tests in one but not the other of these two fitted models, at a significance level of 5%.

Problem 2

The `cheddar` data from the `faraway` library has data on samples of cheddar cheese. For each of 30 samples, a subjective `taste` score is given along with checmical composition measurements labeled `Acetic`, `H2S`, and `Lactic`.

a) Fit a multiple linear regression model for predicting `taste` from the three chemical composition measurements. Which of the variables have statistically significant coefficients at level 0.05?

Answer:

```
library(faraway)
fit = lm(taste ~ Acetic + H2S + Lactic, data = cheddar)
summary(fit)
```

```
##
## Call:
## lm(formula = taste ~ Acetic + H2S + Lactic, data = cheddar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
```

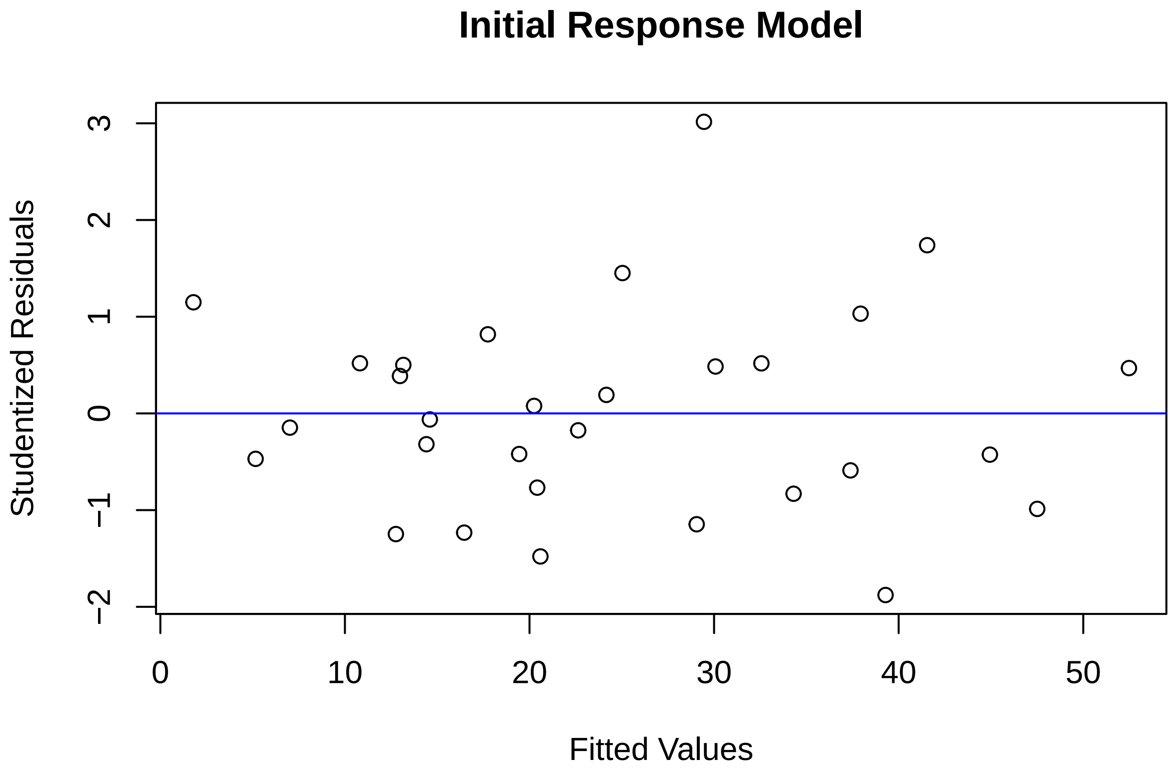
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
## Acetic        0.3277     4.4598   0.073  0.94198
## H2S           3.9118     1.2484   3.133  0.00425 **
## Lactic       19.6705     8.6291   2.280  0.03108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.13 on 26 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6116
## F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

H2S and Lactic have statistically significant coefficients at the the 0.05 significance level.

b) Plot studentized residuals versus fitted values for the model in a), including a horizontal line at height 0 for reference. Is there any evidence of outliers, heteroscedasticity or curvature in the plot? Explain briefly.

Answer:

```r
p = plot(fit$fitted.values, rstudent(fit),
     xlab = "Fitted Values",
     ylab = "Studentized Residuals",
     main = "Initial Response Model") +
  abline(h = 0, col = "blue")
```



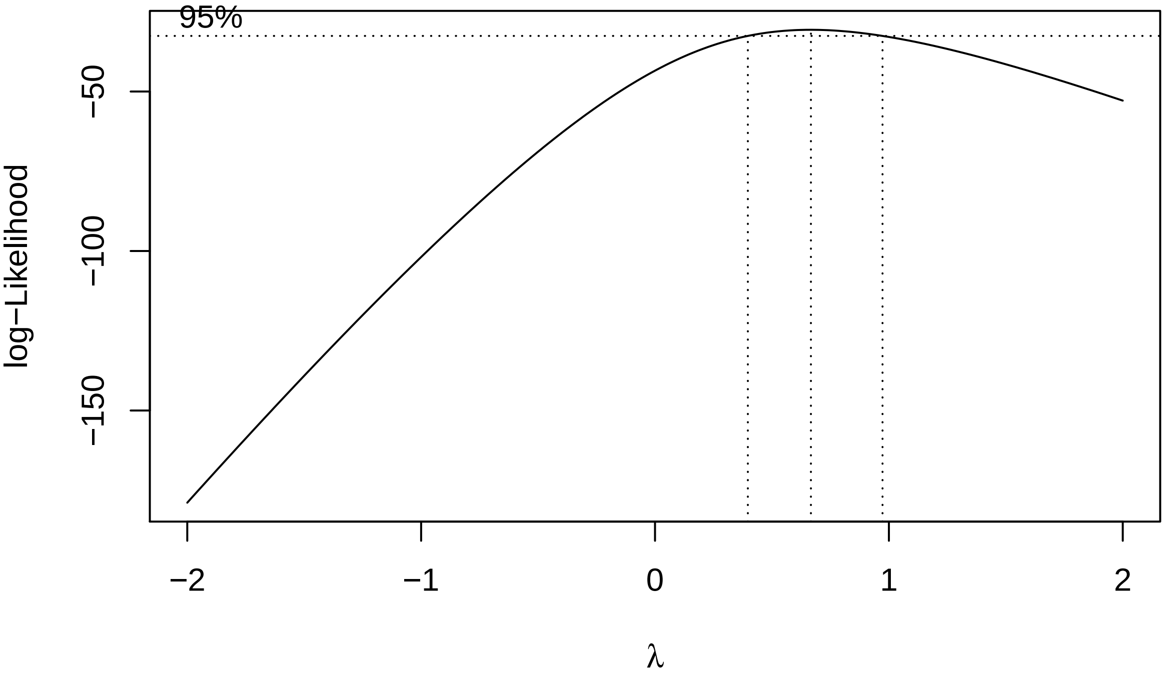
There seems to be a point with a studentized residual value of about 3 that may be an outlier.

The studentized residuals seem to have slightly higher variance with greater fitted values, so that may be evidence for some slight heteroscedasticity. Other than that, the points seem to be distributed fairly randomly, so there is no evidence for a clear curve in the plot.

c) Use the Box-Cox method to find an optimal transformation of the response. What is the optimal power $\hat{\lambda}$ of the response, based on the log-likelihood?

Answer:

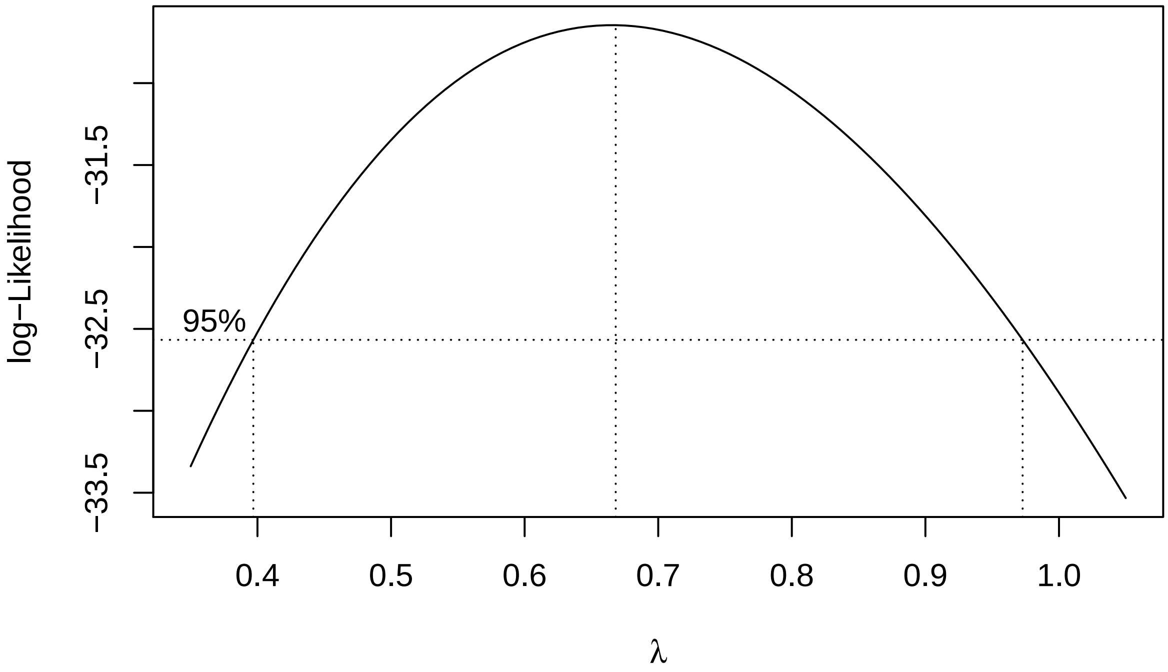
```
library(MASS)
trans = boxcox(fit, lambda=seq(-2, 2, length=400))
```



```
trans$x[trans$y == max(trans$y)]
```

```
## [1] 0.6666667
```

```
boxcox(fit,plotit=T,lambda=seq(0.35,1.05,by=0.1))
```



The optimal lambda value for the response is .6667 or 2/3.

d) Does the 95% confidence interval for the Box-Cox transformation parameter λ include $\lambda = 1$? What is your interpretation of this?

Answer:

```
tmp=trans$x[trans$y > max(trans$y) - qchisq(0.95, 1)/2]
range(tmp)
```

```
## [1] 0.4060150 0.9674185
```

This confidence interval doesn't include the value of lambda = 1, so an appropriate transformation is encouraged.

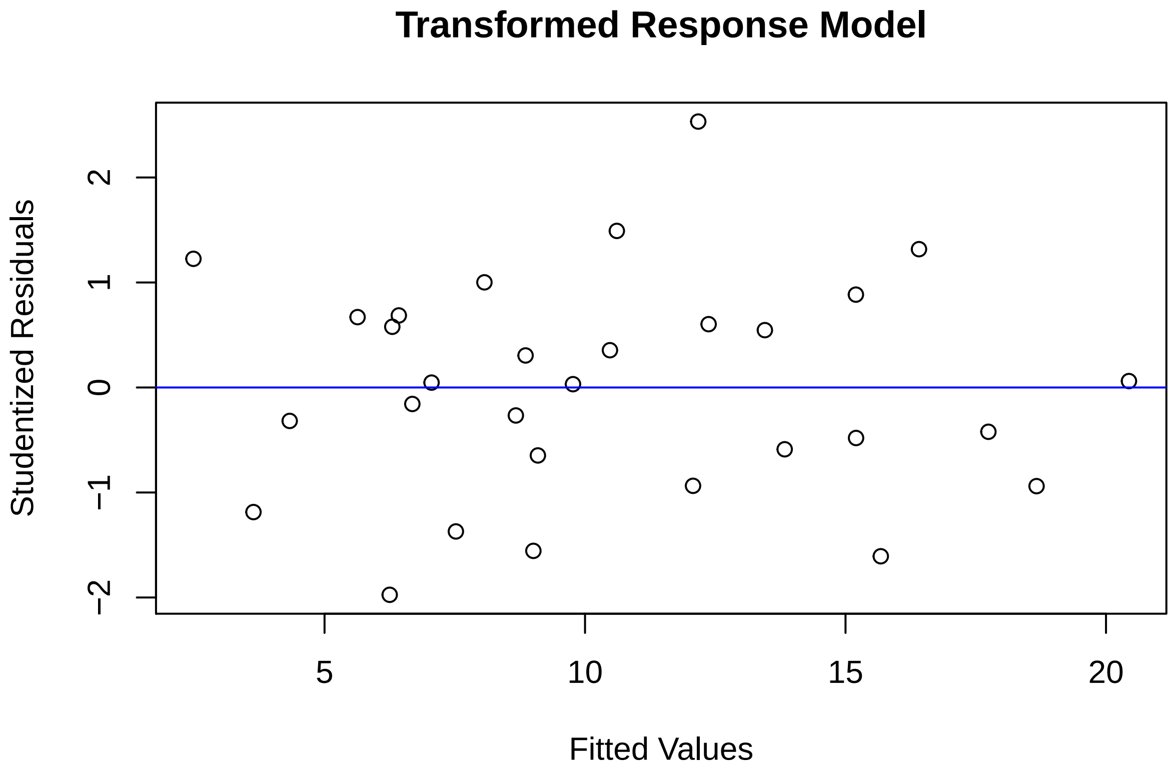
e) Use the optimal transformation found in c) to refit the data with the transformed response depending on the three chemical composition measurements. Plot studentized residuals versus fitted values for this transformed response model. Compare the graph qualitatively to the graph in Part b). Note: you can include a transformation of a variable in a formula by enclosing the transformation inside the function `I( )`, as in lm(I(yˆ3)~x, . . .).

Answer:

```
lambda = 2/3
fit_2 = lm(I((taste^(lambda) - 1)/lambda) ~ Acetic + H2S + Lactic, data = cheddar)

p2 = plot(fit_2$fitted.values, rstudent(fit_2),
     xlab = "Fitted Values",
     ylab = "Studentized Residuals",
     main = "Transformed Response Model") +
```

```r
abline(h = 0, col = "blue")
```



After the transformation, the single point that may have been an outlier seems less extreme. The variance across studentized residuals is fairly even regardless of the fitted values. This would suggest the transformed model has reduced heteroscedasticity when compared to the initial model. The curvature appears to be relatively the same.

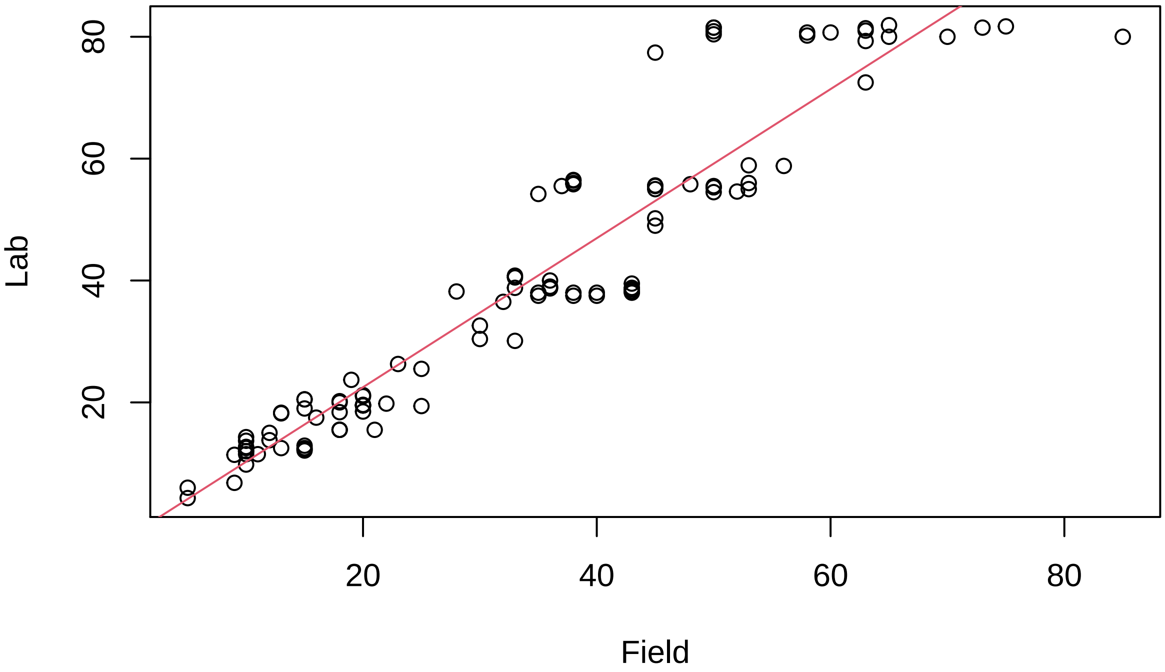
Problem 3:

The `pipeline` data in the `faraway` library consist of ultrasonic measurements of defects in the Alaska pipeline in the `Lab` and in the `Field`, i.e., on site.

a) Make a scatter plot of the `Lab` measurements versus the `Field` measurements, including the least squares regression line on the plot.

Answer:

```r
library("faraway")
model = lm(Lab ~ Field, data = pipeline)
plot(pipeline$Field, pipeline$Lab, xlab = "Field", ylab = "Lab")
abline(model$coefficients, col = 2)
```



b) Fit the linear regression of `Lab` on `Field` and show the model summary. How strong is the linear correlation between the lab and field measurements?

Answer:

```
summary(model)
```

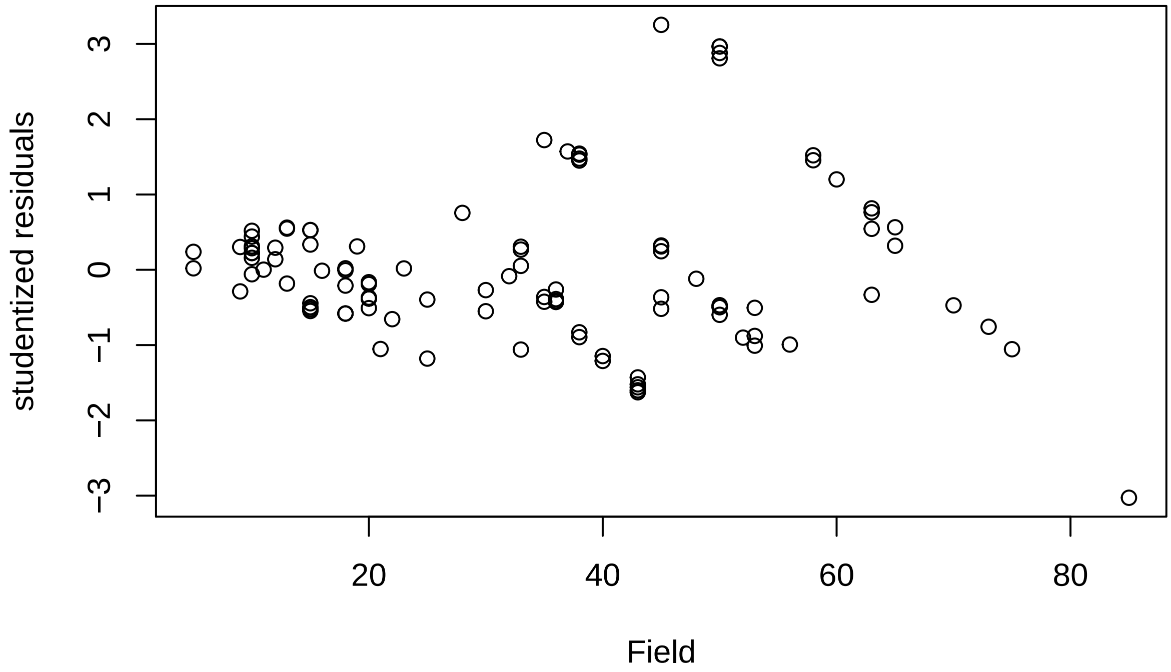
```
##
## Call:
## lm(formula = Lab ~ Field, data = pipeline)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.985  -4.072  -1.431   2.504  24.334
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.96750    1.57479  -1.249    0.214
## Field        1.22297    0.04107  29.778   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.865 on 105 degrees of freedom
## Multiple R-squared:  0.8941, Adjusted R-squared:  0.8931
## F-statistic: 886.7 on 1 and 105 DF,  p-value: < 2.2e-16
```

The multiple R^2 is 0.8941 according the model summary. We can see the linear correlation is pretty strong between the lab and field.

c) Plot studentized residuals versus `Field` for the model in b). Does the graph show any evidence of heteroscedascticity or curvature? Describe briefly.

Answer:

```
plot(pipeline$Field, rstudent(model),
     xlab = "Field",
     ylab = "studentized residuals")
```



We somehow can see the trend here but apparently the variance of error is not constant according to the plot. This suggests heteroscedasctic error variance.

d) Fit a linear model regressing the log of `Lab` on the log of `Field`. Show the model summary and compare the linear correlation with that of the model in b).

Answer:

```
modellog = lm(I(log(Lab))~I(log(Field)), data = pipeline)
summary(modellog)
```

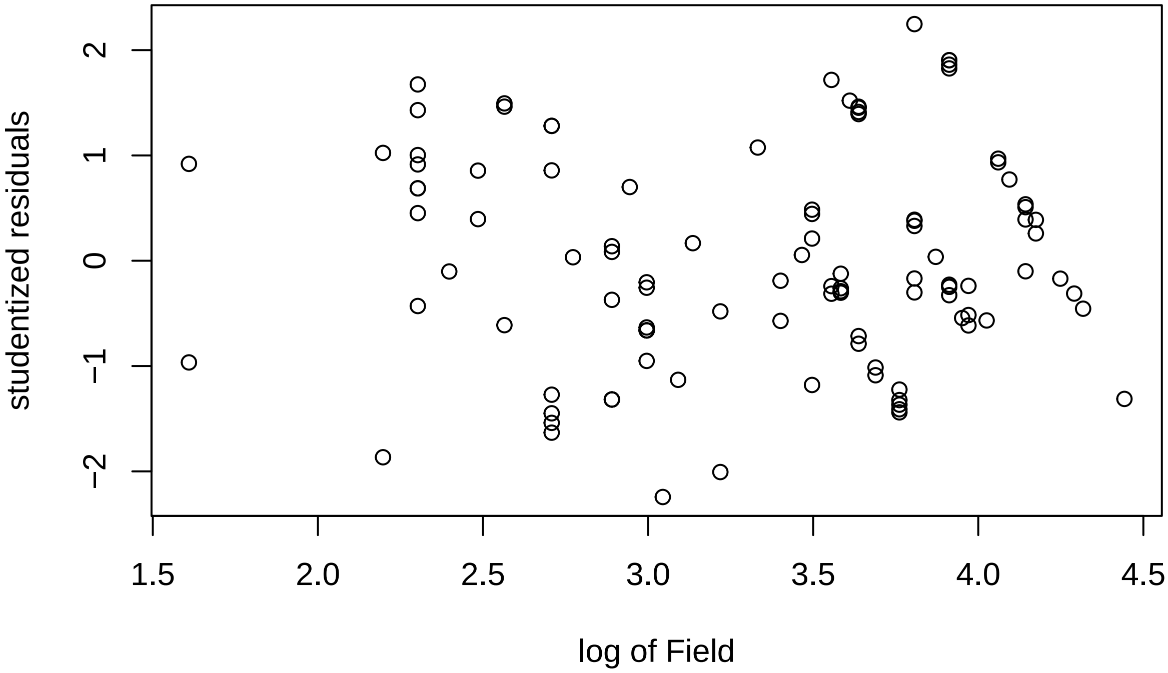
```
##
## Call:
## lm(formula = I(log(Lab)) ~ I(log(Field)), data = pipeline)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40212 -0.11853 -0.03092  0.13424  0.40209
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.06849    0.09305  -0.736    0.463
## I(log(Field))  1.05483    0.02743  38.457   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1837 on 105 degrees of freedom
## Multiple R-squared:  0.9337, Adjusted R-squared:  0.9331
## F-statistic:  1479 on 1 and 105 DF,  p-value: < 2.2e-16
```

The multiple R^2 value 0.9337 is larger than 0.8941. We can see that the linear correlation here is stronger.

e) For the model in d), plot studentized residuals versus the log of `Field`. Does this graph show any evidence of heteroscedascticity or curvature? Describe briefly.

Answer:

```
plot(log(pipeline$Field), rstudent(modellog),
     xlab = "log of Field",
     ylab = "studentized residuals")
```



The graph looks much better now. No obvious curvature is shown and the variance of error seems to be consistent.