Exam 2

STAT 021

Swarthmore College

2019/11/22

Name: _____

Instructions:

There are five questions on this exam and they are meant to be organized by the amount of time each question may take with the longer problems at the very end. The points allotted for each question are given at the end of the problem. Please don't write an entire page response for any of the answers. Rather, answer these questions to the best of your ability with succinct, informative statements or observations. You may or may not use the following formulas and definitions. The R output for Problems 3, 4 and 5 is provided at the very end for ease of reference (you can tear these sheets off if you'd like).

Formulas and Definitions To standardize a vector of numerical variables (say $w_1, w_2, \ldots w_n$) we subtract the mean and divide by the standard deviation for each value; e.g. $z_i = \frac{w_i - a}{b}$, where

$$a = \bar{w} = \frac{1}{n}\sum_{i=1}^{n} w_i \quad \text{and} \quad b = \frac{1}{n-1}\sum_{i=1}^{n}(w_i - \bar{w})^2.$$

Linear model with p predictor variables: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$.

In the model above, if we assume that the mean of ϵ is 0 and the variance of ϵ is some unknown number, σ^2, then the mean of the random variable Y is $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ and the variance of Y is σ^2.

Fitted/estimated model: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$

In the fitted model above, the values for each $\hat{\beta}_j$, $j = 1, \ldots, p$, are functions of Y that minimize the distance from the regression line (plane) to the data points (y_{obs}).

Definition of residuals: $\hat{y} - y = e$

Regression model sums of squares: $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

Residual sums of squares: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Total sums of squares: $\sum_{i=1}^{n}(y_i - \bar{y})^2$

Relationship among the sums of squares terms: $SS_{tot} = SS_{reg} + SS_{res}$

The sums of squares terms are used to calculate the following statistics:

$$\hat{\sigma} = \sqrt{\frac{SS_{res}}{n-2}}$$

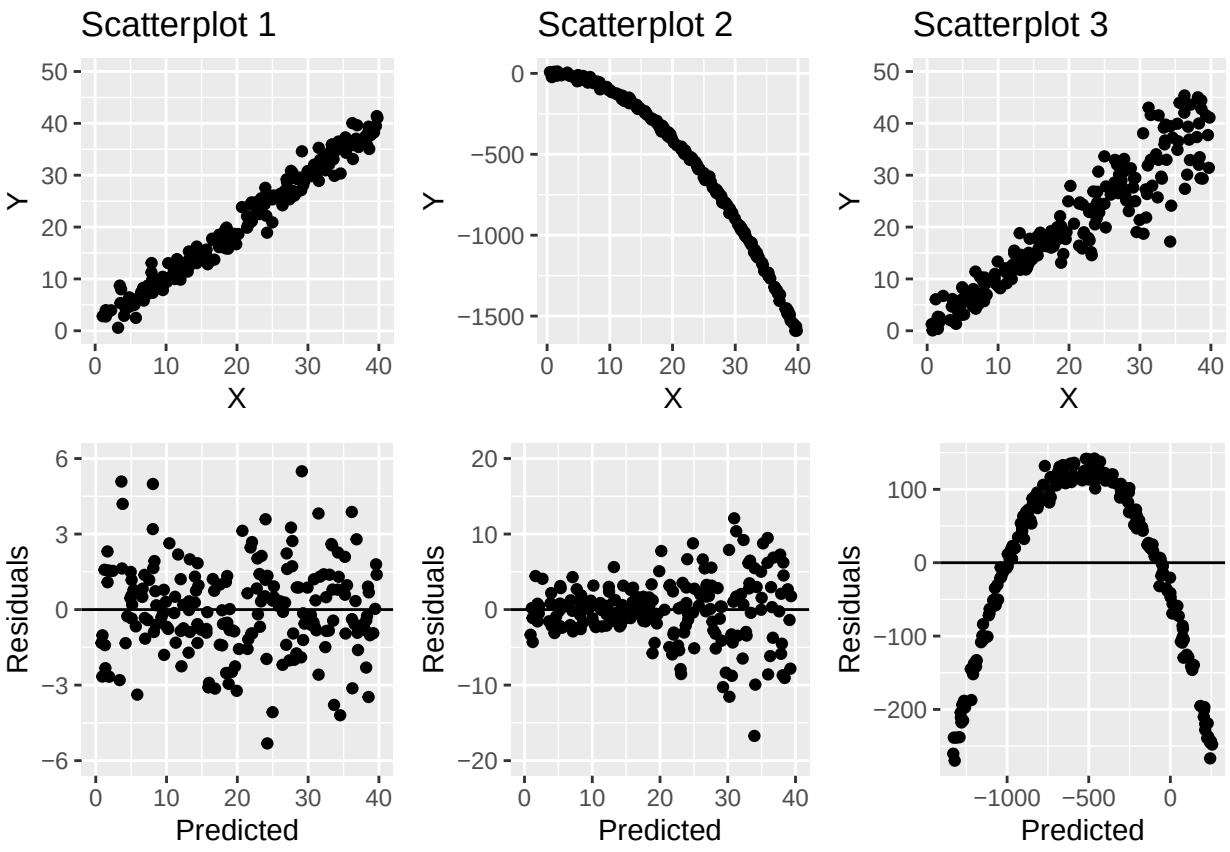
$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}}$$

We can calculate the correlation between two vectors $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ with the formula

$$Cor(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Problem 1

Below are three scatter plots for SLR models and three residual plots. Draw three lines connecting each scatterplot from the top row to the corresponding residual plot in the bottom row. (10 points)



Problem 2

The R-output below contains the results from a regression model fit to a dataset concerning life expectancy in years, alcohol consumption in drinks per day, and smoking status (smoker - Yes vs. non-smoker - No). Answer the following questions pertaining to the R output below. (24 points total)

```
## Call:
## lm(formula = Lifespan ~ Alcohol + Smoker)

## Residuals:
## Min 1Q Median 3Q Max
## -30.796 -7.139 0.125 6.949 19.578

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 93.6818  2.1024     44.56   <2e-16 ***
## Alcohol     -3.2656  0.3147    -10.38   <2e-16 ***
## SmokerYes  -23.4392  1.9922    -11.77   <2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 9.959 on 97 degrees of freedom
## Multiple R-squared: 0.7129, Adjusted R-squared: 0.7069
## F-statistic: 120.4 on 2 and 97 DF, p-value: < 2.2e-16
```

a) How does drinking alcohol affect average life expectancy? (4 points)

b) Suppose someone consumpes about 2.5 alcoholic dreaks per day and smokes regularly. Calculate (provide the formula for) the average life expectancy for this individual. (4 points)

c) What is the average difference in lifespan between smokers and non-smokers? (4 points)

d) Supposing we fix the amount of alcohol consumption; is there a statistically significant relationship between life expectancy and smoking status? Justify your answer. (4 points)

e) How much of the variation in lifespan can be explained by smoking status and alcohol consumption? (4 points)

f) Which of the questions above are questions of statistical inference? (4 points)

Problem 3

Let's examine the impact of Vitamin C from two sources and at two different dosages on the growth of teeth in Guinea pigs. The variables in the dataset are:

- *len* - the length of growth in the teeth
- *supp.VC* - a binary categorical variable that is 1 if the Vitamin C supplement type is ascorbic acid and is 0 if the Vitamin C supplement type is orange juice.
- *dose.2.0* - is a binary categorical variable that is 1 if the amount of Vitamin C is 2.0 mg and is 0 if the amount of Vitamin C is 1.0 mg.

Reference the R ouput corresponding to Problem 3 at the end of this exam to answer these three questions. (16 points total)

a) Write the down the main effects equations (in numbers) for predicting the average tooth growth for Guinea pics who:

   1) Have supplement type absorbic acid and dosage of 1.0 mg (1 point)
   2) Have supplement type absorbic acid and dosage of 2.0 mg (1 point)
   3) Have supplement type orange juice and dosage of 1.0 mg (1 point)
   4) Have supplement type orange juice and dosage of 2.0 mg (1 point)

Problem 3 (continued)

b) Write the down the interaction effects equations (in numbers) for predicting the average tooth growth for Guinea pics who:

1) Have supplement type absorbic acid and dosage of 1.0 mg (1 point)
2) Have supplement type absorbic acid and dosage of 2.0 mg (1 point)
3) Have supplement type orange juice and dosage of 1.0 mg (1 point)
4) Have supplement type orange juice and dosage of 2.0 mg (1 point)

c) Based on the R output, which model do you think is a better choice, the one **with** interactions or the one **without** interactions? Justify your answer. (4 points)

d) For whichever model you chose in part (c), explain the effect of supplement type variable on tooth growth and try to use words that are understandable to someone who doesn't know statistics. (4 points)

Problem 4

For **problems 4 and 5**, we are trying to understand how the aerial biomass (response variable) production of a certain type of marsh grass is related to the three predictor variables pH, which measures the pH level of the soil (from 0-14), K, which measures the potassium level of the soil (in ppm), and the categorical variable *location* which can be one of three different spots ("OI" is short for Oak Island - this is the reference level, "SI" is short for Smith Island, and "SM" is short for Snows Marsh). Aerial biomass is measured with units gm^{-2}.

Consider the fit of a MLR regression model that includes the main effects of location, pH, and K with the response. Reference the R ouput and plots corresponding to Problem 4 at the end of this exam to answer these three questions. (30 points total)

a) What is the average (absolute) difference in biomass (in gm^{-2}) for moss found in Smith Island versus moss found in the Snows Marsh? (13 points)

b) What is the average (absolute) difference in biomass (in gm^{-2}) if the potassium (variable K) in the soil increases by 500 ppm? (13 points)

c) Based on this output, is it correct to say that the effect of changing pH level of the soil has a greater impact on the biomass than changing the potassium levels? Why or why not? (4 points)

Problem 5

Suppose instead of looking at the relationship between biomass and the predictor variables, we decide to look at the relationship between **the logarithm of the biomass** and the predictor variables. Reference the R ouput and plots corresponding to Problem 5 at the end of this exam to answer these questions. As a reminder, the logarithmic function is defined by the relationship that if $y = log_a(x)$ then $a^y = x$ (and vice versa). (30 points total)

a) Does this model look better or worse compared to the model in Problem 3? Justify your answer. (13 points)

b) What is the average difference in the response variable based on this linear regression model when the potassium in the soil increases by 500 ppm? (4 points)

c) Are the answers to part (b) of Problem 5 and part (b) of Problem 4 directly comperable? Why or why not? (13 points)

R output for Problem 3

```
teeth_SLR_main<-lm(len ~ supp + dose)
summary(teeth_SLR_main)

## Call:
## lm(formula = len ~ supp + dose)

## Residuals:
##    Min     1Q Median     3Q    Max
## -6.697 -2.765 -1.005  2.178  9.262

## Coefficients:
##                     Estimate Std. Error t value   Pr(>|t|)
## (Intercept)          21.197      1.061  19.976    < 2e-16 ***
## supp.aa              -2.925      1.225  -2.387     0.0222 *
## dose.2.0              6.365      1.225   5.195   7.72e-06 ***
## ---
## Signif. codes:  0 Õ***Õ 0.001 Õ**Õ 0.01 Õ*Õ 0.05 Õ.Õ 0.1 Õ Õ 1

## Residual standard error: 3.875 on 37 degrees of freedom
## Multiple R-squared:  0.469,  Adjusted R-squared:  0.4403
## F-statistic: 16.34 on 2 and 37 DF,  p-value: 8.202e-06
```

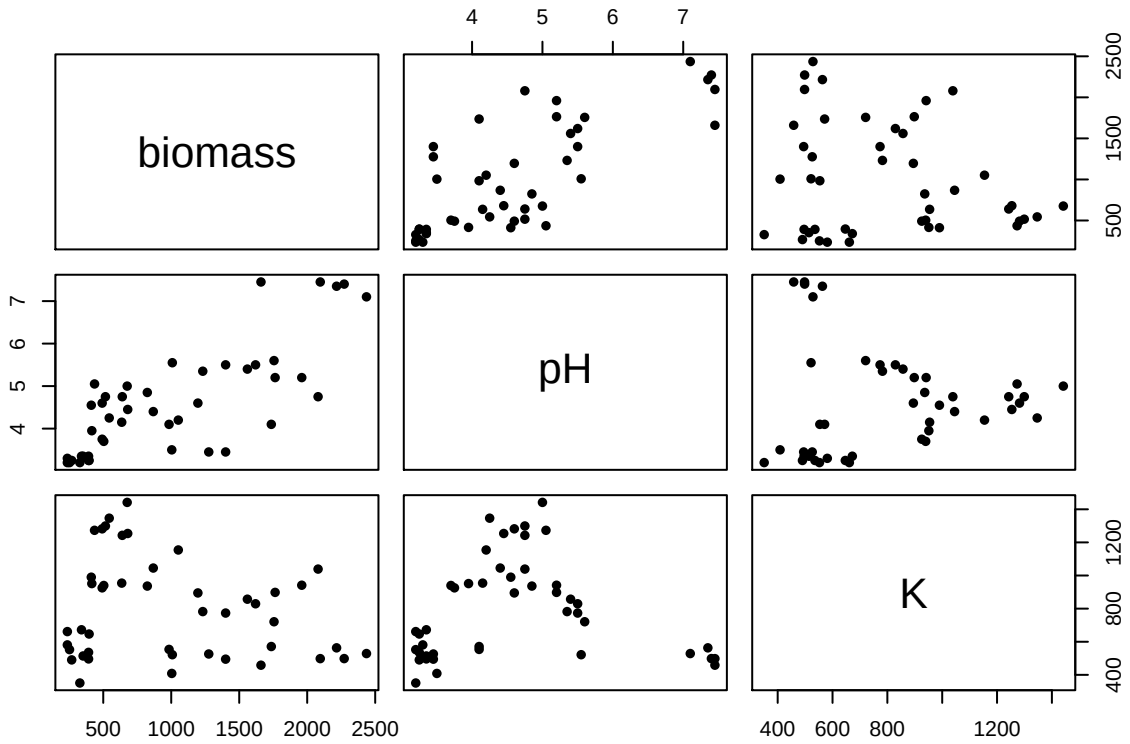
```
teeth_SLR_int<-lm(len ~ supp + dose + supp*dose)
summary(teeth_SLR_int)

## Call:
## lm(formula = len ~ supp * dose)

## Residuals:
##    Min     1Q Median     3Q    Max
## -8.200 -2.337 -0.005  2.147  7.760

## Coefficients:
##                         Estimate Std. Error   t value   Pr(>|t|)
## (Intercept)               22.700      1.137    19.969    < 2e-16 ***
## supp.aa                   -5.930      1.608    -3.689    0.00074 ***
## dose.2.0                   3.360      1.608     2.090    0.04374 *
## supp.aa:dose.2.0           6.010      2.274     2.643    0.01208 *
## ---
## Signif. codes:  0 Õ***Õ 0.001 Õ**Õ 0.01 Õ*Õ 0.05 Õ.Õ 0.1 Õ Õ 1

## Residual standard error: 3.595 on 36 degrees of freedom
## Multiple R-squared:  0.5553, Adjusted R-squared:  0.5183
## F-statistic: 14.99 on 3 and 36 DF,  p-value: 1.717e-06
```

R output for the model in Problem 4

```
biomass <- read_table2("~/Google Drive Swat/Swat docs/Stat 21/Data/biomass_data.txt", col_names = TRUE)
```

```
## Parsed with column specification:
## cols(
##   ID = col_double(),
##   Location = col_character(),
##   Type = col_character(),
##   BIO = col_number(),
##   SAL = col_double(),
##   pH = col_double(),
##   K = col_number(),
##   Na = col_number(),
##   Zn = col_double()
## )
```

```
biomass %>% select(BIO, pH, K) %>% pairs(labels=c("biomass","pH","K"),pch=16)
```




```
cor(biomass$pH, biomass$K)
```

```
## [1] 0.01922804
```

```
as_factor(biomass$Location)
```

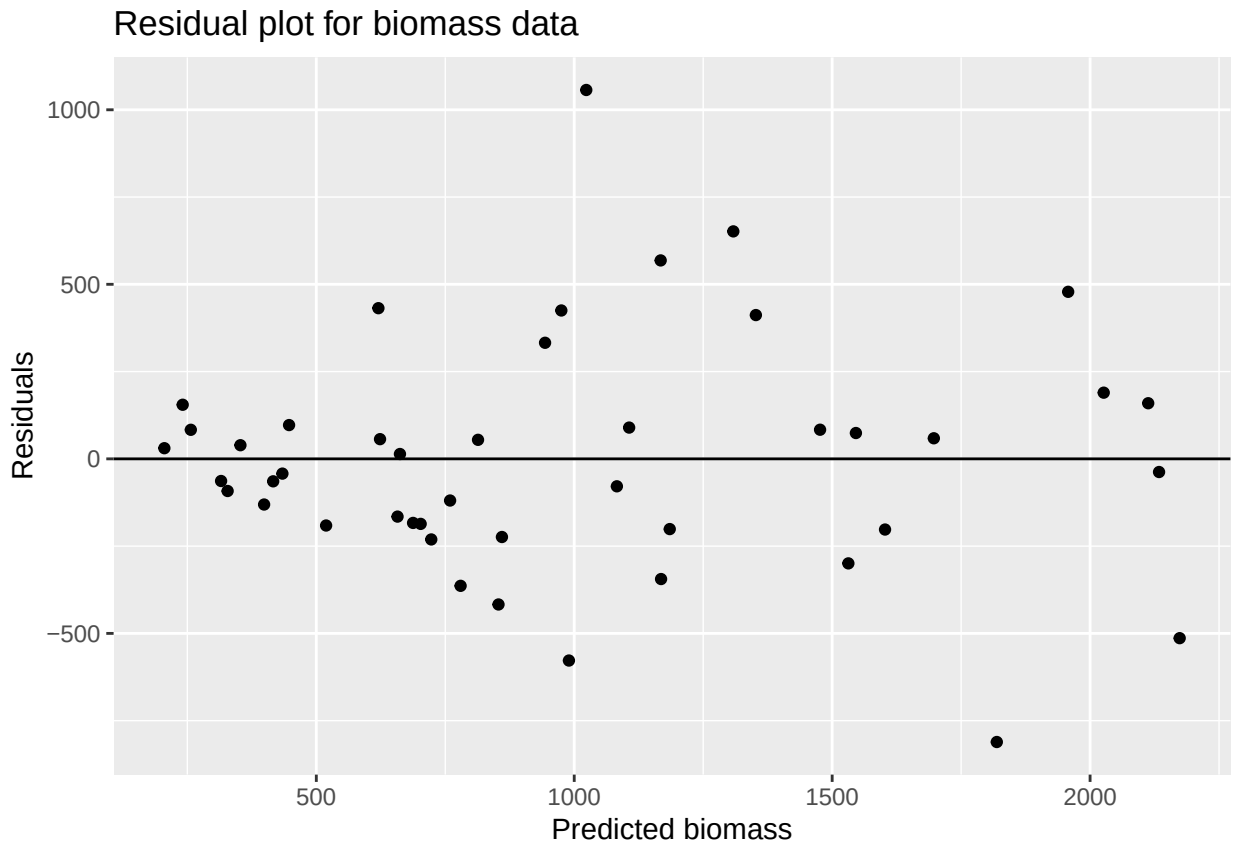
```
##  [1] OI OI OI OI OI OI OI OI OI OI OI OI OI OI OI OI SI SI SI SI SI SI SI SI
## [24] SI SI SI SI SI SI SI SM SM SM SM SM SM SM SM SM SM SM SM SM SM SM SM
## Levels: OI SI SM
```

R output for the model in Problem 4 (continued)

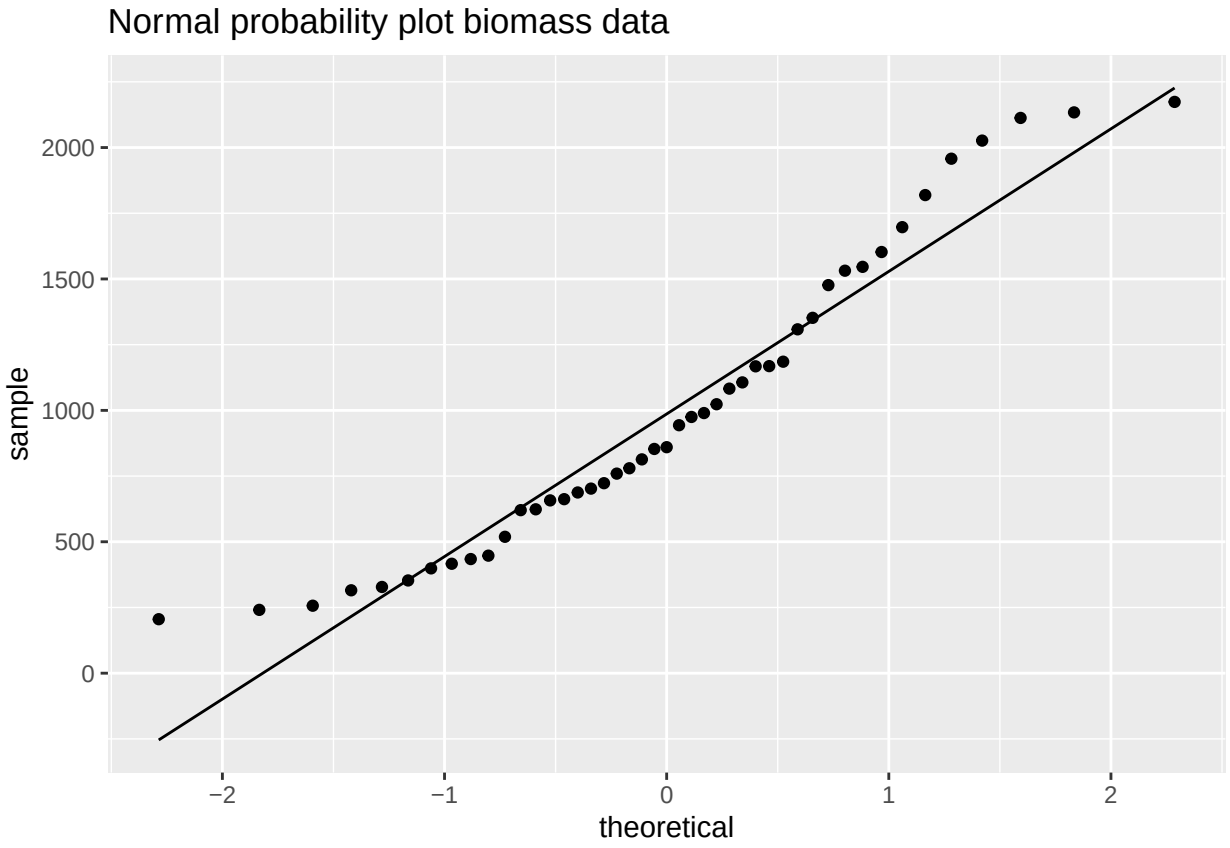
```
MLR_bio <- lm(BIO ~ as_factor(Location) + pH + K , data = biomass)
MLR_bio_sum <- summary(MLR_bio)
MLR_bio_sum
```

```
##
## Call:
## lm(formula = BIO ~ as_factor(Location) + pH + K, data = biomass)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -811.15 -190.99  -37.70   96.78 1056.67
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              43.0122   299.1682   0.144   0.8864
## as_factor(Location)SI -497.6490   163.4252  -3.045   0.0041 **
## as_factor(Location)SM   58.1814   131.6870   0.442   0.6610
## pH                     414.9021    43.3381   9.574 6.68e-12 ***
## K                       -1.0095     0.2324  -4.344 9.32e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 354.2 on 40 degrees of freedom
## Multiple R-squared:  0.7382, Adjusted R-squared:  0.712
## F-statistic: 28.19 on 4 and 40 DF,  p-value: 3.613e-11
```

R output for the model in Problem 4 (continued)

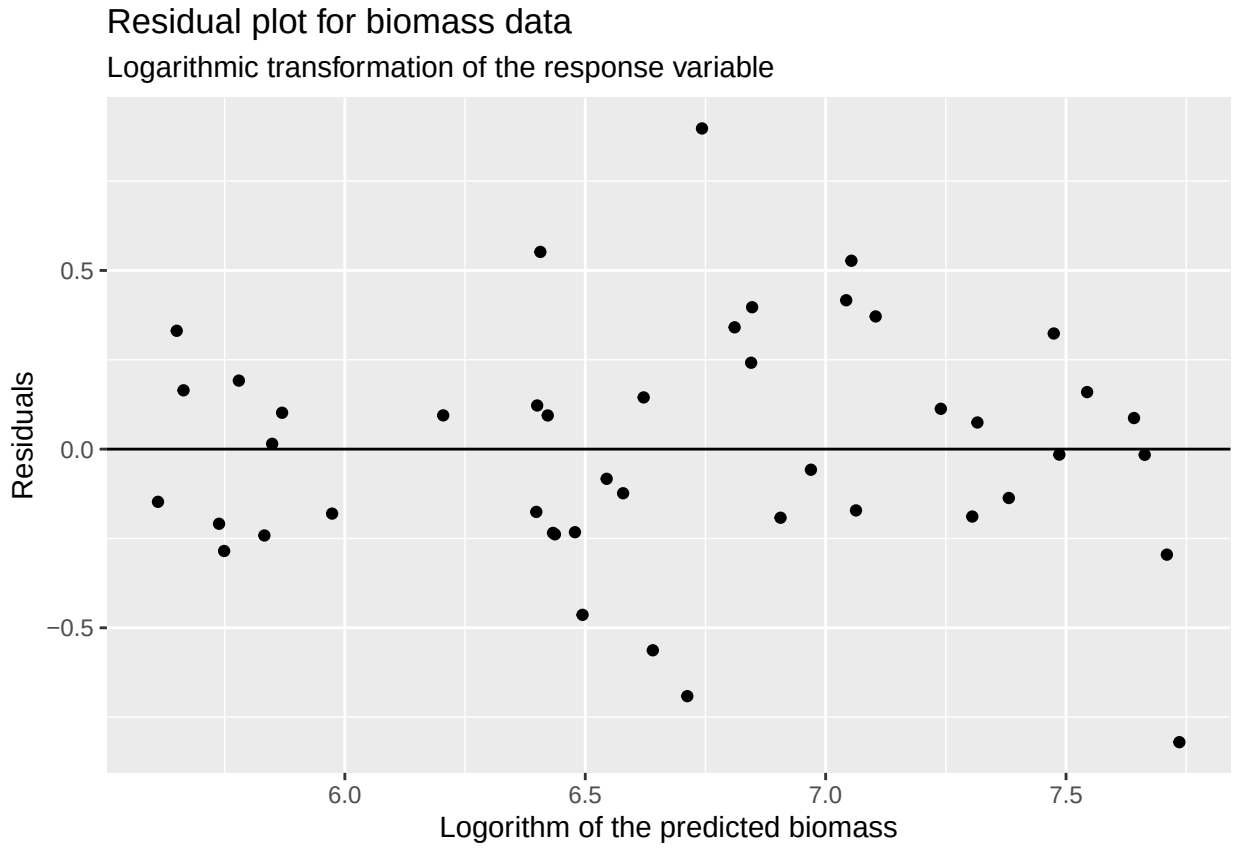


R output for the model in Problem 4 (continued)



R output for the model in Problem 5

```
##
## Call:
## lm(formula = log_biomass ~ as_factor(Location) + pH + K, data = biomass2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82000 -0.19199 -0.01526  0.16459  0.89733
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.9125341  0.2917904  20.263  < 2e-16 ***
## as_factor(Location)SI -0.9320328  0.1593950  -5.847 7.75e-07 ***
## as_factor(Location)SM -0.0394140  0.1284395  -0.307    0.761
## pH                       0.4380929  0.0422694  10.364 6.81e-13 ***
## K                       -0.0011660  0.0002267  -5.144 7.47e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3455 on 40 degrees of freedom
## Multiple R-squared:  0.7868, Adjusted R-squared:  0.7655
## F-statistic:  36.9 on 4 and 40 DF,  p-value: 6.309e-13
```



R output for the model in Problem 5 (continued)