

Exam 2

STAT 021

Swarthmore College

2019/11/22

Name: _____

Instructions:

There are seven questions on this exam. 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.

Formulas and Definitions To standardize a vector of numerical variables (say w_1, w_2, \dots, 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: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$.

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 $\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$ and the variance of Y is σ^2 .

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

In the fitted model above, the values for each $\hat{\beta}_j$, $j = 1, \dots, p$, 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, \dots, x_n) and (y_1, \dots, 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

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). (35 points)

```
## 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
```

Answer the following questions pertaining to the output above.

- 1) How does drinking alcohol affect average life expectancy?

Each additional drink per day decreases life expectancy by 3.2656 years on average.

- 2) Suppose the mean alcohol consumption in both groups combined is 2.5 drinks per day. Calculate (provide the formula for) the estimate of the average life expectancy for **non-smokers**.

$$93.6818 - 3.2656 \times 2.5 = 85.5178$$

- 3) What is the average difference in lifespan between smokers and non-smokers?

23.4392

- 4) Adjusting for the effects of alcohol consumption, is there a statistically meaningful relationship between life expectancy and smoking status? Provide the support for your conclusion.

Yes because the p-value is nearly 0.

- 5) How much of the variation in lifespan can be explained by smoking status and alcohol consumption.

71.29%

- 6) Which of the questions above are questions of statistical inference?

Problem 2 In an article published in Academic Press, Chester Bliss examined the impact of Vitamin C from two sources on the growth of teeth in Guinea pigs. In this experiment, the researchers measured the length of the teeth of the Guinea pigs as the experimental outcome and each Guinea pig was randomly assigned to one of two possible supplement treatments and one of three possible doses. (35 points)

The variables in the dataset are:

- length describes the length of growth in the teeth
- supp.VC indicates that the supplement type for Vitamin C is ascorbic acid (the baseline supplement type is OJ for orange juice)
- dose.2.0 indicates the amount of Vitamin C is 2.0 mg (the baseline is 1.0 mg.)

```

TGmod2.lm<-lm(len~supp*dose])
summary(TGmod2.lm)

## 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.VC          -5.930      1.608   -3.689  0.00074 ***
## dose.2.0          3.360      1.608    2.090  0.04374 *
## supp.VC:dose.2.0    6.010      2.274    2.643  0.01208 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 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

```

- Write the 4 prediction equations in their most simplified form with actual numbers from the R output (round estimates to the nearest 100th) for the model WITHOUT interactions.
- Write the 4 prediction equations in their most simplified form with actual numbers from the R output (round estimates to the nearest 100th) for the model WITH interactions.
- Based on the R output, which model do you think is a better choice, the one WITH interactions or the one WITHOUT interactions? Why? (Be specific and cite evidence to back up your choice.)
- Interpret the coefficient value in front of suppVC for the model you chose in part c. (Use words and describe it to someone who doesn't know statistics.)
- The R² statistics for both models are high, over 0.45. Is this alone sufficient evidence to assess that the linear models are adequate fits to the data? Briefly explain.

For problems 3 and 4, we are trying to understand how the aerial biomass (response variable) production of a certain type of marsh grass is related to the two predictor variables pH , which measures the pH level of the soil (from 0-14), and K , which measures the potassium level of the soil (in ppm). Aerial biomass is measured with units gm^{-2} .

Problem 3 Reference the R output and plots corresponding to Problem 3 at the end of this exam to answer this question. Consider the fit of a MLR regression model that includes the main effects of pH and K with the response and answer the following questions: (10 points)

- What is the average (absolute) difference in biomass (in gm^{-2}) for moss found in the “SI” location versus moss found in the “SM” location?
- What is the average (absolute) difference in biomass (in gm^{-2}) if the potassium (variable K) in the soil increases by 500 ppm?
- 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?

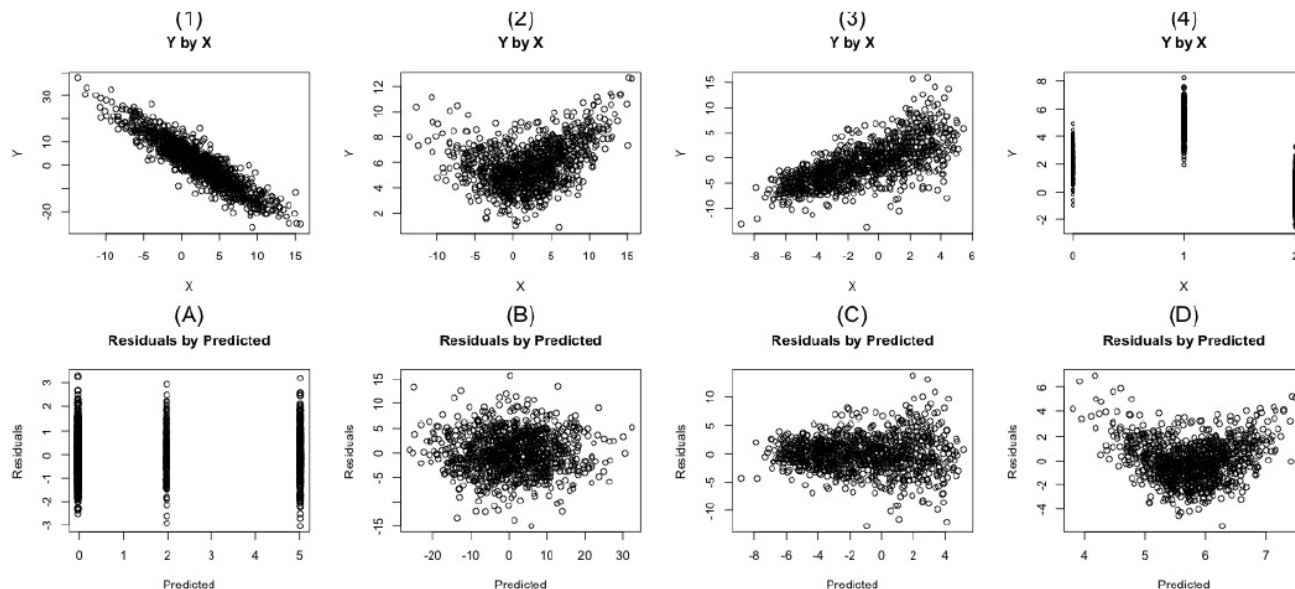
Problem 4 Reference the R output and plots corresponding to Problem 4 at the end of this exam to answer this question. Suppose instead of looking at the relationship between biomass and the predictor variables, we

decide to look at the relationship between the logarithm of the biomass and the predictor variables. Now fitting a linear regression model to this transformed response variable answer the following questions:

- Does this model look better or worse compared to the model in Problem 3? Justify your answer.
- 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?
- Are the answers to part(b) of Problem 4 and part (b) of Problem 3 directly comparable? Why or why not?

As a reminder, the logarithmic function is defined by the relationship that if $y = \log_a(x)$ then $a^y = x$ (and vice versa). (10 points)

Problem 5



Here are four plots of X and Y and four residual plots. I estimated a linear regression for each set of data in the Y by X plots. Match each Y by X plot with the residual plot that would result from the linear regression. (10 points)

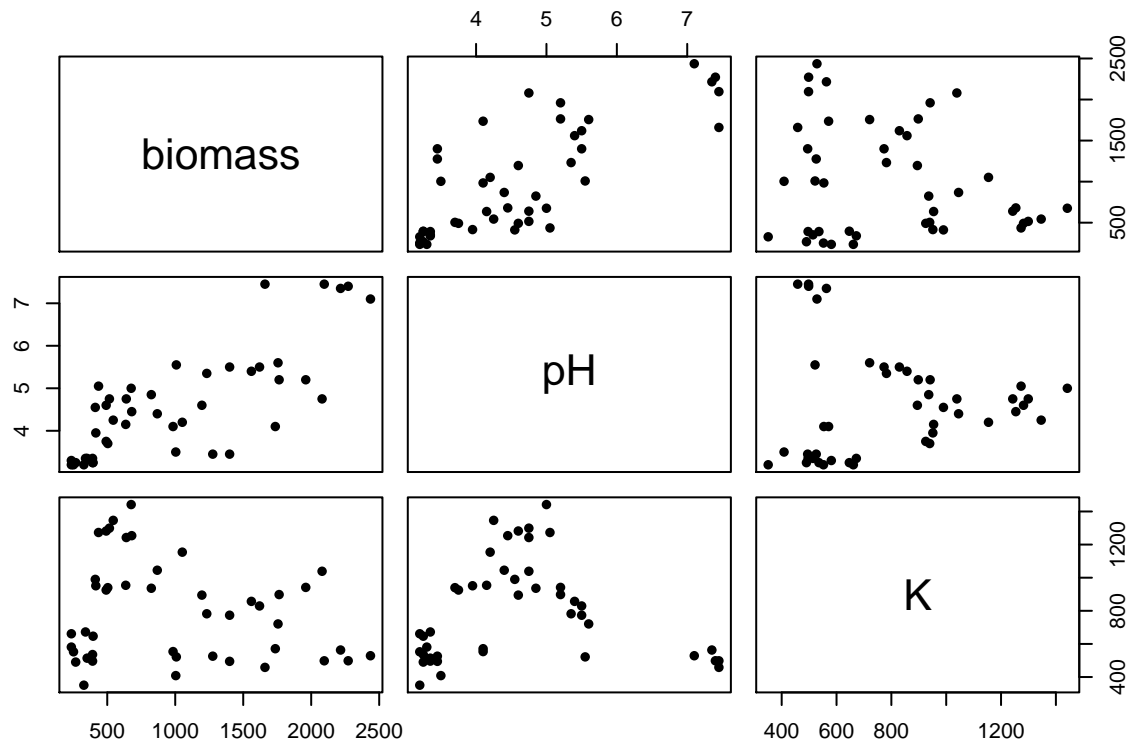
R output for Problem 3

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

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

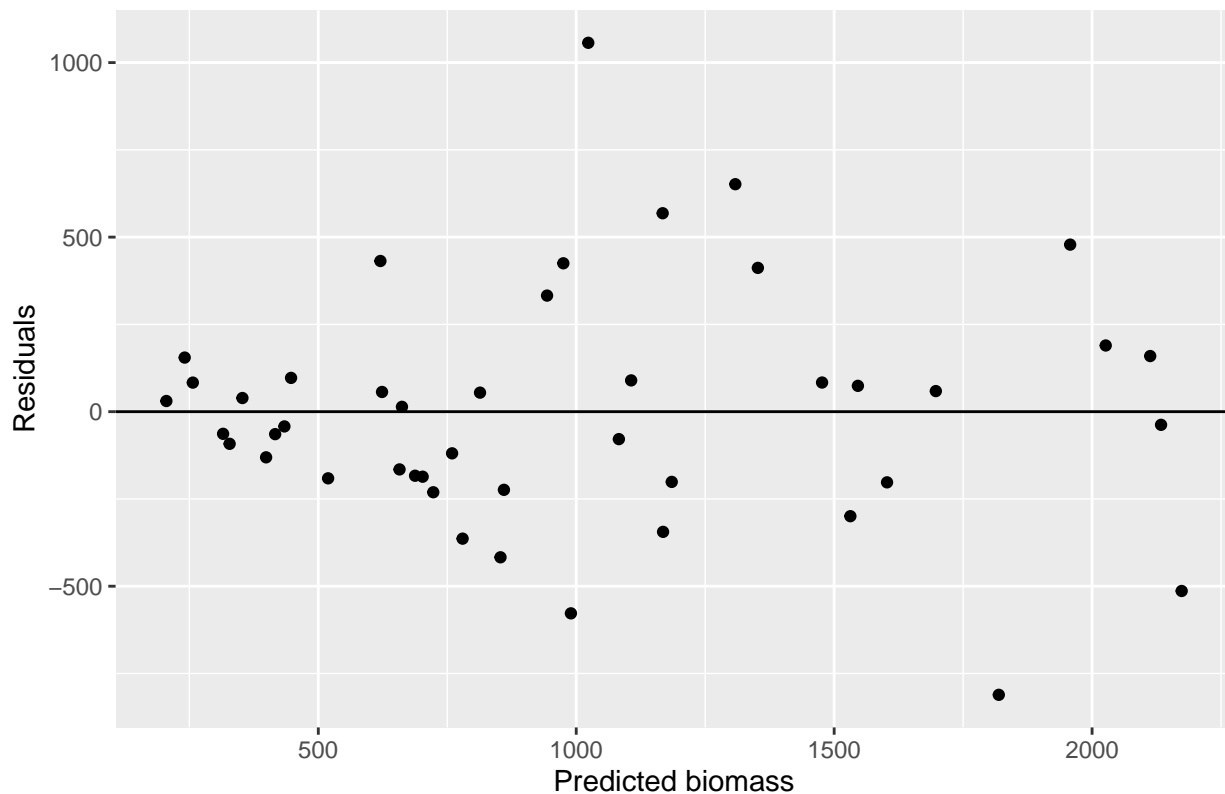
```
MLR_bio_sum <- summary(MLR_bio)
```

```
MLR_bio_sum
```

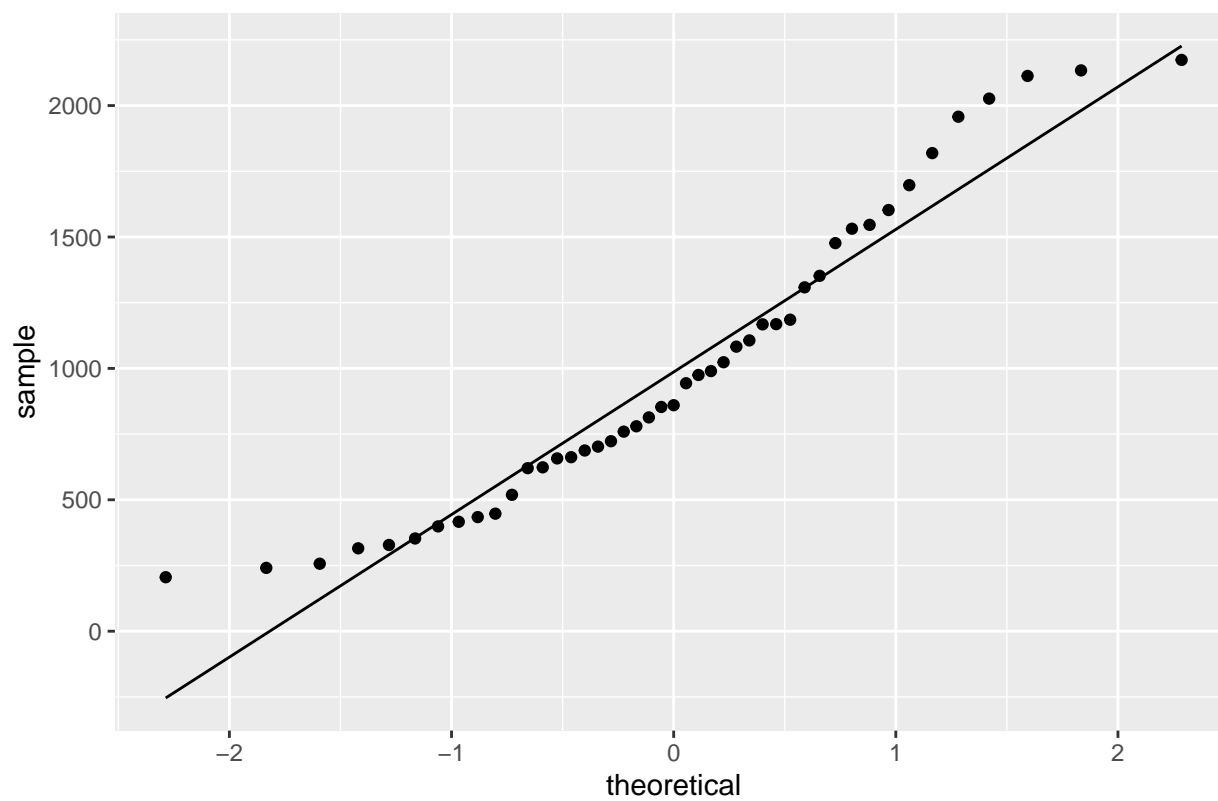
```
##
```

```
## 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
```

Residual plot for biomass data



Normal probability plot biomass data

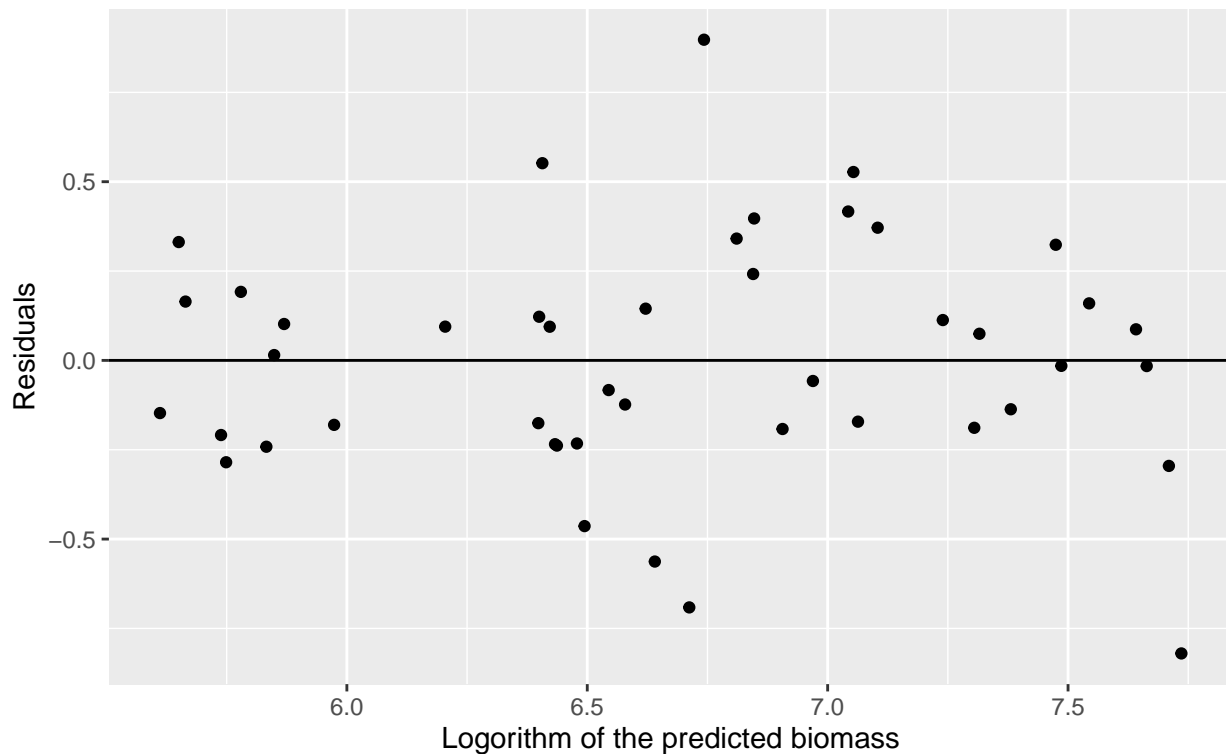


R output for Problem 4

```
##
## 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
```

Residual plot for biomass data

Logarithmic transformation of the response variable



Normal probability plot biomass data

Logarithmic transformation of the response variable

