Stat 21 Test 3

Due: May 16, 2020 by 12:00pm EST

This test is due on to be submitted on Gradescope by **May 16** at **12:00pm EST**. Please use the `#test3_questions` channel on Slack to post any clairfication questions. Do not ask questions like "Is [this] the right answer?" or "How can I get R to produce [this] plot?"

You must submit your solutions as a single **PDF** document uploaded to **Gradescope**. You may use R markdown to write up your solutions alone or you may use R markdown and hand-written solutions. **You must show all of your work**, including code input and output. Please make sure each problem is **clearly labeled** and that any handwritten components (such as pictures or equations) are easily readable in the PDF document. You may want to use a service like CamScanner (https://www.camscanner.com/) to help you upload handwritten pages and Small PDF (https://smallpdf.com/merge-pdf) to merge multiple PDFs into a single document.

You are permitted to reference all class material and use the internet. You are not permitted however, to get assistance from any other person, online or otherwise.

- Your file should contain the code to answer each question in its own code block. Your code should produce plots/output that will be automatically embedded in the output pdf file.
- Each answer must be supported by written statements and relevant plots.
- In order to knit this document, make sure you have installed the following packages in your version of RStudio: `ggplot2`, `tidyverse`, `gridExtra`, `knitr`.
- If you are running into issues with the Swarthmore RStudio server, you may submit your soulutions to me as an RMarkdown document in Slack via DM. Name your file using the convention: `[SwatID]_stat21_test3.rmd`. You may use this online R compiler to double check the the R code chunks of your document: https://rdrr.io/snippets/.
- If at any point you want to use the 'select' function to select specific columns from a data object, use 'dplyr::select' rather than 'select'.

You may write mathematical equations out with words like this: y_hat = beta0_hat + beta1_hat*x1

Or you can write a mathematical equation between dollar signs like this: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$. (Just make sure there is no white space space immediately after the first dollar sign or immediately before the second dollar sign!)

Problem 1 (20 points)

The R-output below contains the results from a regression model fit to a data set 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.

```
## 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? (2 points)

b) Suppose someone consumes about 2.5 alcoholic drinks per day and smokes regularly. Calculate (provide the formula for) the average life expectancy for this individual. (2 points)

c) What is the average difference in lifespan between smokers and non-smokers? (2 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 question(s) above are questions of statistical inference? (4 points)

Problem 2 (20 points)

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 data set are:

- *len* - the length of growth in the teeth
- *suppVC* - 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* - 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.

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

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

## Residuals:
##    Min     1Q Median     3Q    Max
## -6.600 -3.700  0.373  2.116  8.800

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.2725     1.2824   7.231 1.31e-09 ***
## suppVC       -3.7000     1.0936  -3.383   0.0013 **
## dose          9.7636     0.8768  11.135 6.31e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Residual standard error: 4.236 on 57 degrees of freedom
## Multiple R-squared:  0.7038, Adjusted R-squared:  0.6934
## F-statistic: 67.72 on 2 and 57 DF,  p-value: 8.716e-16
```

a) Show all your work and find the estimated tooth growth based on the main effects for Guina pigs who: (5 points)

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

```
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)   11.550      1.581   7.304 1.09e-09 ***
## suppVC        -8.255      2.236  -3.691 0.000507 ***
## dose           7.811      1.195   6.534 2.03e-08 ***
## suppVC:dose    3.904      1.691   2.309 0.024631 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Residual standard error: 4.083 on 56 degrees of freedom
## Multiple R-squared:  0.7296, Adjusted R-squared:  0.7151
## F-statistic: 50.36 on 3 and 56 DF,  p-value: 6.521e-16
```

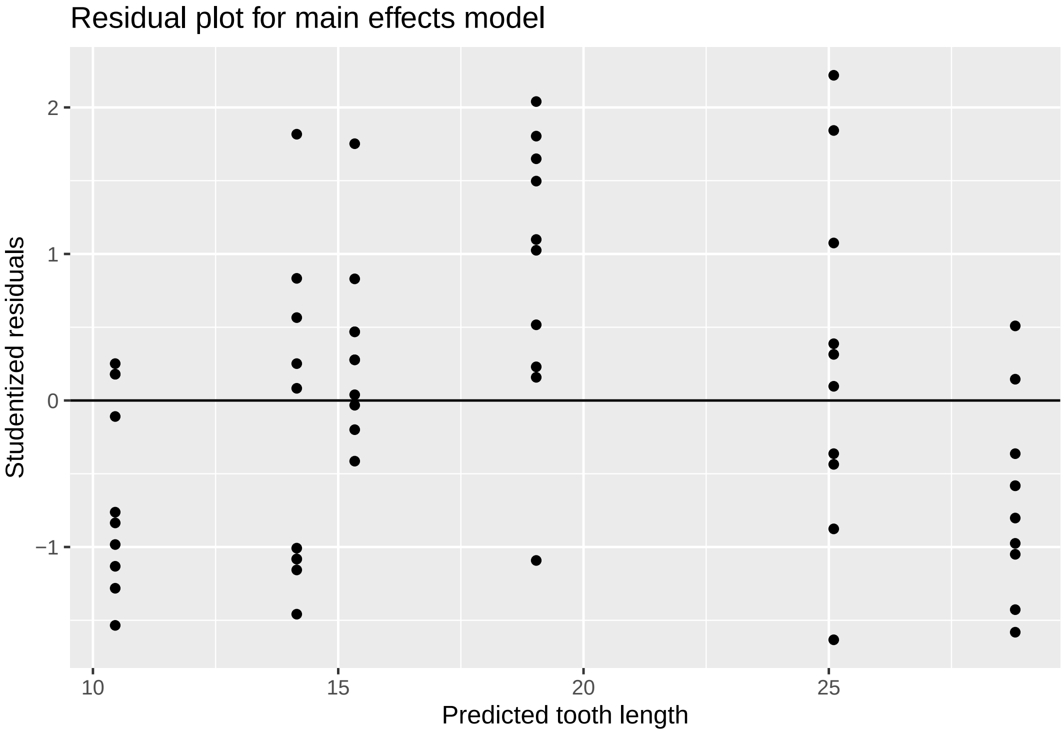
b) Write the down the interaction effects equations (in numbers) for predicting the average tooth growth for Guinea pics who: (5 points)

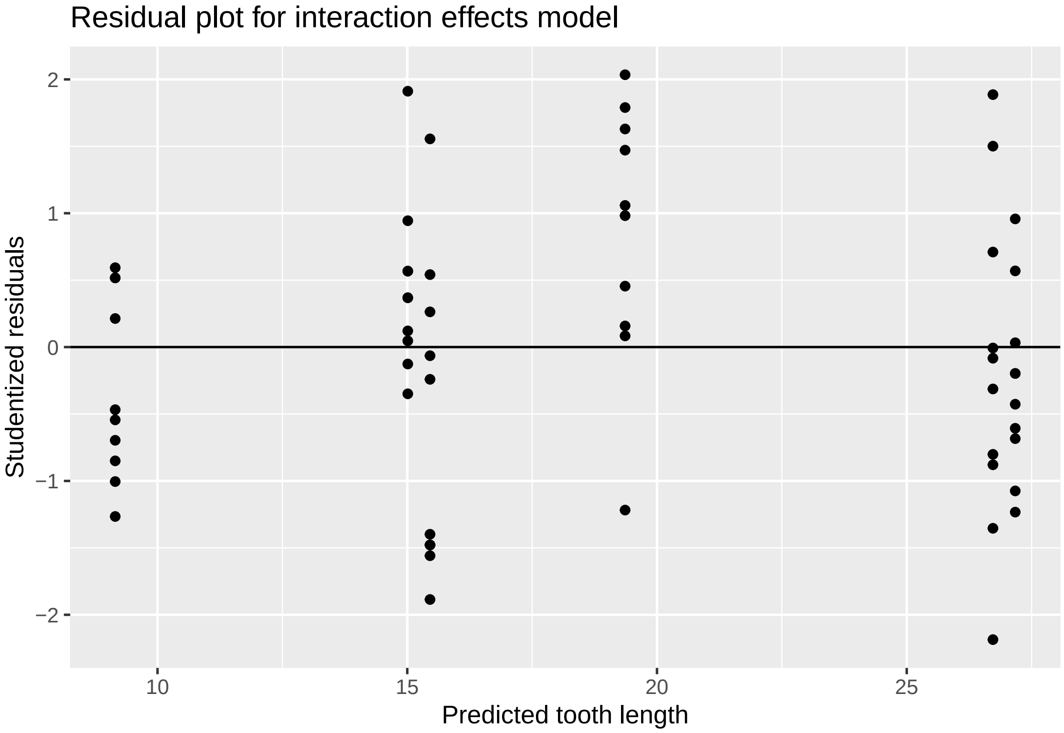
1) Have supplement type absorbic acid and dosage of 1.0 mg
2) Have supplement type absorbic acid and dosage of 2.0 mg
3) Have supplement type orange juice and dosage of 1.0 mg
4) Have supplement type orange juice and dosage of 2.0 mg

c) Based on the R output above and the residual plots on the next page, which model do you think is a better choice, the one **with** interactions or the one **without** interactions? Justify your answer. (5 points)

d) For whichever model you chose in part (c), explain the relationship between supplement type and tooth growth in language that can be understood by an average high school student. Be specific and provide an actual estimate for the effect of this predictor. (5 points)

Problem 2 Plots





Problem 3 (15 points)

Suppose 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}.

```
biomass <- read_table2(url(
  "http://www.swarthmore.edu/NatSci/sthornt1/DataFiles/biomass_data.txt"), col_names = TRUE)
```

Consider the fit of a MLR regression model that includes the main effects of location, pH, and K with the response. The residual plot and Normal probability plot for this model are show on the next page.

```
biomass1 <-  biomass %>% mutate(Loc_fct = factor(Location))
MLR_bio <- lm(BIO ~ Loc_fct + pH + K , data = biomass1)
summary(MLR_bio)
```

```
##
## Call:
## lm(formula = BIO ~ Loc_fct + pH + K, data = biomass1)
##
## 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
## Loc_fctSI   -497.6490   163.4252  -3.045   0.0041 **
## Loc_fctSM     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
```

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? (2 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? (2 points)

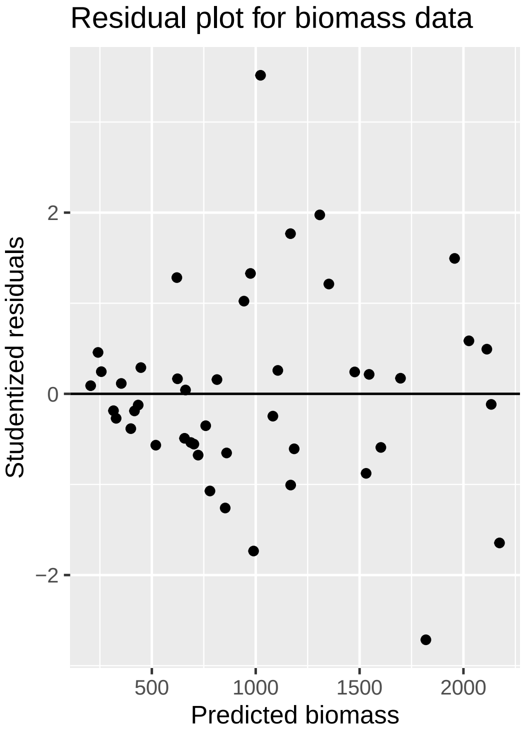
c) The code below considers the same predictor, fit by the same exact data only now the quantitative predictor variables have been standardized. (The residual plot and Normal probability plots for this model are also shown on the next page. Read the titles carefully!) Based on this version of the model, what is the average (absolute) difference in biomass (in gm^{-2}) if the potassium (variable K) in the soil increases by 500 ppm? (Hint: Recall that the `scale()` function in R standardizes the vector input to have a sample mean of zero and sample variance of one.) (7 points)

```
biomass2 <- biomass %>% mutate(Loc_fct = factor(Location),
                               pH_std = scale(pH),
                               K_std = scale(K))
MLR_bio_std <- lm(BIO ~ Loc_fct + pH_std + K_std, data = biomass2)
summary(MLR_bio_std)
```

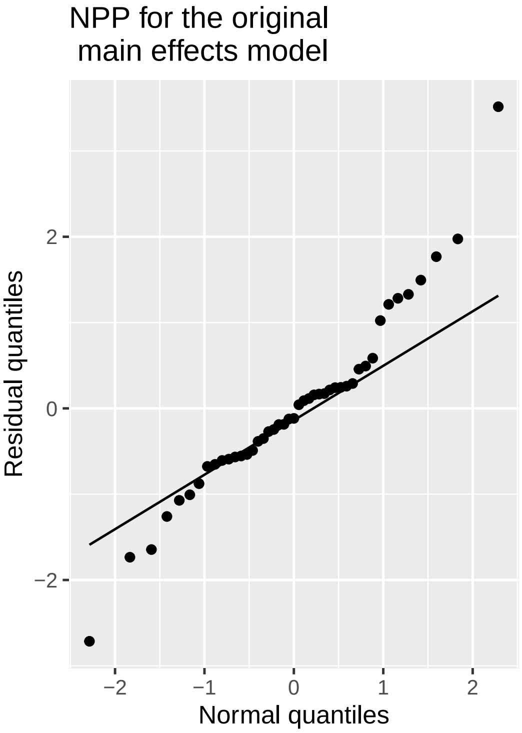
```
##
## Call:
## lm(formula = BIO ~ Loc_fct + pH_std + K_std, data = biomass2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -811.15 -190.99  -37.70   96.78 1056.67
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1147.29      99.73  11.504 2.92e-14 ***
## Loc_fctSI    -497.65     163.43  -3.045   0.0041 **
## Loc_fctSM      58.18     131.69   0.442   0.6610
## pH_std        517.38      54.04   9.574 6.68e-12 ***
## K_std        -300.43      69.17  -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
```

(d) Why might we want to consider the standardized model in part (c) rather than the original model? (6 points)

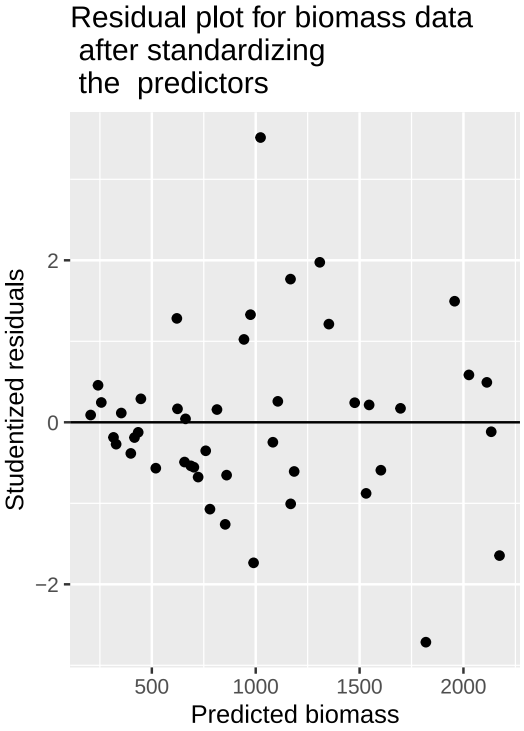
Problem 3 Plots



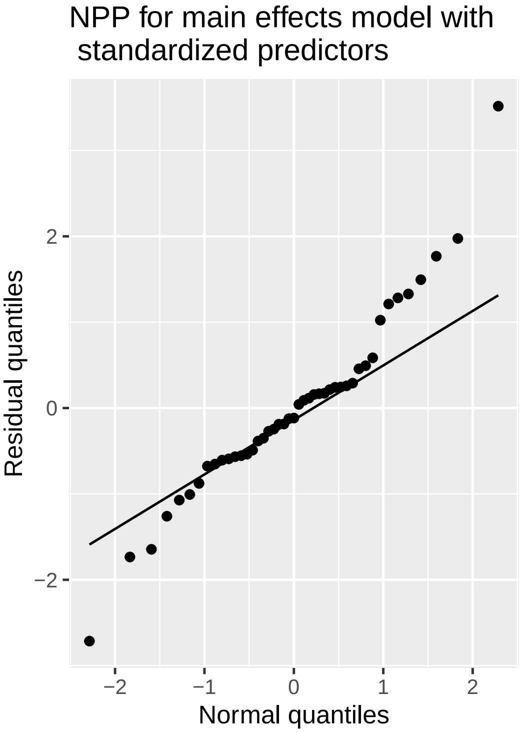
Residual plot for biomass data



NPP for the original
main effects model



Residual plot for biomass data
after standardizing
the predictors



NPP for main effects model with
standardized predictors

Problem 4 (15 points)

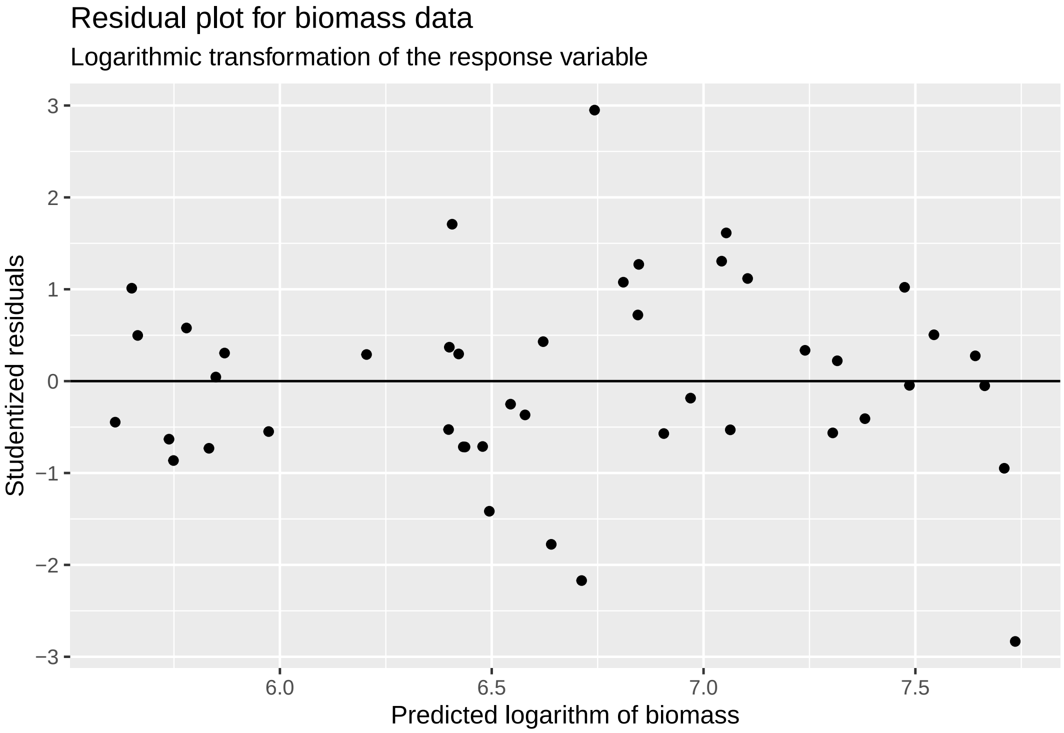
Suppose instead of looking at the relationship between biomass and the predictor variables, we decide to look at the relationship between **the logarithm of the biomass** and the predictor variables. The residual plot and Normal probability plots based on this model are show on the next page. As a reminder, the logarithmic function is defined by the relationship that if $y = log_a(x)$ then $a^y = x$ (and vice versa).

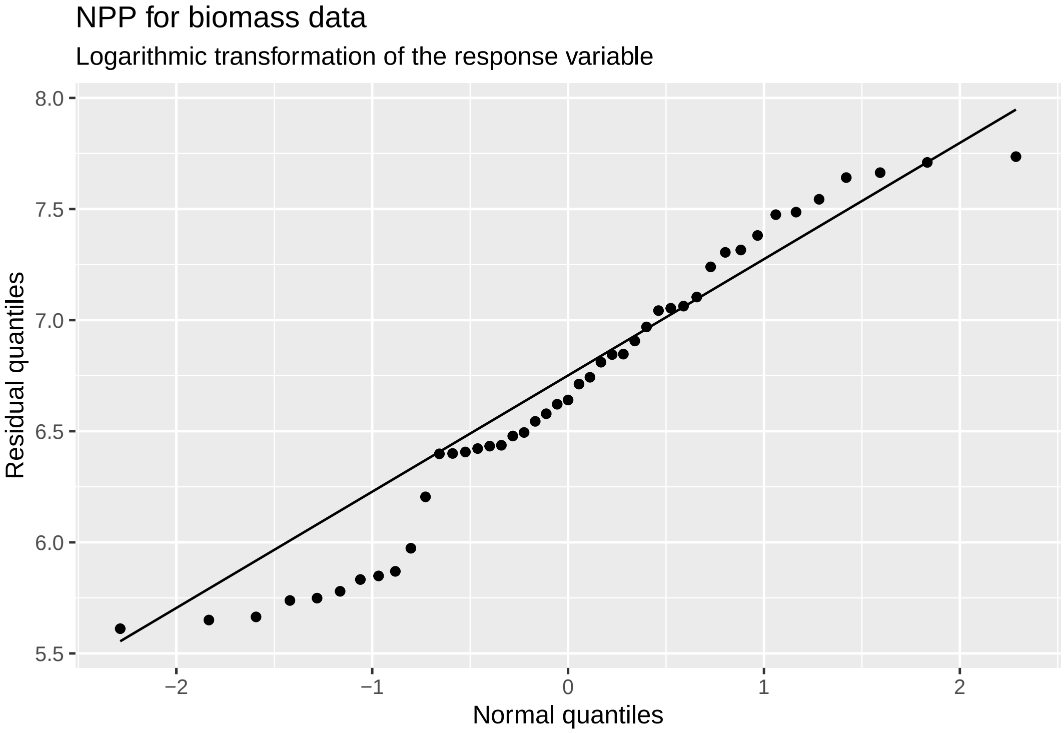
```
##
## Call:
## lm(formula = log_biomass ~ Loc_fct + pH + K, data = biomass3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82000 -0.19199 -0.01526  0.16459  0.89733
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.9125341  0.2917904  20.263  < 2e-16 ***
## Loc_fctSI   -0.9320328  0.1593950  -5.847 7.75e-07 ***
## Loc_fctSM   -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
```

a) Based on this model of the transformed response, what is the average difference in biomass when the potassium in the soil increases by 500 ppm? Show your work. (5 points)

b) Are the answers to part (a) of Problem 4 and part (b) of Problem 3 directly comparable? Why or why not? (10 points)

Problem 4 Plots





Problem 5 (30 points)

```
coasters <- read_table2(url(
  "http://www.swarthmore.edu/NatSci/sthornt1/DataFiles/roller_coasters.txt"),
  col_types = "f?????f") %>%
  filter(!is.na(Duration) & !is.na(Inversions))
head(coasters)
```

```
## # A tibble: 6 x 7
##   Track Speed Height  Drop Length Duration Inversions
##   <fct> <dbl>  <dbl> <dbl>  <dbl>    <dbl> <fct>
## 1 Steel  35     51.8  49.8  1312.       95 0
## 2 Steel  41     94    66    2423        60 1
## 3 Steel  42     70    64    3100       120 0
## 4 Steel  45     56    47     635        66 1
## 5 Steel  49.7  108.   95.2  2625.      130 1
## 6 Steel  49.7  108.   95.2  2625.      130 1
```

Suppose we have randomly surveyed 91 roller coasters across the US. We are interested in investigating the relationship among a set of predictor variables with the quantitative response variable, the maximum speed of the coaster (mph). The quantitative predictor variables we are going to consider include

- the length of the track (in feet),
- the duration of the ride (in seconds),
- the highest climb the ride reaches (in feet), and
- the lowest drop the ride reaches (in feet) - coasters can extend below ground.

The categorical predictor variables we are going to consider are

- the type of track (wooden or steel) and
- whether or not the ride has a loop/inversion (1 for yes, 0 for no).

Note: For this data set, we are given that the data is a random sample so we can assume that it is representative of all roller coasters in the US. You can also assume that the data is independent for this example.

(a) Create a scatter plot matrix showing only the quantitative variables (predictors and response) and create box plots for each of the categorical variables (with coaster speed on the vertical axis). Does there appear to be any evidence of multicollinearity among the quantitative predictors? (5 points)

(b) Fit the following four regression models to this data and write out the corresponding estimated regression equations as well as the adjusted coefficient of determination for each model. (8 points)

Model 1: $E[Y|\text{length, height}] = \beta_0 + \beta_1\text{length} + \beta_2\text{height} + \epsilon$

Model 2: $E[Y|\text{length, height, track}] = \beta_0 + \beta_1\text{length} + \beta_2\text{height} + \beta_3 w_{track} + \epsilon$, where $w_{track} = \begin{cases} 1, & \text{if steel} \\ 0, & \text{otherwise} \end{cases}$

Model 3: $E[Y|\text{duration, drop}] = \beta_0 + \beta_1\text{duration} + \beta_2\text{drop} + \epsilon$

Model 4: $E[Y|\text{duration, drop, loop}] = \beta_0 + \beta_1\text{duration} + \beta_2\text{drop} + \beta_3 w_{loop} + \epsilon$, where $w_{loop} = \begin{cases} 1, & \text{if has inversion} \\ 0, & \text{otherwise} \end{cases}$

(c) Perform a thorough analysis of the studentized residuals for each of the four models you fit in part (b). Make sure every plot is clearly labeled and include at least a one sentence statement explaining the relevance of each plot. (Hint: Make sure you load "MASS" library to find the studentized residuals.) (7 points)

(d) Suppose your friend's favorite ride is the Wicked Twister and they want to know what is the estimated maximum speed this coaster reaches. This a steel roller coaster that reaches a height of 215 ft, has a drop of 206 ft, is 675 ft long, lasts for 40 seconds, and has no inversions. Based on your answers to parts (a)- (c), which of the regression models are appropriate to use to answer your friend's question about the Wicked Twister? (5 points)

(e) Wikipedia lists the speed of the Wicked Twister as 72 mph. Suppose we want to determine if there is a statistically significant difference between the speed reported by Wikipedia and the estimated maximum speed based on one of our models. Explain how we could do this and state which models from part (b) (if any) are appropriate to use to answer this question. Explain your answer. (5 points)