Beckham Carver

Homework 4   
STAT 4025/5025 – Due Sunday, February 13th 11:00 pm   
27 points   
   
1. Utilize the trout\_hatchery\_22.csv data file. Recall that there is interest in comparing Diets in terms of length gain after six weeks for rainbow trout fry.   
   
a. One way to think about comparing groups is to use a regression model with indicator variables. Write out a regression model with an indicator variable that will allow for you to determine if there is a difference in Diets A and B – make Diet B your reference group. Explain the symbols from your regression model and explicitly   
define your indicator variable. (5 pts – 2 points for model and 2 points for explain   
symbols, 1 point for explicitly defining the indicator variable)

Our indicator variable is the Diet, with Diet B as our reference category. This means Diet B will be assigned ‘0’ for our model and Diet A will be assigned ‘1’ which is referenced as ‘indicator’. In mathematical terms our model is:

Y = β ₁ + β ₂ \* indicator + error

Where ‘β ₁’ is the mean length of diet B, and ‘β ₂’ is the difference in means between Diets A and B. And Y is the mean length, our dependent variable.

In R this would look like:

p1\_lm <- lm(mn\_length~Diet\_Ind, data=trout\_p1)

Where Diet\_Ind is a column containing the indicator variables for each experimental unit assigned as described earlier.

b. Explicitly provide models for the mean increase in length for Diets and for the mean increase in length for Diet B using the regression model symbols from part a. Note that I don’t want numbers here but symbols. (2 pts)

mn\_length = β ₁ + β ₂ \* Diet\_Ind + error

Where beta 1 is diet B and beta 2 is diet A, the total of these two is the difference in means.

diet B is the reference category!!  
c. Use the lm() function in R in an appropriate manner to test whether the mean length of trout raised with Diet A is significantly different from the mean length of trout raised with Diet B. State your hypotheses in terms of the beta(s) in your regression model from part a; State your conclusion based on a significance level of 0.05. Keep Diet B as the reference category. (5 pts – 2 pts for correctly using lm() – be sure to provide your code for credit; 1 pt for correct hypotheses; 2 pts for correct conclusion)

Hypothesis:

* H₀ : β ₁ = 0
* H₁ : β ₂ ≠ 0

R Output:

Residuals:

Min 1Q Median 3Q Max

-4.740 -2.450 -1.890 1.915 10.160

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 32.600 2.287 14.257 5.71e-07 \*\*\*

Diet\_Ind 6.940 3.234 2.146 0.0642 .

Conclusions:

By our regression model Diet B produced a change in mean length of 32.6 whereas Diet A produced a change of 25.66. However, our significance level indicated by our Diet\_Ind was above 0.05, which is not within our cutoff limit meaning we cannot conclude the difference in diets caused the difference in means.

Code:

trout <- read.csv("trout\_hatchery\_22.csv",header=TRUE)

trout\_p1 <- trout %>%

group\_by(Diet,raceway)%>%

summarise(count=n(),mn\_length=mean(Length\_6w\_mm,na.rm=T))%>%

filter(Diet=="A" || Diet=="B")%>%

mutate(count = NULL)

trout\_p1$Diet\_Ind <- ifelse(trout\_p1$Diet == 'A', 1,0)

p1\_lm <- lm(mn\_length~Diet\_Ind,data=trout\_p1)

summary(p1\_lm)

2. Refer to Exercise 3.3 on p. 60 of the Oehlert text and do the following

Look at page 60 in the olwork text  
   
a. Using R, construct Normal QQ plots of the moisture content of the silage for each   
of the treatment (note, you should have 4 separate plots – place the plots in a 2   
row by 2 column format). 2 pts

Chart, line chart

Description automatically generated  
b. Are you suspect of the Normality assumption from the plot in ‘a’? Explain 1 pt

The plots were constructed from the minimum of 3 data points each, with such a small sample size having confidence in normality is difficult.

c. Produce side-by-side boxplots of the moisture content for each treatment group. 2   
pts

Diagram

Description automatically generated

d. What do you notice from the boxplots in terms of whether or not you think there   
is any difference between the treatment groups in terms of mean moisture   
content? 2 pts

The box plots have greatly differing minimums and maximums, notably the CON and NACL plots are skewed to the lower   
e. Give a practical explanation of the experimental error in terms of the silage in the   
sodium chloride group (if you don’t know what orange pulp silage is, consider the   
following link to learn more - https://www.feedipedia.org/node/12416).– 2 pts.   
What are some things that could contribute to the experimental error variance in   
this group? 2 pts

I do not completely understand the question here, however a greater sample size would reduce the experimental error given additional samples did not skew further from the observed average. This is because the equation becomes less sensitive as the sample size increases.

f. By hand, compute the estimate of the experimental error variance using the data   
from the sodium chloride group...show me your work – 3 pts.

A picture containing text, whiteboard, document

Description automatically generated  
g. For this example, show how you might use R to randomly assign piles of silage to   
the different treatment groups – 3 pts.

Graphical user interface, text

Description automatically generated with medium confidence

Table, calendar

Description automatically generated with medium confidence