140.616.01.HW.10.Delahanty.Jeremy

Jeremy Delahanty

2024-05-01

1

The mouse PH dataset available as object `mouse.ph` in the SPH.140.615 package contains the blood pH readings on male mouse litter mates of two strains that had been selected for high and low blood pH. Only litters with at least 4 males were considered, and males were selected at random whenver more than 4 males were present in the litter. Data are presented on seven litters for each strain. Plot the data. Write down and explain your model, analyze the data, and explain your results.

```
# Call the library from class
library(SPH.140.615)
```

With the available data frame from the class, we should use the following procedure for testing our assumptions to see if they are satisfied for our ANOVA. The key assumptions are that the dataset is modeled approximately by a Gaussian distribution and that sampled groups have approximately homogeneous variances. We can first examine whether these assumptions are true through first examining the object and then graphing the dataset at different levels. I will use an significance value $\alpha = 0.05$ for this test. Our Null Hypothesis H_0 is that there is no difference in blood pH between strains.

```
# Examine the dataset.
str(mouse.ph)
```

```
## 'data.frame':    56 obs. of  3 variables:
##  $ pH    : num  7.43 7.38 7.49 7.49 7.39 7.46 7.5 7.55 7.53 7.5 ...
##  $ strain: Factor w/ 2 levels "pHH","pHL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ litter: Factor w/ 14 levels "387","388","389",..: 1 1 1 1 2 2 2 2 3 3 ...
```

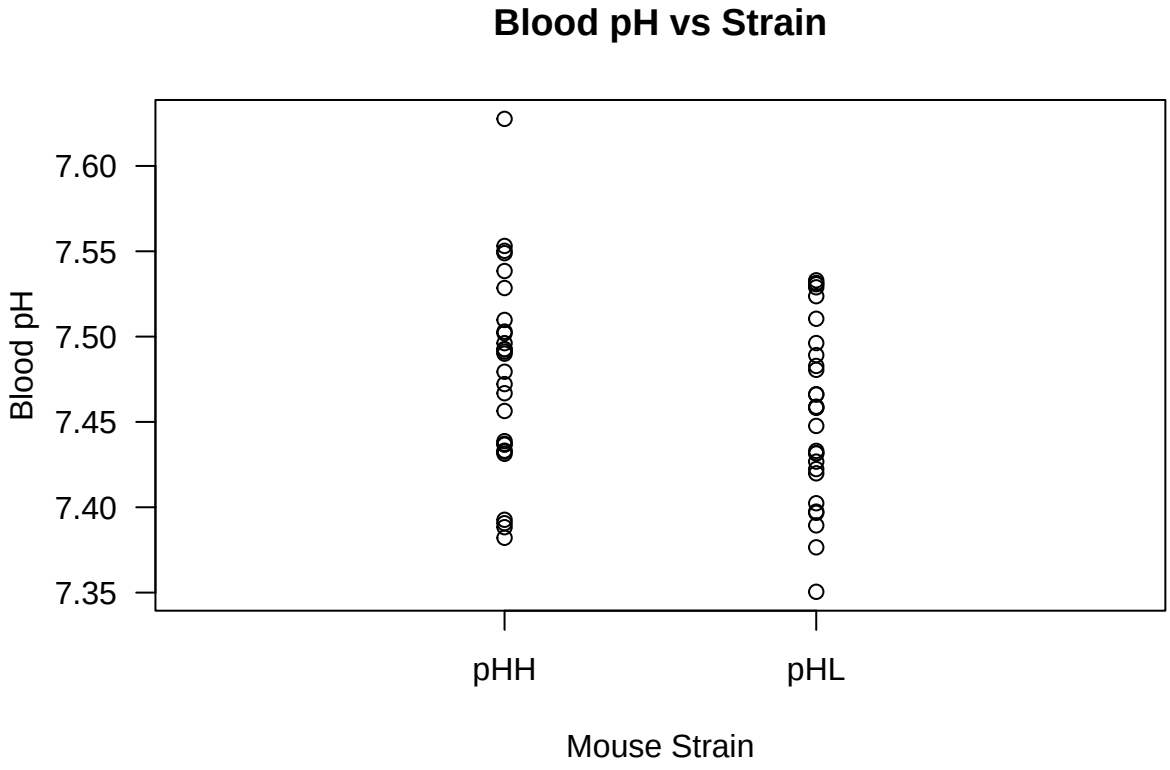
```
head(mouse.ph)
```

```
##     pH strain litter
## 1 7.43    pHH    387
## 2 7.38    pHH    387
## 3 7.49    pHH    387
## 4 7.49    pHH    387
## 5 7.39    pHH    388
## 6 7.46    pHH    388
```

```r
summary(mouse.ph)
```

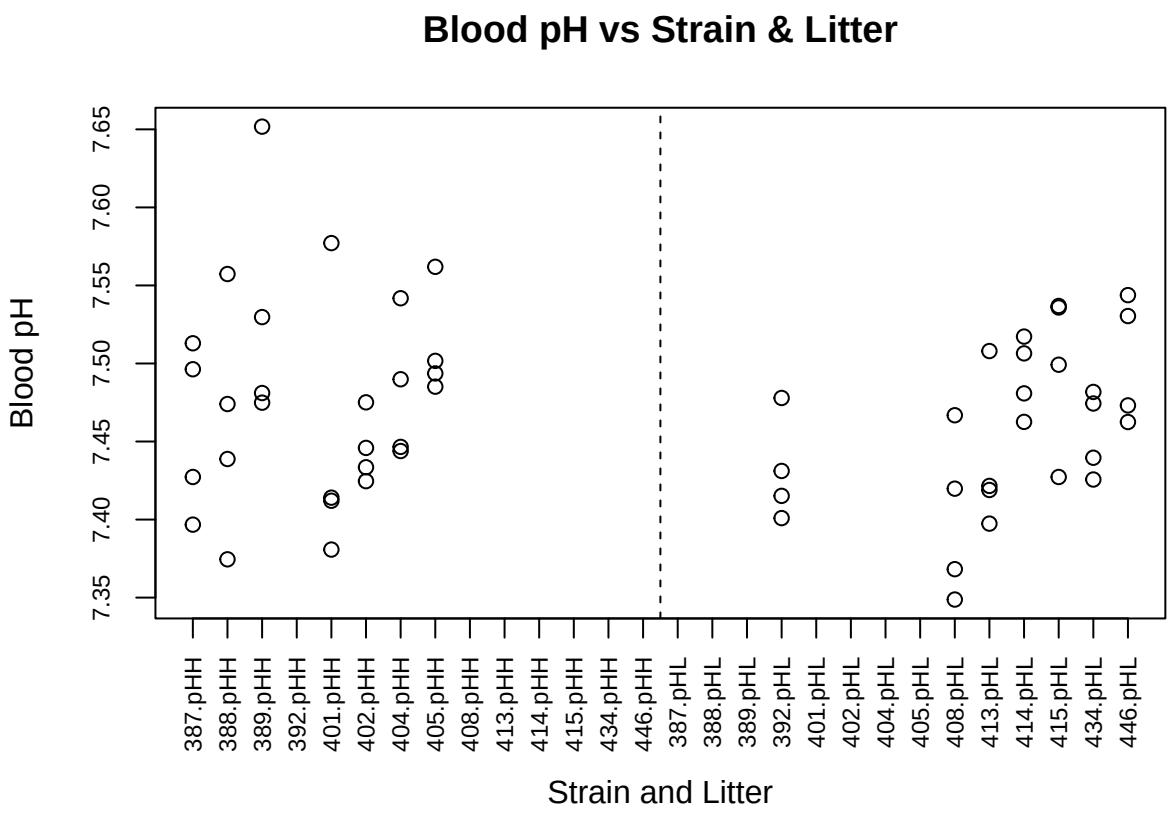
```
##        pH        strain       litter
##  Min.   :7.350   pHH:28   387    : 4
##  1st Qu.:7.430   pHL:28   388    : 4
##  Median :7.470            389    : 4
##  Mean   :7.466            392    : 4
##  3rd Qu.:7.500            401    : 4
##  Max.   :7.630            402    : 4
##                           (Other):32
```

```r
# We see we have factors of strain, litter, and measurements for pH for each
# mouse.
# Plot the data with a strip chart as shown in lab, first for just strain
# differences
par(las=1) # style to axis labels horizontally
stripchart(jitter(mouse.ph$pH, factor=2) ~ mouse.ph$strain, ylab="Blood pH",
           xlab="Mouse Strain", main="Blood pH vs Strain",pch=1,
           vertical=TRUE, xlim=c(0, 3))
```




```r
# The data appears to have somewhat similar mean and variance at the level of strains
# We should also check for similarities between each litter for each strain.
par(las=3) # style to label axes vertically
stripchart(split(
```

```
  jitter(mouse.ph$pH, factor=14), list(mouse.ph$litter, mouse.ph$strain)),
  cex.axis=0.75, ylab="Blood pH", main="Blood pH vs Strain & Litter", pch=1, vertical=TRUE,)
# Without specifying this separately, the vertical ticks will overlap upon xlab
title(xlab = "Strain and Litter", mgp = c(4, 0, 0))
abline(v=14.5, lty=2)
```




```
# We can also use Bartlett's test for determine whether the variances between
# groups are equal. We can do this across the factor of strain and litter on pH.
bartlett.test(mouse.ph$pH ~ mouse.ph$strain)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  mouse.ph$pH by mouse.ph$strain
## Bartlett's K-squared = 0.54021, df = 1, p-value = 0.4623
```

```
bartlett.test(mouse.ph$pH ~ mouse.ph$litter)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  mouse.ph$pH by mouse.ph$litter
## Bartlett's K-squared = 10.266, df = 13, p-value = 0.672
```

```r
# The data appear to have similar variances in this case as well overall.
# We want to analyze this data using a *nested* AOV as described in lab.
# This takes the output from an ANOVA and performs corrections of ratio of
# mean squares for accurate p-value calculations in nested datasets.
# The nesting is: Strain[Litter]
nested.anova(aov(mouse.ph$pH ~ mouse.ph$strain / mouse.ph$litter))
```

```
## Analysis of Variance Table
##
## Response: mouse.ph$pH
##                                 Df   Sum Sq   Mean Sq F value  Pr(>F)
## mouse.ph$strain                  1 0.006645 0.0066446  1.2638 0.28292
## mouse.ph$strain:mouse.ph$litter 12 0.063093 0.0052577  2.2221 0.02822 *
## Residuals                       42 0.099375 0.0023661
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that when using the nested ANOVA (a form of Mixed Effects model provided in the course) where the fixed effect α_i corresponds to mouse strain and random effect $\beta_i j$ corresponds to litter effect that the strain does not have a significant effect. We fail to reject our null hypothesis. The influence of strain upon litter did have a statistically significant effect. We used Bartlett's test for homogeneity of variances to see that there was no difference in variance between strains as well as between litters. We were

2

We are investigating the effect of a treatment in a group of mice, by comparing it to a control group.For both the treatment and the control group, we have five mice (so a total of ten mice are used), and we take two replicate measurements per mouse. The group (treatment/control) sum of squares is 4.37, the mouse within group sum of squares is 43.43, and the error sum of squares is 7.90. Carry out an the appropriate analysis of variance, give an explanation whether you consider a factor to be a fixed or a random effect, and comment on the significance of the results.

This question requires us to assume that the data are normally distributed and are independent and identically distributed. Without the original dataset and just these summary statistics, we will have to perform this calculation by hand in R.

```r
# Define sample size for each group
n_per_group <- 5
# Define total size of the sample
n_tot <- 10
# Define number of replicates per mouse
n_replicates <- 2
# Define the number of groups
n_groups <- 2
# Sum of Squares Between Groups
between_ssq <- 4.37
# Sum of Squares Within Groups
within_ssq <- 43.43
```

```r
# Error of Sum of Squares
error_ssq <- 7.90

# Now, calculate group means and compute the F-statistic as from class
group_mean <- between_ssq / (n_groups - 1)
within_mean <- within_ssq / (n_tot - n_groups)
critical_value <- group_mean / within_mean
p_value <- pf(critical_value, df1 = n_groups -1, df2 = n_tot - n_groups,
              lower.tail =FALSE)

# Display the calculated P-value
cat("The p-value associated with this AOV is:", p_value)
```

```
## The p-value associated with this AOV is: 0.3958075
```

In this model of a balanced experimental design, our fixed effect is the treatment applied to the animal while the random effect is the mouse that is chosen. With a significance value of $\alpha = 0.05$, we would fail to reject the null hypothesis and conclude that there is insufficient evidence that there is a treatment effect in the mice.

3

A research study was conducted to examine the impact of eating a high protein breakfast on adolescents performance during a physical education physical fitness test. Half of the subjects received a high protein breakfast and half were given a low protein breakfast. All of the adolescents, both male and female, were given a fitness test with high scores representing better performance.

For this question, I will use a significance value of $\alpha = 0.05$.

```r
# First, we need to create the dataset
protein <- data.frame(
  Scores = c(10,7,9,6,8,5,4,7,4,5,5,4,6,3,2,3,4,5,1,2),
  Sex = factor(rep(c("M", "F"), rep(10, 2))),
  Food = factor(rep(c("HP", "LP"), rep(5,2)))
)

# Next, display the structure of protein
str(protein)
```

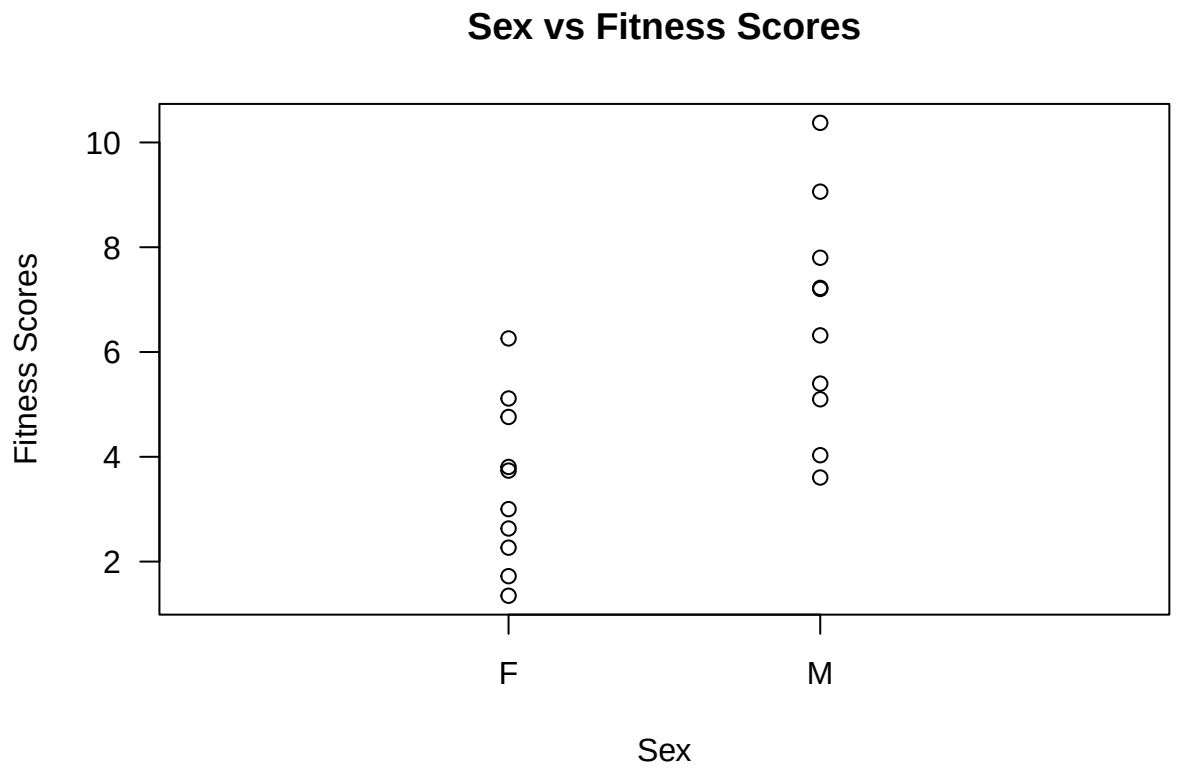
```
## 'data.frame':    20 obs. of  3 variables:
##  $ Scores: num  10 7 9 6 8 5 4 7 4 5 ...
##  $ Sex   : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Food  : Factor w/ 2 levels "HP","LP": 1 1 1 1 1 2 2 2 2 2 ...
```

```r
# Create a summary of the dataset
summary(protein)
```

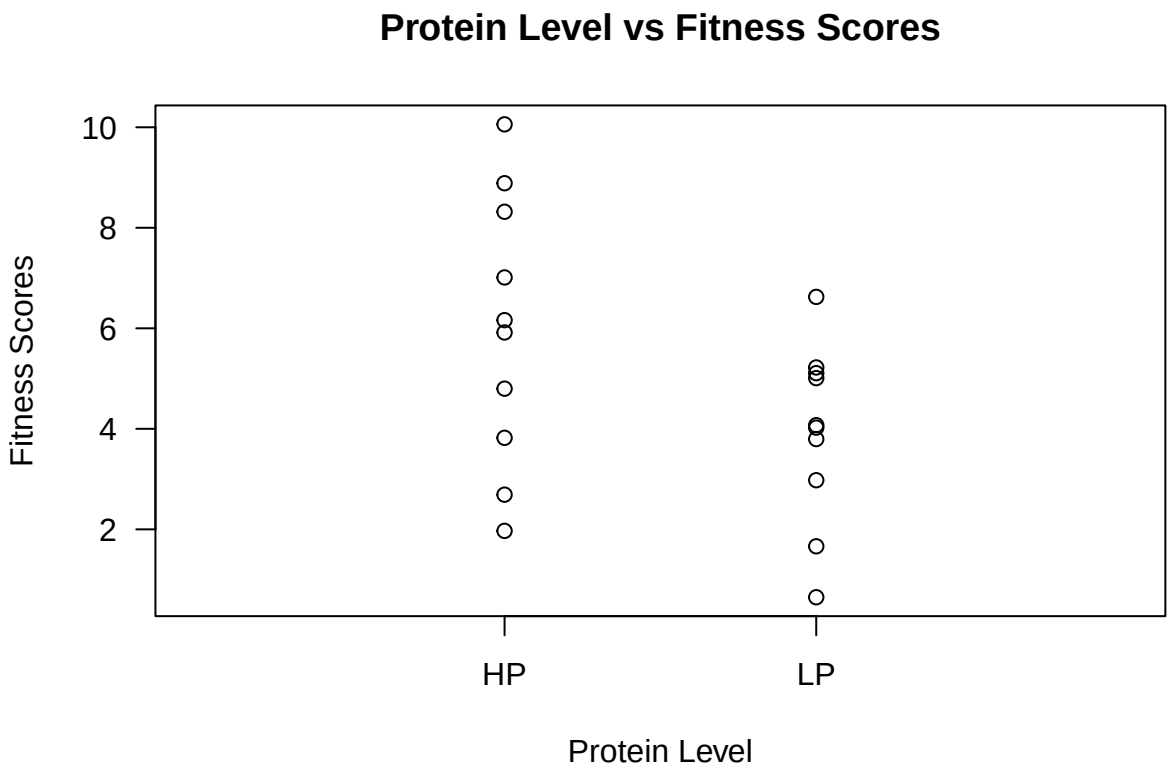
```
##      Scores      Sex    Food
```

```
##  Min.   : 1.00   F:10   HP:10
##  1st Qu.: 3.75   M:10   LP:10
##  Median : 5.00
##  Mean   : 5.00
##  3rd Qu.: 6.25
##  Max.   :10.00
```

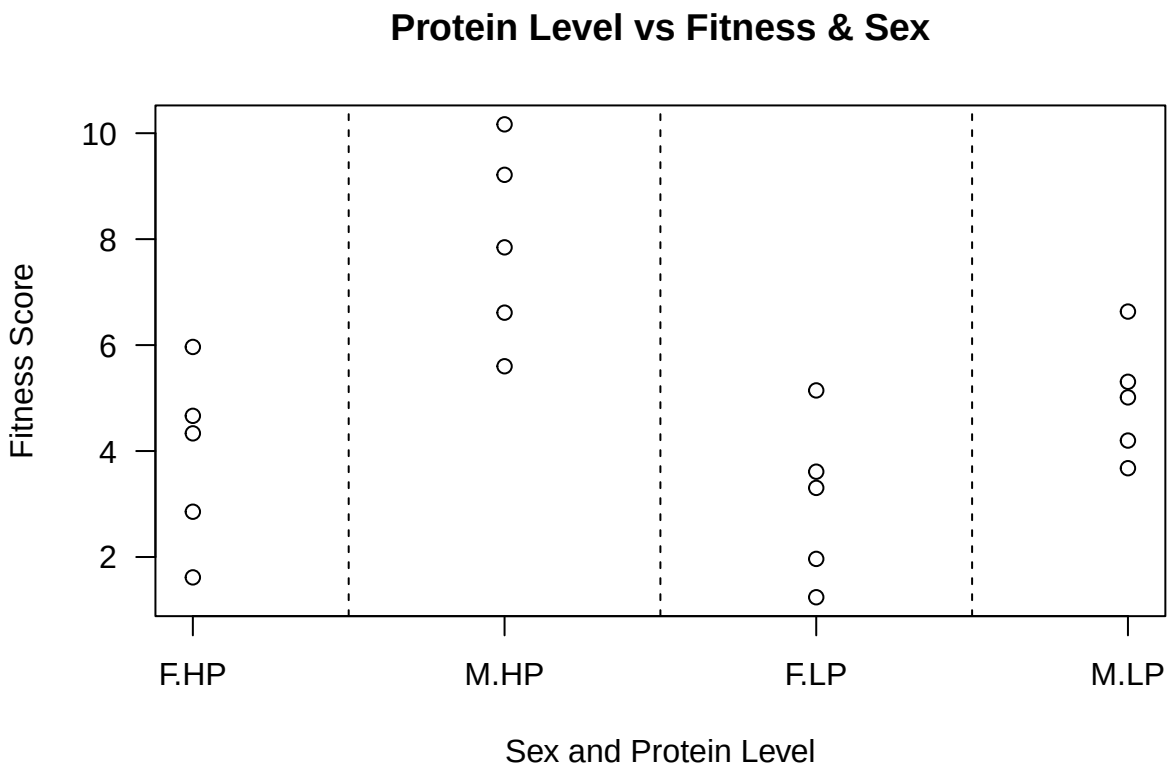
```r
# Now, plot the data across each factor
# First, across sex
par(las=1)
stripchart(jitter(Scores, factor =2 ) ~ Sex, data = protein,
           xlab="Sex", ylab="Fitness Scores", main="Sex vs Fitness Scores",
           pch = 1, vertical = TRUE, xlim=c(0, 3))
```




```r
# Now across protein level
par(las=1)
stripchart(jitter(Scores, factor = 2) ~ Food, data = protein,
           xlab="Protein Level", ylab="Fitness Scores",
           main="Protein Level vs Fitness Scores", pch = 1, vertical = TRUE,
           xlim=c(0,3))
```




```
# Now protein level nested with sex
stripchart(split(jitter(protein$Scores, factor = 2), list(protein$Sex, protein$Food)),
           ylab="Fitness Score", main="Protein Level vs Fitness & Sex",
           xlab="Sex and Protein Level",pch = 1, vertical = TRUE,)
abline(v=c(1.5:3.5), lty=2)
```




```
# The observations of our strip charts demonstrate that the variances appear
# approximately equal. We can continue forward with our AOV. We will use
# both factors (sex and protein content) in a 2 way ANOVA as demonstrated
# in the lab lecture. This will include an interaction term between sex
# and protein content.
summary(aov(Scores ~ Sex * Food, data = protein))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Sex          1     45   45.00  20.000 0.000385 ***
## Food         1     20   20.00   8.889 0.008814 **
## Sex:Food     1      5    5.00   2.222 0.155487
## Residuals   16     36    2.25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see in our 2-way ANOVA investigating the fixed effects of sex and protein content in breakfast demonstrates significant evidence at $\alpha = 0.05$ that sex and breakfast protein content do change fitness test scores in this population. There is insufficient evidence to conclude that the interaction between sex and protein content influences fitness test scores.

4

In an experiment, the effect of oxygen level on fermentation end products was examined.Four oxygen concentrations and two sugar types were used. The amount of ethanol was measured for each oxygen-sugar combination. The data are available here. Analyze the data, and clearly state your assumptions. Plot the data in a meaningful way, and check your model assumptions (use a data transformation if necessary). Summarize your findings in a sentence or two, and include the ANOVA table. Hint: if you can't take a logarithmic transformation because you have some zeros in your data, you can instead take the square root, or add one 'pseudo-count' before taking the log (as is done for RNA-seq analyses)

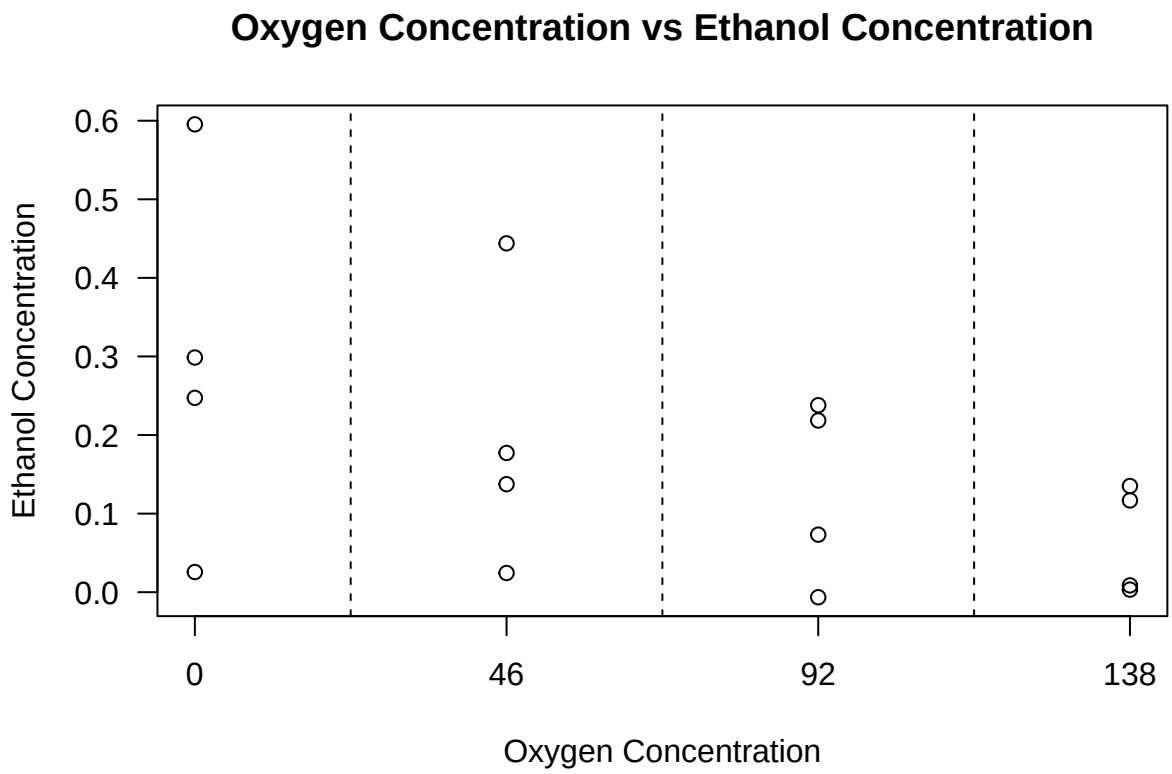
```
# First, we need to download the dataset
download.file(
  "http://biostat.jhsph.edu/~iruczins/teaching/140.615/data/fermentation.csv",
  destfile = "./fermentation.csv", cacheOK=TRUE)
fermentation <- read.csv("./fermentation.csv")

# Inspect the data
head(fermentation)
```

```
##   ethanol oxygen     sugar
## 1    0.59      0 Galactose
## 2    0.30      0 Galactose
## 3    0.25      0   Glucose
## 4    0.03      0   Glucose
## 5    0.44     46 Galactose
## 6    0.18     46 Galactose
```

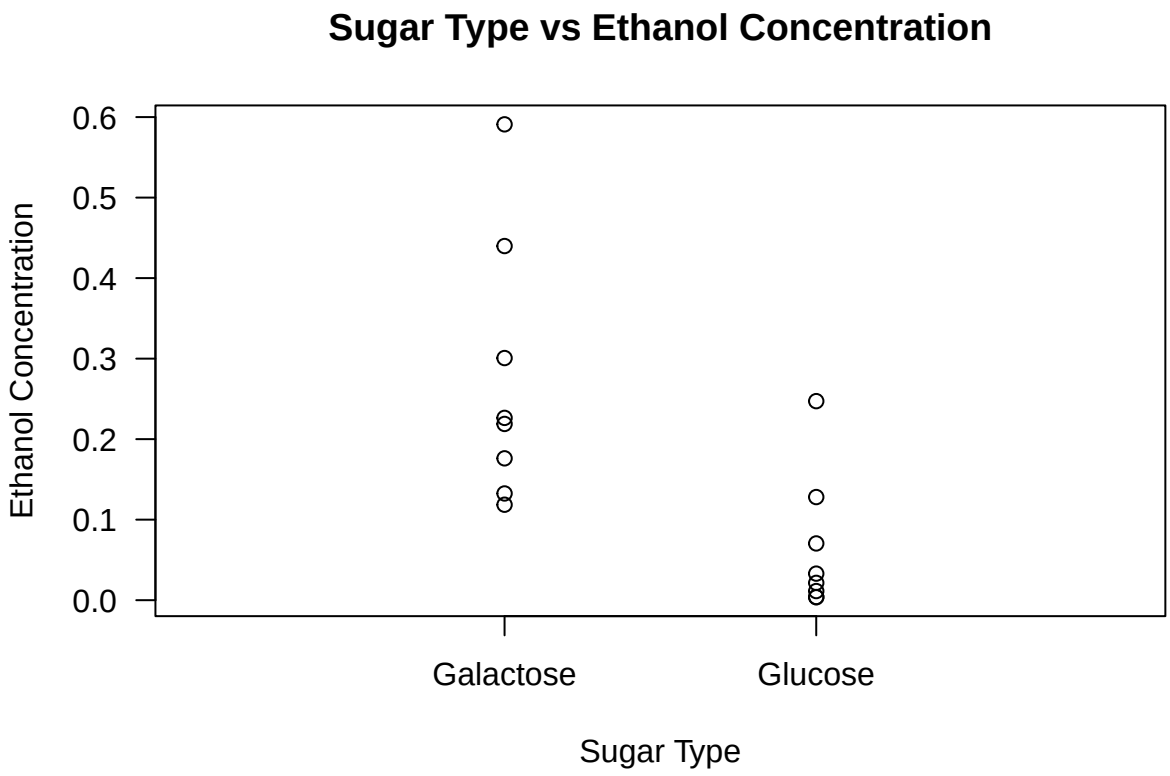
```
str(fermentation)
```

```
## 'data.frame':    16 obs. of  3 variables:
##  $ ethanol: num  0.59 0.3 0.25 0.03 0.44 0.18 0.13 0.02 0.22 0.23 ...
##  $ oxygen : int  0 0 0 0 46 46 46 46 92 92 ...
##  $ sugar  : chr  "Galactose" "Galactose" "Glucose" "Glucose" ...
```

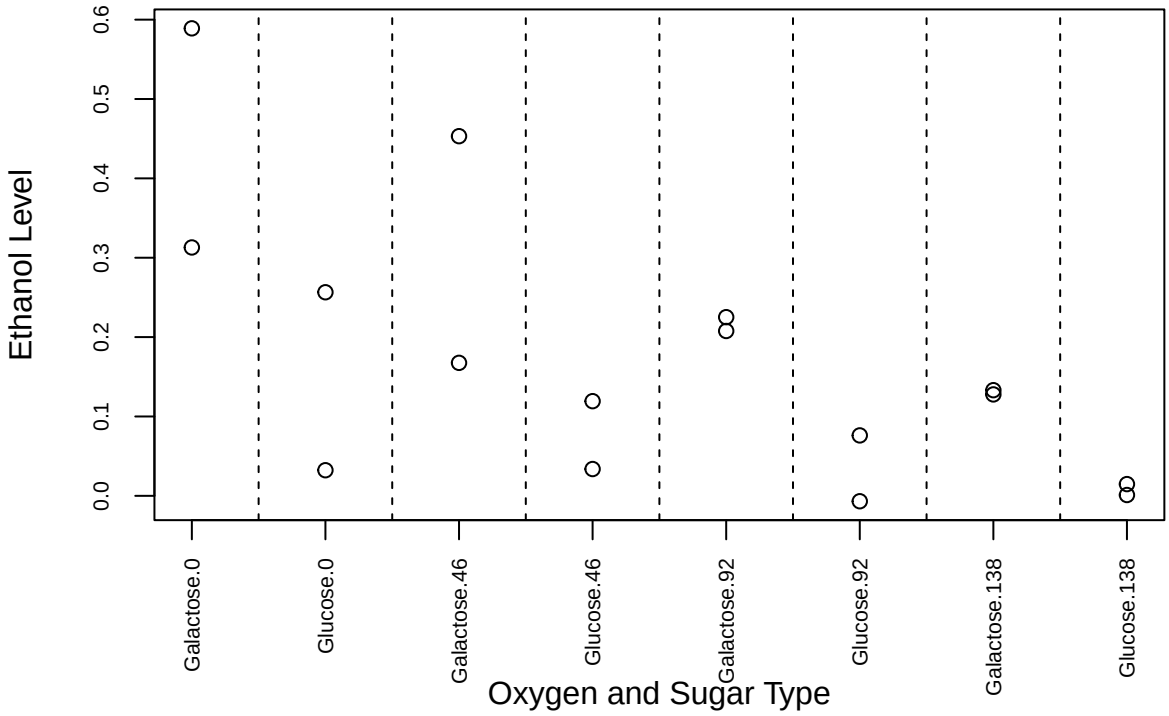
```
# Our result is ethanol value, we want to turn oxygen concentration and sugar
# type into a factor
oxygen_concentration <- factor(fermentation$oxygen)
sugar_type <- factor(fermentation$sugar)

# Plot the datasets as before
# First, by Oxygen Concentration
par(las=1)
stripchart(jitter(ethanol, factor=4) ~ oxygen_concentration,
           data=fermentation, xlab="Oxygen Concentration",
           ylab="Ethanol Concentration",
           main="Oxygen Concentration vs Ethanol Concentration",
           vertical=TRUE, pch=1)
abline(v=c(1.5:3.5), lty=2)
```




```
# Next, by sugar type
stripchart(jitter(ethanol, factor=2) ~ sugar_type,
           data=fermentation, xlab="Sugar Type",
           ylab="Ethanol Concentration",
           main="Sugar Type vs Ethanol Concentration",
           vertical=TRUE, pch=1, xlim=c(0,3))
```




```r
# Next, by sugar type and oxygen concentration
par(las=3)
stripchart(split(
  jitter(fermentation$ethanol, factor=8), list(fermentation$sugar, fermentation$oxygen)),
  cex.axis=0.65, ylab="Ethanol Level",
  main="Ethanol Concentration vs Oxygen Concentration & Sugar Type",
  pch=1, vertical=TRUE)
# Without specifying this separately, the vertical ticks will overlap upon xlab
title(xlab = "Oxygen and Sugar Type", mgp = c(4, 1, 0))
abline(v=c(1.5:8.5), lty=2)
```



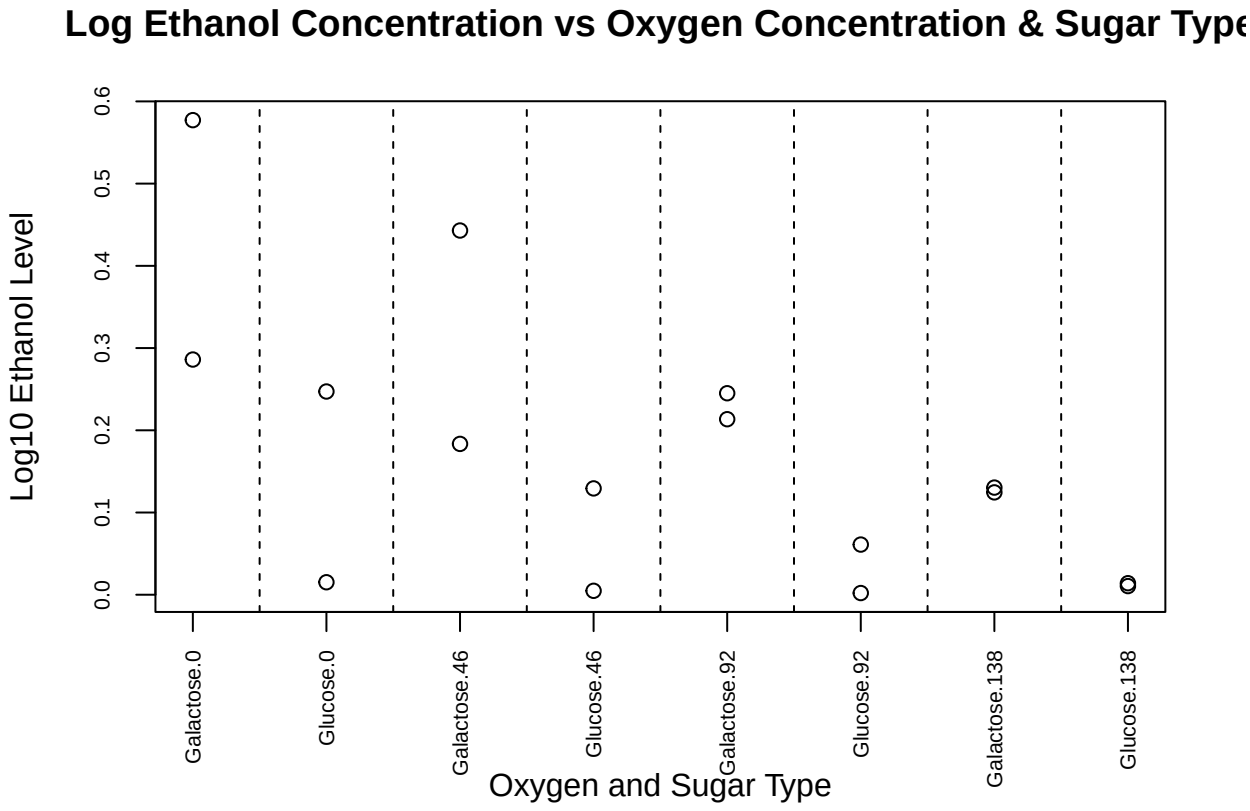

```
# It certainly appears that there's some different variances in the dataset
# We can perform a transformation like a log transformation. However, doing so
# requires zero values to be changed to something small like 0.001.
# First, create  a new dataset that is log transformed
# It turns out that copies in R are actually pretty cool and  that copying data
# should leave the original object intact
# See: https://stackoverflow.com/a/58392556
log_fermentation <- fermentation
# Replace any zero values in the ethanol concentration
log_fermentation$ethanol <- replace(
  log_fermentation$ethanol, log_fermentation$ethanol == 0, 0.0001)
head(log_fermentation)
```

```
##   ethanol oxygen     sugar
## 1    0.59      0 Galactose
## 2    0.30      0 Galactose
## 3    0.25      0   Glucose
## 4    0.03      0   Glucose
## 5    0.44     46 Galactose
## 6    0.18     46 Galactose
```

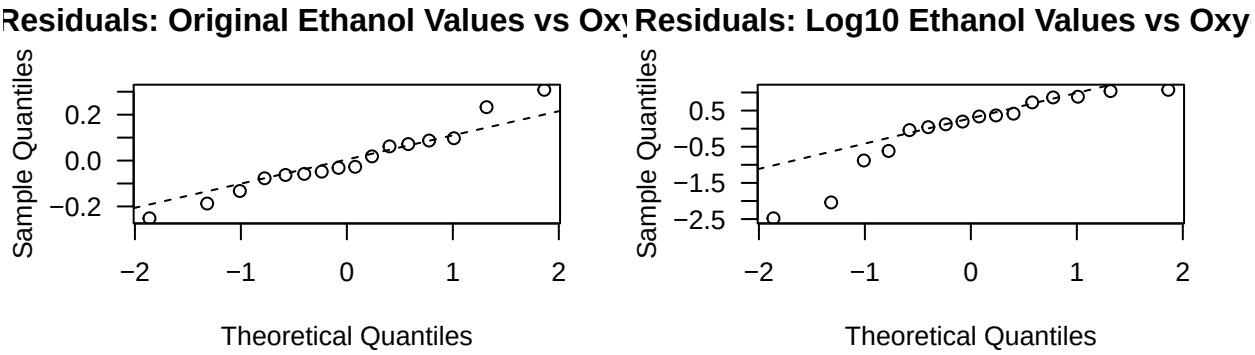
```
# Take log transform
log_fermentation <- cbind(
  log_fermentation,
  logEtOH=log10(log_fermentation$ethanol))
head(log_fermentation)
```

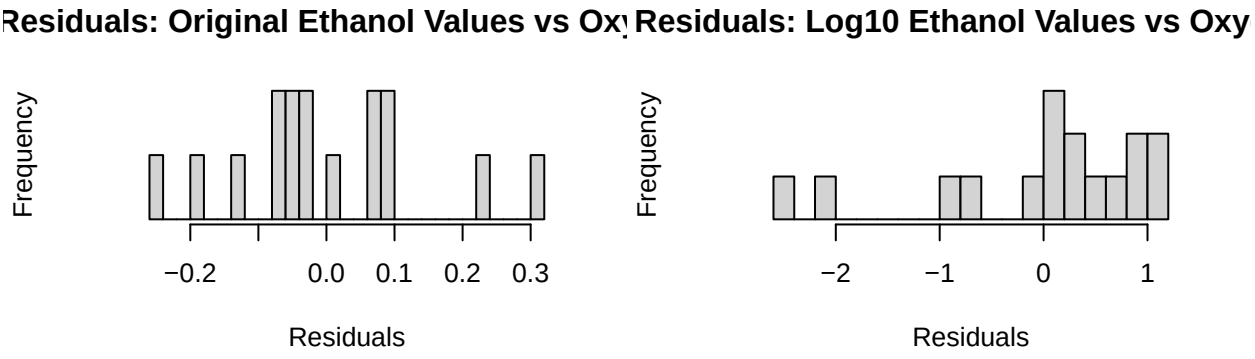
```
##   ethanol oxygen     sugar    logEtOH
## 1    0.59      0 Galactose -0.2291480
## 2    0.30      0 Galactose -0.5228787
## 3    0.25      0   Glucose -0.6020600
## 4    0.03      0   Glucose -1.5228787
## 5    0.44     46 Galactose -0.3565473
## 6    0.18     46 Galactose -0.7447275
```

```
# Plot the log transformed data against all factors to see how variance has
# changed
# Next, by sugar type and oxygen concentration
par(las=3)
stripchart(split(
  jitter(log_fermentation$ethanol, factor=8),
  list(log_fermentation$sugar,
       log_fermentation$oxygen)),
  cex.axis=0.65, ylab="Log10 Ethanol Level",
  main="Log Ethanol Concentration vs Oxygen Concentration & Sugar Type",
  pch=1, vertical=TRUE)
# Without specifying this separately, the vertical ticks will overlap upon xlab
title(xlab = "Oxygen and Sugar Type", mgp = c(4, 1, 0))
abline(v=c(1.5:8.5), lty=2)
```



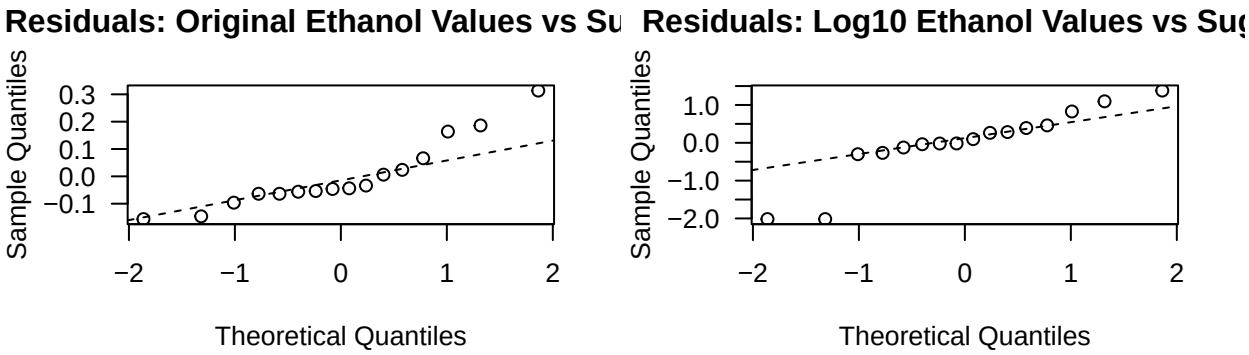
```r
# The data does not appear to be extremely well transformed, but the variances
# within each level appears to be some what more evenly distributed.
# We can confirm this with the use of QQ plots/histograms

# First, oxygen concentration
par(las=1, mfcol=c(2,2))
etoh.aov <- aov(ethanol ~ oxygen, data=log_fermentation)
qqnorm(etoh.aov$residuals, main="Residuals: Original Ethanol Values vs Oxygen")
qqline(etoh.aov$residuals, lty=2, lwd=1)
hist(etoh.aov$residuals, breaks=20, yaxt="n", xlab="Residuals",
     main="Residuals: Original Ethanol Values vs Oxygen")

logetoh.aov <- aov(logEtOH ~ oxygen, data=log_fermentation)
qqnorm(logetoh.aov$residuals, main="Residuals: Log10 Ethanol Values vs Oxygen")
qqline(logetoh.aov$residuals, lty=2, lwd=1)
hist(logetoh.aov$residuals, breaks=20, yaxt="n", xlab="Residuals",
     main="Residuals: Log10 Ethanol Values vs Oxygen")
```

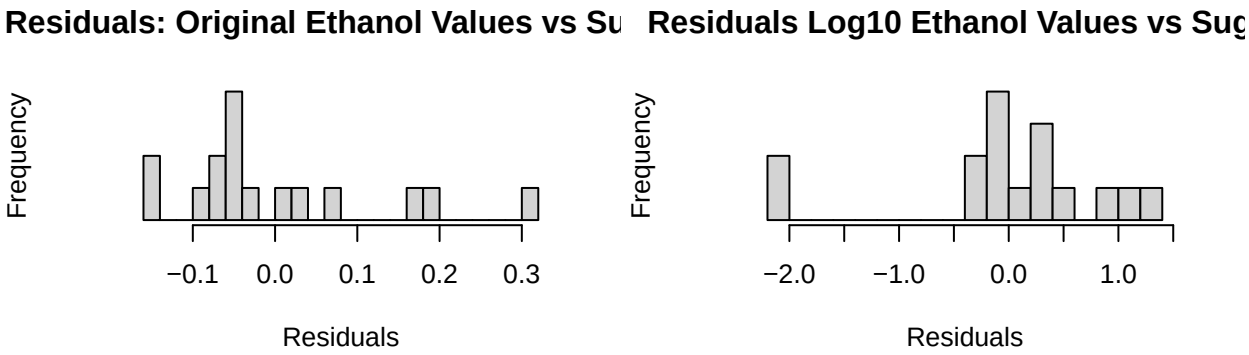





```r
# Now, sugar types
par(las=1, mfcol=c(2,2))
etoh.aov <- aov(ethanol ~ sugar, data=log_fermentation)
qqnorm(etoh.aov$residuals, main="Residuals: Original Ethanol Values vs Sugar")
qqline(etoh.aov$residuals, lty=2, lwd=1)
hist(etoh.aov$residuals, breaks=20, yaxt="n", xlab="Residuals",
     main="Residuals: Original Ethanol Values vs Sugar")
```

```
logetoh.aov <- aov(logEtOH ~ sugar, data=log_fermentation)
qqnorm(logetoh.aov$residuals, main="Residuals: Log10 Ethanol Values vs Sugar")
qqline(logetoh.aov$residuals, lty=2, lwd=1)
hist(logetoh.aov$residuals, breaks=20, yaxt="n", xlab="Residuals",
     main="Residuals Log10 Ethanol Values vs Sugar")
```






```
# It will be better to use the original data given our plots. We can
# perform an AOV simply with the original data
summary(aov(ethanol ~ oxygen_concentration * sugar_type, data = fermentation))
```

```
##                                Df  Sum Sq Mean Sq F value  Pr(>F)
## oxygen_concentration            3 0.11255 0.03752   2.761 0.11147
## sugar_type                      1 0.18063 0.18063  13.293 0.00653 **
## oxygen_concentration:sugar_type 3 0.01813 0.00604   0.445 0.72769
## Residuals                       8 0.10870 0.01359
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It appears that our dataset was not especially helped by the log transformation undertaken in our analysis. There appears to be several somewhat extreme values in our log transformation data that yields a more skewed appearing distribution of residuals in both the oxygen concentration and sugar type. Further, the data appears to be somewhat better centered around zero with the use of the original dataset.

Thus, I decided to perform a two-way ANOVA including an interaction between Oxygen and Sugar Type as these are each fixed effects that are independent of one another.

Thus, at a significance $\alpha = 0.05$, the sugar type has an effect upon ethanol concentration while oxygen concentration does not have a significant effect. There is also insufficient evidence of an interaction between sugar type and oxygen concentration upon sugar levels. Thus, we can reject the null hypothesis that there is no influence of oxygen or sugar type upon ethanol concentration.