Lesson 16 Polyphenols and Grapes (4.2-4.3)

CDT I.M.Smrt

```
# Load packages
library(tidyverse)
library(ggResidpanel)
```

Review

Up to this point we have been using statistical models of the form:

Grape Seeds:

Note here our researchers are interested in explaining the variation in the amount of proanthocyanidin (PC) in a grape seed and the sources of the explained variation might be thought of as the percentage of ethanol.

```
grapes <- read.table("https://raw.githubusercontent.com/jkstarling/MA376/main/Polyphenols.txt", header =

# setwd("C:/Users/james.starling/OneDrive - West Point/Teaching/MA376/JimsLessons_AY23-1/Block III/LSN1
# grapes <-read.table("Polyphenols.txt", header = TRUE, stringsAsFactors = TRUE)

grapes <- grapes %>% mutate(Ethanolf = as.factor(Ethanol.))
```

0) Lets plot the data. Plot `PC` vs `Ethanol.`. What do you see?

```
# gr.means <- grapes %>% group_by(Ethanolf) %>% summarise(mn = mean(PC))
#
# grapes %>% ggplot(aes(x=Ethanolf, y = PC)) + geom_point() +
#   geom_point(aes(x=Ethanolf, y = mn), color="red", size=4, data = gr.means)
```

1) We can see from the plot that there may be different means for the PC values at different ethanol values. Create a multiple-means model to see if we can determine there is a difference. What can we determine about the results?

```
# contrasts(grapes$Ethanolf) = contr.sum
# grape.lm <- grapes %>% lm(xxx ~ xxx, data = .)
# summary(grape.lm)
# anova(grape.lm)
```

2) What else could we try...? If our means had a linear relationship, what model could we try? Write down the model here. What is the fitted (or predicted model)?

3) Fit the linear model using R. How do we interpret the outcome?

```
# grape.reg.lm <- grapes %>% lm(xxx ~ xxxx, data = .)
# summary(grape.reg.lm)
```

4) One often overlooked aspect of using regression vs ANOVA is that the linear regression model is actually more restrictive than the separate means model. Compare the anova tables for the regression model with the multiple-means model. How does our R^2/residual standard error/SSError compare?

```
# anova(xxxx)
# anova(xxxx)
```

Note: When we use a regression model vs a separate means model we are making a trade-off. We are actually building a simpler model, but hoping that the additional complexity a multiple means model provides minimal difference in explaining variability. All things being equal we always prefer a simpler model. If we have two models that have similar SSError values, choose the model that uses fewer degrees of freedom. It isn't always clear whether a separate means model outperforms a linear regression model.

5) The assumption that we are making is that our group means are linearly related to each other. If we have a scientific reason for believing that Ethanol is linearly related to PC than it certainly would be advantageous to use a linear regression model over a group means (or cell means) model.

What is our null and alternative hypothesis under this model?

H_0:

H_a:

6) Another way to think about this hypothesis is that we are testing whether the linear regression model explains more variation (statistically speaking) than the null model. Recall that the null model is:

One statistic that our book starts with that can be used to test this hypothesis is our R^2 statistic. Remember that R^2 is a measurement of how much of our variation can be explained through our explanatory variables.

If Ethanol didn't matter (H_0 is true) how can we see how rare it would be that we observed our R^2?

ANS: SIMULATE!!!

```
# my.rsq <- summary(grape.reg.lm)$'r.squared'
#
# set.seed(376)
# M <- 1000
# stat <- rep(NA, M)  # create empty vector
# for(j in 1:M){
#   grapes$eth.shuff <- sample(grapes$Ethanol.)
#   shuff.lm <- grapes %>% lm(PC ~ eth.shuff, data = .)
#   stat[j] <- summary(shuff.lm)$'r.squared'
# }
#
# statt <- data.frame(stat)
```

So… how rare is our observed statistic?

```
# statt %>% ggplot(aes(x=stat)) + geom_histogram() +
#   geom_vline(xintercept = my.rsq, color="red")
#
# xxxxxx
```

7) Can we think of a better statistic to use? Write out the null/alternative hypothesis for a different statistic.

H_0:

H_a:

8) Run a simulation to estimate the new statistic. How rare is our statistic? (note: we assume a two-sided test).

```
# my.beta1 <- unname(coef(grape.reg.lm)[2])

# set.seed(376)
# M <- 1000
# stat <- rep(NA, M)  # create empty vector
# for(j in 1:M){
#   grapes$eth.shuff <- sample(grapes$Ethanol.)
#   shuff.lm <- grapes %>% lm(PC ~ eth.shuff, data = .)
#   stat[j] <- coef(shuff.lm)[2]
# }
#
# statt <- data.frame(stat)
#
# statt %>% ggplot(aes(x=stat)) + geom_histogram() +
#   geom_vline(xintercept = my.beta1, color="red") + geom_vline(xintercept = -my.beta1, color="red")
#
# (sum(stat >= my.rsq) + sum(stat <= -my.rsq))/ M
```

9) Can we estimate an approximate $t-$statistic from the simulation?

t = (sample slope - hypothesized value of slope) / (std. deviation of null distribution of slope) = (2.502 - 0) / 1.418 = 1.764

```
# sd(xxx)
# 2.502 / sd(xxx)
```

10) But.... if our validity conditions are met, we can actually know the distribution of $\hat{\beta}_1$. What are the validity conditions?

L -

I -

N -

E -

11) Of these, the ones that really matter are outliers and independence (in my opinion), we are relatively robust otherwise. The validity conditions can be wrapped up into $\epsilon_{ij}\ldots$ but we can never actually observe these so we can estimate it from the residuals ($r_{ij} = y_{ij} - \hat{y}_{ij}$). Check the residuals and plot below. Explain what you see.

```
# resids <- xxx
# yhat <- xxxx
# qqnorm(resids)
# qqline(resids)
```

12) What if we use ANOVA? What do we see?

Using the f-stat, calculate the p-value. Can you compare the results of a t-test with the F-statistic from the ANOVA table?

note: $t = \hat{\beta}_1/SE(\hat{\beta}_1)$ where $SE(\hat{\beta}_1) = SE\ of\ residuals/(\sqrt{n-1}SD(X))$.

```
# anova(grape.reg.lm)
#
# # f-stat
# 1 - pf(xxxx, xxx, xx)
#
```

```
#
# # t-stat
# se.beta1 <- summary(grape.reg.lm)$sigma / xxxxx
# my.tstat <- xxx / xxxx
#
# my.tstat
```

13) Finally, we can calculate the confidence interval of $\hat{\beta}_1$ by

$$\hat{\beta}_1 \pm t^*(SE\ of\ \hat{\beta}_1)$$

```
# tstar <- xxxx
#
# c(xxxxx, xxxx)
#
# # OR
#
# confint(xxxx)
```