Lsn 17 - AY19-2

Clark

Admin

Continuing with our grape seeds.

Recall that the model we were considering was:

The assumption that we are making is that our group means are linearly related to each other. As we mentioned last class, 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 regresison model over a group means (or cell means) model. Note that our H_0 and H_a can be written in words as:

Another way to think about this set of hypothesis (hypotheses?) 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. **Note that we are NOT (at least now) comparing the linear regression model to the group means model, we are only comparing the linear regression model to the null model.**

So, we have our H_0 and H_a, we have our test statistic, now from our data we can calculate the observed statistic.

```
grape<-read.table("http://www.isi-stats.com/isi2/data/Polyphenols.txt",header=T)
grape <- grape %>% mutate(Ethanol=Ethanol.) %>% select(-Ethanol.)
grape <- grape %>% mutate(Time.hrs=Time.hrs.)%>% select(-Time.hrs.)
reg.lm<-lm(PC~Ethanol,data=grape)
our.stat<-summary(reg.lm)$'r.squared'
```

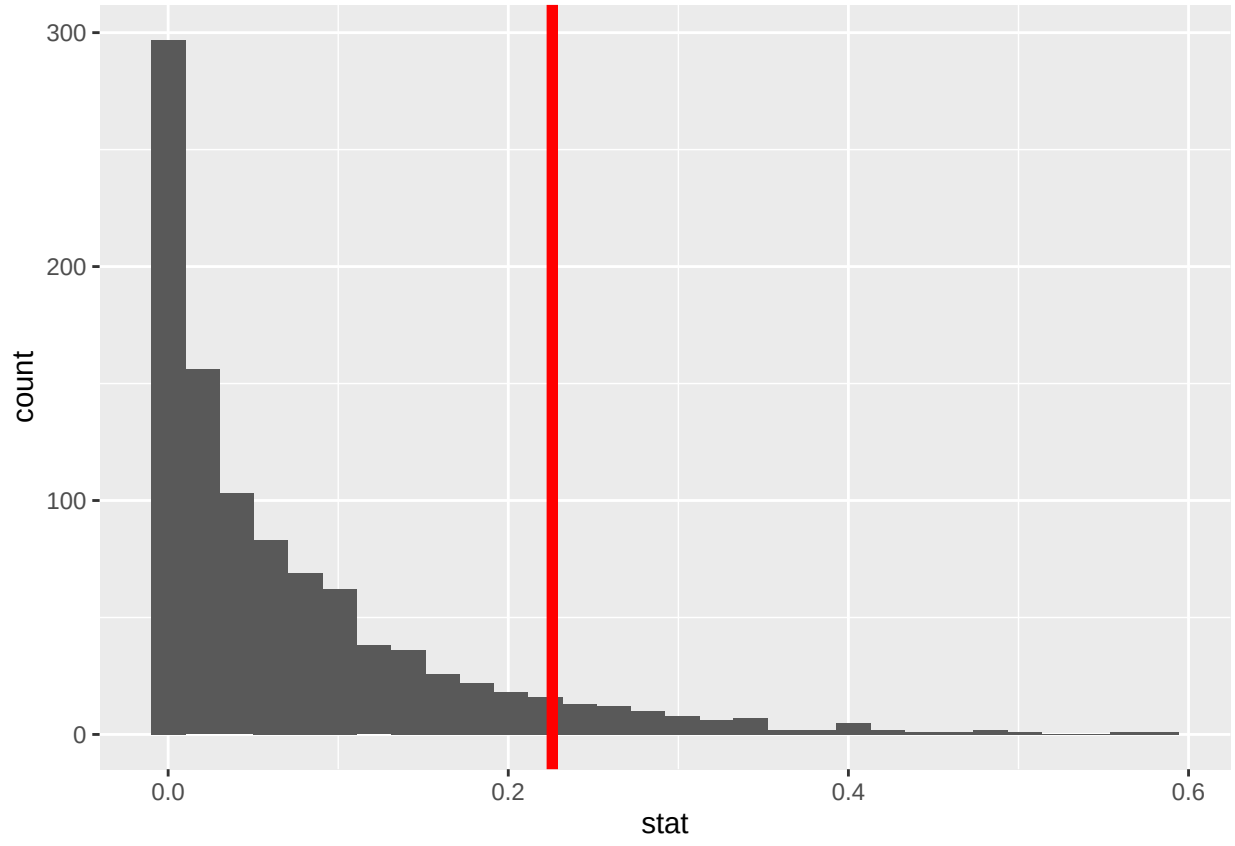
We then can see how rare it would be, if Ethanol didn't matter, that we would have observed *our* R^2.

```
M<-1000
empirical.dist<-data.frame(trial=seq(1,M),stat=NA)
for(i in 1:M){
  grape.shuff<-grape %>% mutate(Ethanol.shuff=sample(Ethanol))
  shuff.lm<-lm(PC~Ethanol.shuff,data=grape.shuff)
  empirical.dist[i,]$stat=summary(shuff.lm)$'r.squared'
}
```

So how rare is our statistic?

```
empirical.dist %>% ggplot(aes(x=stat))+
  geom_histogram()+geom_vline(xintercept=our.stat,color="red",lwd=2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
empirical.dist%>%filter(stat>our.stat)%>%summarize(pval=n()/M)
```

```
##    pval
## 1 0.081
```

This is all well and good, but the distribution of R^2 changes depending on our data (why?). So perhaps this isn't the best statistic to use.

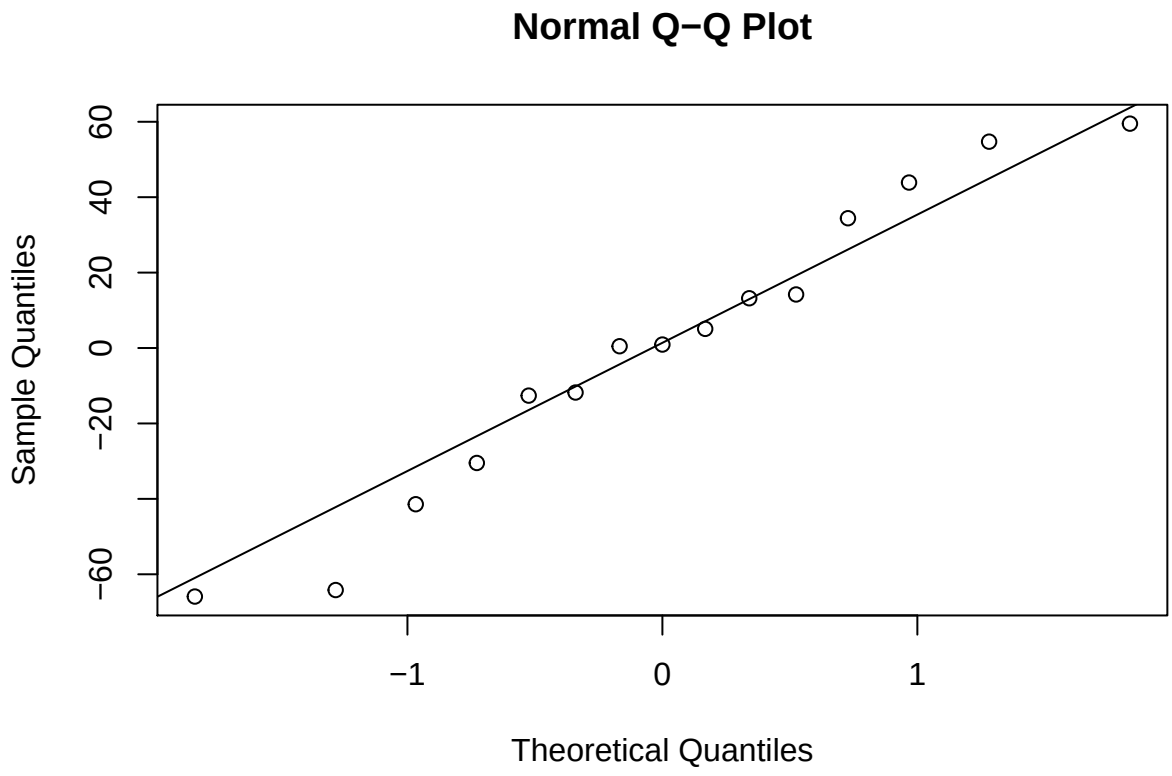
Note that from our model a test comparing the null model to the linear regression model is also a test of:

Therefore, it might make sense to use $\hat{\beta}_1$ as our test statistic. It just so happens that, assuming our validity conditions are met, we **know** the distribution of $\hat{\beta}_1$. The validity conditions are LINE (cute, huh...). 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 $\epsilon_{i,j}$ as:

We can never actually oberve $\epsilon_{i,j}$ but we can estimate it from $r_{i,j} = y_{i,j} - \hat{y}_{i,j}$ where $\hat{y}_{i,j}$ is found from:

This allows us to check Normality and equal variance. To check linearity we can plot the residuals vs. predicted values:

```
resids=reg.lm$residuals
yhat=reg.lm$fitted.values
qqnorm(resids)
qqline(resids)
```




```
#hist(resids)
#plot(yhat,resids)
```

If our conditions are met, then we can form the standardized statistic, which has distribution:

But wait a minute... What about ANOVA? Isn't that the framework we were using to test $\alpha_1 = \alpha_2 = \alpha_3$?

```
anova(reg.lm)
```

```
## Analysis of Variance Table
##
## Response: PC
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Ethanol    1   6260  6260.0  3.7935 0.07338 .
## Residuals 13  21453  1650.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As it turns out, an F-statistic with 1,n degrees of freedom is the square of a t-statistic with n degrees of freedom. So, if we look at our ANOVA output, we have an F statistic of 3.79 that has 1,13 degrees of freedom, which yields a p-values of `1-pf(3.79,1,13)=0.073`, but we can do the same test by testing $\beta_1 = 0$ vs $\beta_1 \neq 0$ using $\hat{\beta}_1$ which has a t distribution with 13 degrees of freedom. A picture:

The goodness of using $\hat{\beta}_1$ vs the F statistic is that the Confidence interval for β_1 that can be formed using $\hat{\beta}_1$ is more intuitive. Recall that the general form of a CI is:

Therefore we need the SE of $\hat{\beta}_1$. Using matrix notation the SE is trivial to find. Outside of the world of linear algebra it's cumbersome and not necessarily that insightful (learn Linear Algebra!!)

However, practically in R we can do:

```
confint(reg.lm,level=0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) -27.7982917 253.398292
## Ethanol      -0.2732065   5.277207
```

So our 95% CI for $\hat{\beta}_1$ is: