Lsn 16 AY23

Clark

Reading Critique, Project, WPR…

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

Key to this has been the assumption that our explanatory variable, or independent variable, is categorical. This was nice in that we could think of each of our observations as coming from a group with separate means. For instance we can think of H_0 and H_a from above as:

However, in some studies our explained variation comes from a variable that has a natural ordering to it. In fact we've seen this a bit already (recall Pistachio study).

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.

```
grape<-read.table("http://www.isi-stats.com/isi2/data/Polyphenols.txt",header=T, stringsAsFactors = T)
grape <- grape %>% mutate(Ethanol=Ethanol.) %>% select(-Ethanol.)
grape <- grape %>% mutate(Time.hrs=Time.hrs.)%>% select(-Time.hrs.)
```

Note that the null model here is:

Which can be fit as:

```
null.lm<-lm(PC~1,data=grape)
summary(null.lm)
```

```
##
## Call:
## lm(formula = PC ~ 1, data = grape)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -89.20 -20.25  12.40  27.60  68.90
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   237.90      11.49   20.71  6.7e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.49 on 14 degrees of freedom
```

Perhaps we fit a separate means model:

Which is done through:

```
sep.means<-lm(PC~0+as.factor(Ethanol),data=grape)
summary(sep.means)
```

```
##
## Call:
## lm(formula = PC ~ 0 + as.factor(Ethanol), data = grape)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67.74 -21.60   1.84  23.85  57.66
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## as.factor(Ethanol)40    212.0       18.9   11.22 1.02e-07 ***
## as.factor(Ethanol)50    239.7       18.9   12.69 2.60e-08 ***
## as.factor(Ethanol)60    262.0       18.9   13.86 9.53e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.26 on 12 degrees of freedom
## Multiple R-squared:  0.9756, Adjusted R-squared:  0.9694
## F-statistic: 159.7 on 3 and 12 DF,  p-value: 6.187e-10
```

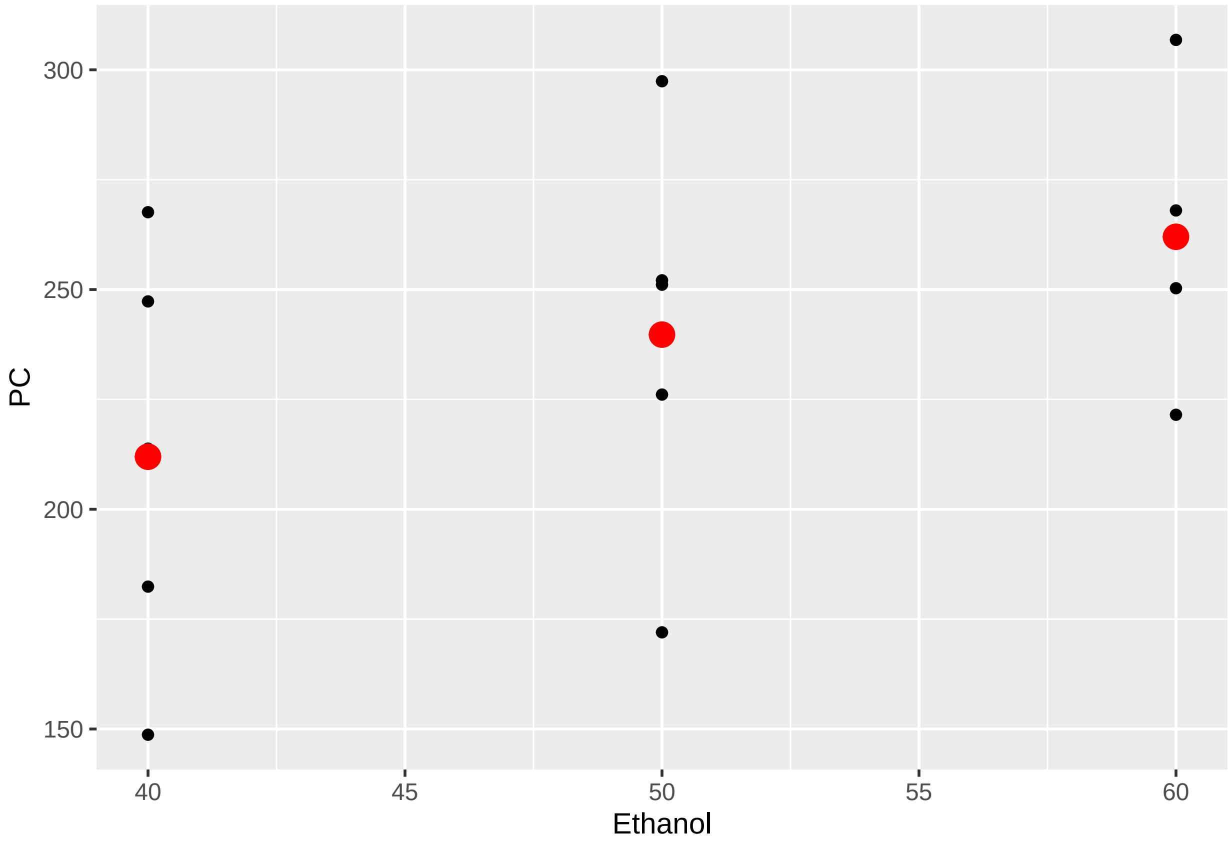
Is this better? Well, perhaps. To test this we can run:

```
grape<-grape %>% mutate(Ethanolf=as.factor(Ethanol))
contrasts(grape$Ethanolf)<-contr.sum
```

```
effects.mod<-lm(PC~Ethanolf,data=grape)
anova(effects.mod)
```

```
## Analysis of Variance Table
##
## Response: PC
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Ethanolf   2  6285.4  3142.7    1.76 0.2137
## Residuals 12 21427.1  1785.6
```

```
gr.means <- grape %>% group_by(Ethanolf)%>%summarize(means=mean(PC))%>%
  mutate(Ethanol=as.numeric(as.character(Ethanolf)))
grape %>% ggplot(aes(x=Ethanol,y=PC))+geom_point()+geom_point(aes(x=Ethanol,y=means),color="red",size=4
```



If our means have a linear relationship, perhaps our model could be:

Which is of course the model for a regression. Note that this is the statistical model. The fitted model or the predicted model is:

Our book uses b_0 and b_1 instead of $\hat{\beta}_0$ and $\hat{\beta}_1$. I prefer the hats, but use what you'd like.

To fit this model we simply run:

```
reg.lm<-lm(PC~Ethanol,data=grape)
summary(reg.lm)
```

```
##
## Call:
## lm(formula = PC ~ Ethanol, data = grape)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -65.90 -21.55   0.92  24.31  59.50
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  112.800     65.081   1.733   0.1067
## Ethanol        2.502      1.285   1.948   0.0734 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.62 on 13 degrees of freedom
## Multiple R-squared:  0.2259, Adjusted R-squared:  0.1663
## F-statistic: 3.793 on 1 and 13 DF,  p-value: 0.07338
```

Note that we obtain $\hat{\beta}$ through the method of least squares:

The interpretation of $\hat{\beta}_0$ and $\hat{\beta}_1$ are:

One often overlooked aspect of using regression vs ANOVA is that the linear regression model is actually more restrictive than the seperate means model:

```
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
```

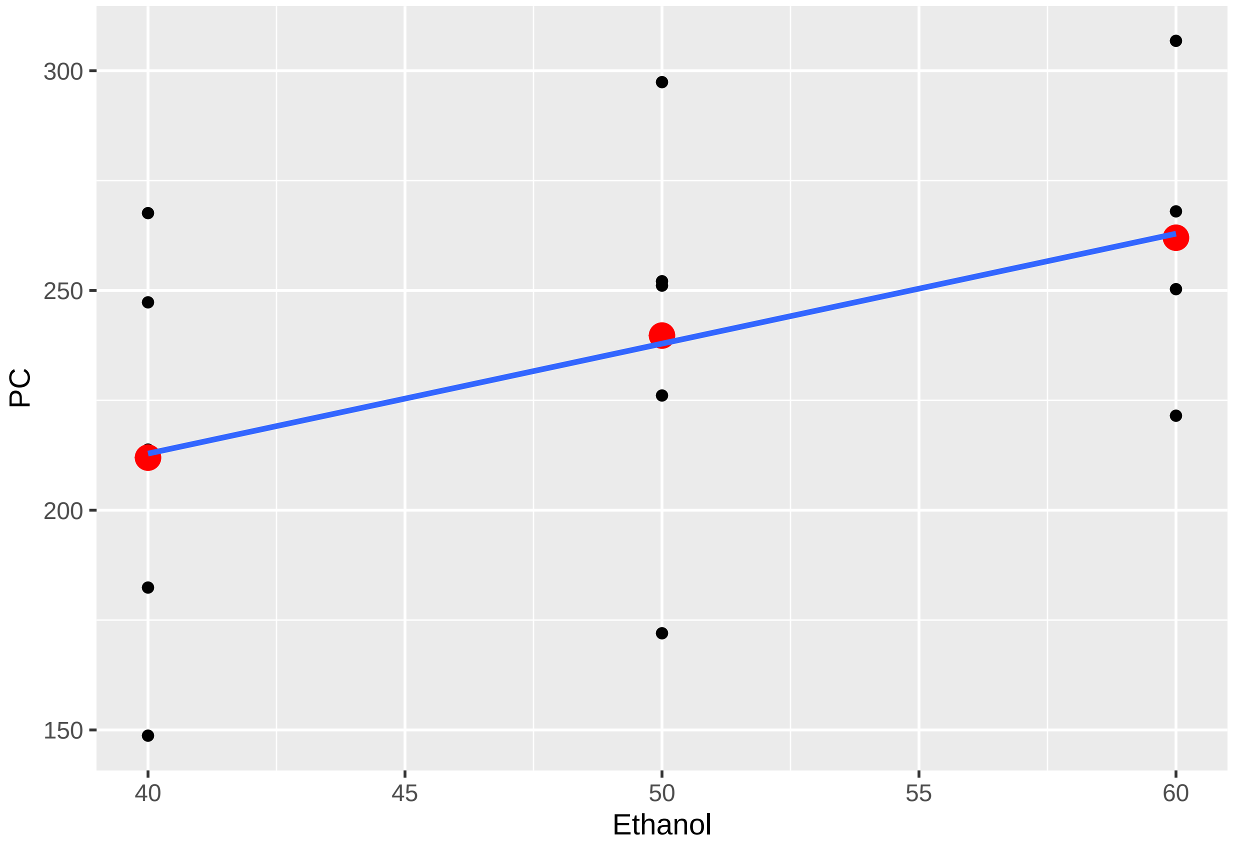
Note that the sums of squares are higher under this model. Why?

However, the SE of the residuals is smaller. When we use a regression model vs a seperate means model we are making a tradeoff. 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 if you don't have science to save you.

To put this another way, previously if we had μ, α_1 and α_2 we automatically knew α_3 (why?). Now, if we know one of our means and β_1 we know all of our other means.

```
grape %>% ggplot(aes(x=Ethanol,y=PC))+geom_point()+
  geom_point(aes(x=Ethanol,y=means),color="red",size=4,data=gr.means)+
  stat_smooth(method="lm", se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

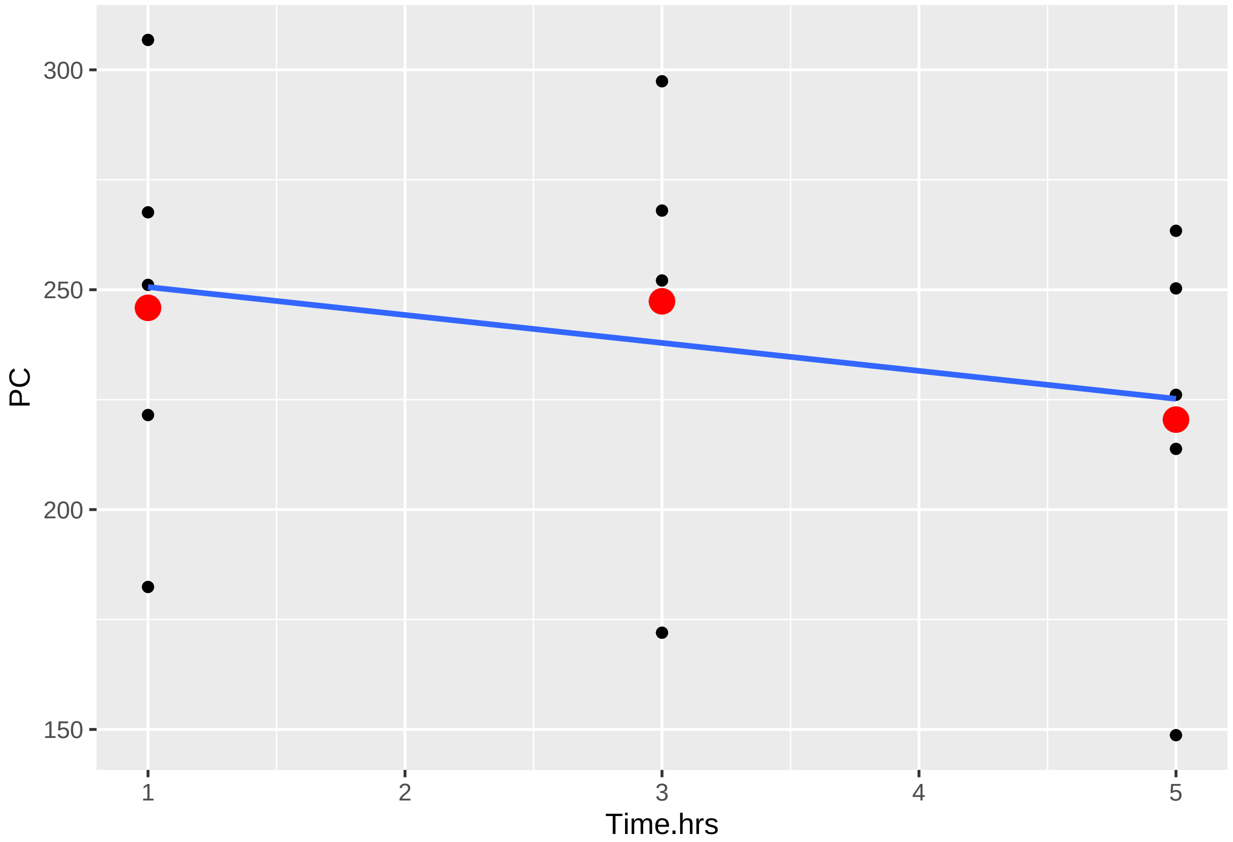


In later lessons we'll explore the statistical properties of $\hat{\beta}$ and the linear regression model. The key point here is that a linear regression model is appropriate if we believe there is a linear relationship between our explanatory variable and our response. It isn't always clear whether a separate means model outperforms a linear regression model though:

```
grape<-grape %>% mutate(Timef=as.factor(Time.hrs))
reg.lm<-lm(PC~Time.hrs,data=grape)
sep.means<-lm(PC~Timef,data=grape)
```

```
gr.means <- grape %>% group_by(Timef)%>%summarize(means=mean(PC))%>%
  mutate(Time=as.numeric(as.character(Timef)))
grape %>% ggplot(aes(x=Time.hrs,y=PC))+geom_point()+
  geom_point(aes(x=Time,y=means),color="red",size=4,data=gr.means)+
  stat_smooth(method="lm",se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



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.

```
reg.lm<-lm(PC~Ethanol,data=grape)
our.stat<-summary(reg.lm)$'r.squared'
```

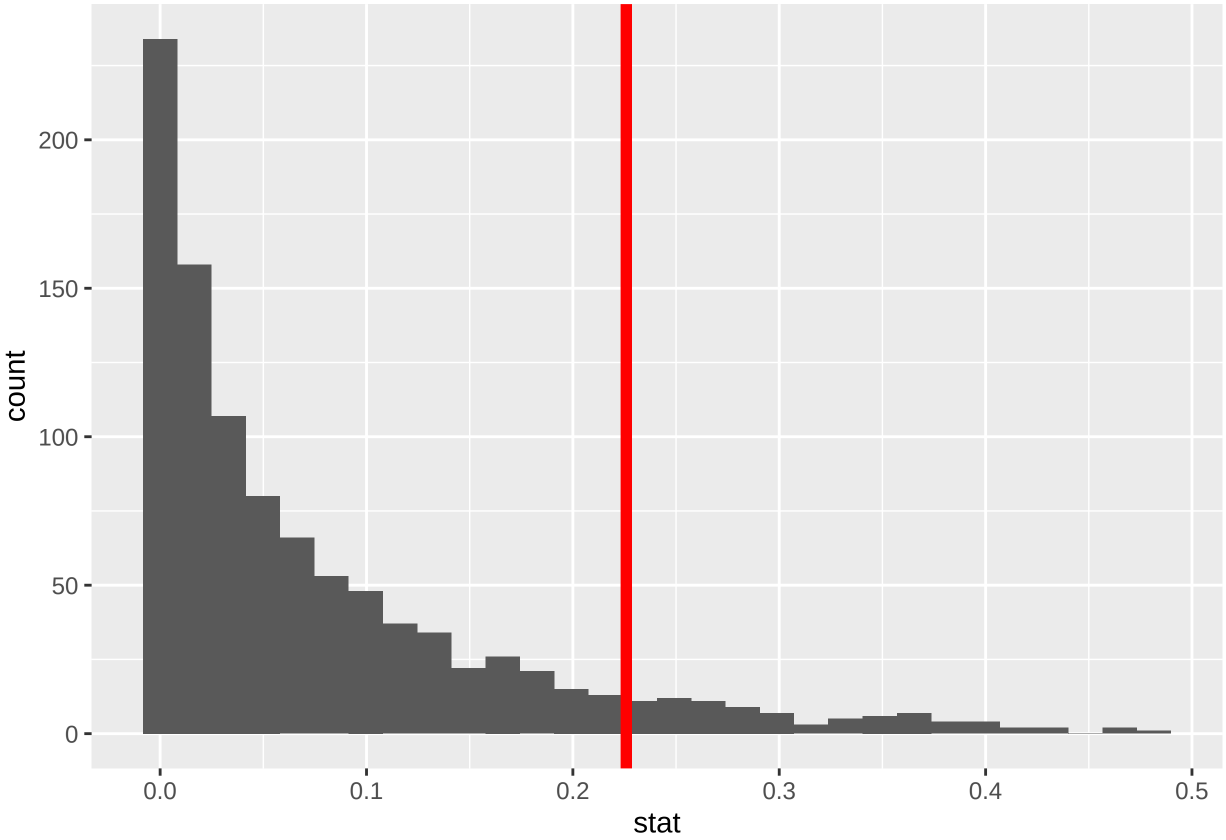
We then can see how rare it would be, if Ethanol didn't matter, that we would have observed *our* R^2.

```
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 0.084
```

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.

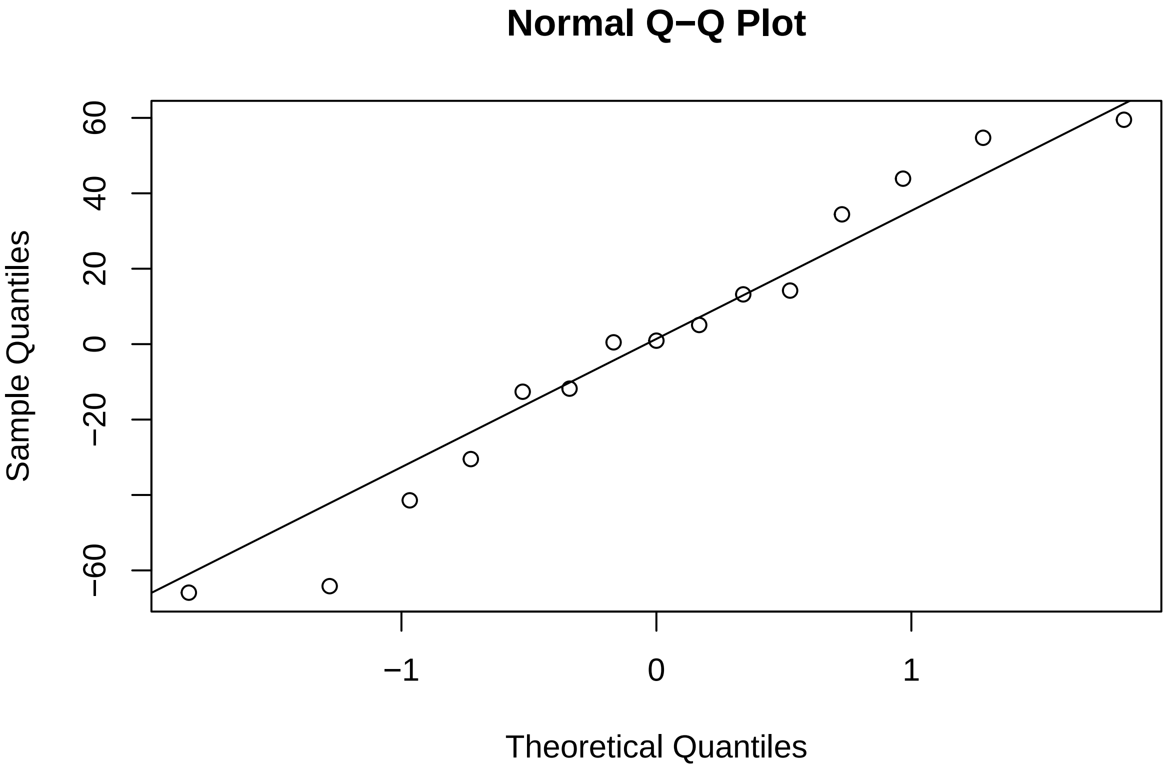
Note that from our model a test comparing the null model to the linear regression model is also a test of:

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: