Lesson 11

Clark

Recall that earlier in the course we discussed confounding, which is

In today's lesson (which adittedly is a bit dense) we are going to go through how confounding can make life difficult and impact our analysis of variance model.

The primary research quesiton we are going to explore is whether wages for blacks differ significantly fromw ages for non-boacks focusing on males who went to college and males who did not go to college.

The initial statistical model we consider is:

We can find the group means by:

```
dat<-read.table("http://www.isi-stats.com/isi2/data/WageSubset.txt",header=T)
dat %>% group_by(race,education)%>%summarise(avg=mean(wage.100))
```

```
## # A tibble: 4 x 3
## # Groups:   race [2]
##   race     education       avg
##   <fct>    <fct>         <dbl>
## 1 black    belowCollege   4.18
## 2 black    beyondCollege  8.47
## 3 nonblack belowCollege   5.41
## 4 nonblack beyondCollege  9.71
```

```
mean(dat$wage.100)
```

```
## [1] 6.062337
```

Note here that the overall mean is a lot closer to nonblack than it is to black. Why?

Therefore we might not want μ in our model to represent the overall average, but rather the average of the group averages, or $(4.52 + 6.21)/2$. In R this is done when we fix our contrasts as contr.sum

```r
dat<-read.table("http://www.isi-stats.com/isi2/data/WageSubset.txt",header=T)
contrasts(dat$race)=contr.sum
contrasts(dat$education)=contr.sum
anova_model2<-lm(wage.100~race,data=dat)
full.bets<-anova_model2$coefficients
full.bets
```

```
## (Intercept)       race1
##   5.3628375  -0.8424549
```

Again μ is NOT the population average, but the **effect average**. Why might we want to do this

Looking at page 175 obviously we might want to explain some of the unexplained variation using college as a factor. The real issue becomes this:

```r
dat %>% group_by(race,education)%>%summarise(num.obs=n())
```

```
## # A tibble: 4 x 3
## # Groups:   race [2]
##   race     education     num.obs
##   <fct>    <fct>           <int>
## 1 black    belowCollege     1301
## 2 black    beyondCollege     112
## 3 nonblack belowCollege    12428
## 4 nonblack beyondCollege    2813
```

So let's do what we did before while ignoring the fact that our samples are unequal.

```r
dat<-read.table("http://www.isi-stats.com/isi2/data/WageSubset.txt",header=T)
dat %>% group_by(education)%>%summarise(avg=mean(wage.100))
```

```
## # A tibble: 2 x 2
##   education       avg
##   <fct>         <dbl>
## 1 belowCollege   5.30
## 2 beyondCollege  9.66
```

Therefore the means of the means is 7.477 and the effect of education is ± 2.181. So perhaps we are tempted to our adjusted statistical model as:

Which we could then analyze via:

```r
dat.adj = dat %>% mutate(adj.val=ifelse(education=="belowCollege",wage.100+2.181,wage.100-2.181))

adj.mod<-lm(adj.val~race,data=dat.adj)

anova(adj.mod)
```

```
## Analysis of Variance Table
##
## Response: adj.val
##              Df Sum Sq Mean Sq F value    Pr(>F)
## race          1   1942 1942.30  129.58 < 2.2e-16 ***
## Residuals 16652 249593   14.99
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Which seems like it should work, right? This is just what we were doing before, what's the problem?

This is, in essence, confounding. When we subtract off the "College effect" we are also subtracting off some part of the education effect. Why?

In the parlance of ANOVA, up to this point we have been calculating what are called "Type I Sums of Squares". These are done sequentally. We first find the Sums of Squares due to factor A and then find the Sums of Squares due to factor B given than factor A is in the model. We can see this because if we run:

```r
forward<-lm(wage.100~race+education,data=dat)
anova(forward)
```

```
## Analysis of Variance Table
##
## Response: wage.100
##              Df Sum Sq Mean Sq F value    Pr(>F)
## race          1   3671    3671  244.92 < 2.2e-16 ***
## education     1  44156   44156 2945.93 < 2.2e-16 ***
## Residuals 16651 249581      15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
backward<-lm(wage.100~education+race,data=dat)
anova(backward)
```

```
## Analysis of Variance Table
```

```
##
## Response: wage.100
##              Df Sum Sq Mean Sq F value    Pr(>F)
## education     1  45873   45873 3060.48 < 2.2e-16 ***
## race          1   1954    1954  130.36 < 2.2e-16 ***
## Residuals 16651 249581      15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our Sums of Squares change. **This is because the only time we are doing "Conditional Sums of Squares" is when our variable is the second variable in the model**

To further see that education and race are covariated, we note that by knowing someone's education we have information on race. Further, by knowing education we have information on wage.

To reflect covariance in our model we draw our diagram like:

Note that our statistical model doesn't change, but to fit this in R we need the `library(car)` installed and we can run:

```r
library(car)
contrasts(dat$race)=contr.sum
contrasts(dat$education)=contr.sum
anova_model2<-lm(wage.100~education+race,data=dat)
anova.table<-Anova(anova_model2,type=3)
anova.table
```

```
## Anova Table (Type III tests)
##
## Response: wage.100
##             Sum Sq    Df  F value    Pr(>F)
## (Intercept) 192915     1 12870.43 < 2.2e-16 ***
## education    44156     1  2945.93 < 2.2e-16 ***
## race          1954     1   130.36 < 2.2e-16 ***
## Residuals   249581 16651
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

An interesting note here is that the sums of squares no longer equal the total sums of squares. The extra sums of squares can be thought of as variation that cannot be disentangled from education or race. Our book calls this SScovariation, which I rather like. It's variability that still exists but we cannot attribute to either factor so we basically shrug our shoulders.

Type I Sums of Squares vs Type III Sums of Squares

- Type I Calculations Make Sense

- Type I prefereable when order matters (More concerned with one of the factors First)

- Type I Order Matters!

- Type III all effects are conditional on *everything else in the model*

- Type III not sample size dependent
- Type III are NOT additive