Lesson 11

Clark

September 25, 2019

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

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

The primary research quesiton we are going to explore si 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 in our model to represent the overall average, but rather the average of the group averages, or . In R this is done when we fix our contrasts as contr.sum

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 is NOT the population average, but the .

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:

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.

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 . So perhaps we are tempted to our adjusted statistical model as:

Which we could then analyze via:

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, covariance. 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:

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

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.

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:

library(car)

## Warning: package 'car' was built under R version 3.5.1

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

## The following object is masked from 'package:purrr':  
##   
## some

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=2)  
anova.table

## Anova Table (Type II tests)  
##   
## Response: wage.100  
## Sum Sq Df F value Pr(>F)   
## 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.