Lsn 15

Clark

Admin

Let's reconsider the Salary Discrimination dataset

```
salary<-read.table("http://www.isi-stats.com/isi2/data/Wages.txt",header=T)
salary.dat<-salary%>%mutate(wage=wage/100)
```

Instead of looking at College educated vs not college educated, we now consider the full dataset.

```
levels(salary.dat$educ)
```

```
## [1] "belowHS"       "beyondCollege" "college"       "HS"
```

```
gr.means=salary.dat%>%group_by(educ,race)%>%summarize(mean.salary=mean(wage))
```

What do we see?

```
gr.means$educ<-factor(gr.means$educ,levels=c("belowHS","HS","college","beyondCollege"))
gr.means %>% ggplot(aes(x=educ,y=mean.salary,color=race))+
  geom_line(aes(group=race),lwd=2)+geom_point()
```

A statistical model:

Shell ANOVA table:

To fit the model we use:

```r
contrasts(salary.dat$race)=contr.sum
contrasts(salary.dat$educ)=contr.sum
inter.lm<-lm(wage~race*educ,data=salary.dat)
coef(inter.lm)
```

```
## (Intercept)       race1       educ1       educ2       educ3 race1:educ1
##  6.05387773 -0.59514079 -2.12189614  3.03256089  0.23755534  0.33916246
## race1:educ2 race1:educ3
## -0.02381393 -0.25316463
```

Getting the fits is a bit of a pain but we can do it:

To fit the ANOVA model we note that we are now interested in Type III Sums of squares. Why?

```r
library(car)
```

```
## 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
```

```r
Anova(inter.lm,type=3)
```

```
## Anova Table (Type III tests)
##
## Response: wage
##             Sum Sq   Df   F value    Pr(>F)
## (Intercept) 121306     1 6926.7761 < 2.2e-16 ***
```

```
## race           1172     1   66.9427 2.924e-16 ***
## educ           9718     3  184.9630 < 2.2e-16 ***
## race:educ       164     3    3.1247   0.02473 *
## Residuals    448727 25623
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Why are there 3 df for education and the interaction?

Consider the ANOVA table without the interaction

```
 no.inter.lm<-lm(wage~race+educ,data=salary.dat)
Anova(no.inter.lm,type=3)
```

```
## Anova Table (Type III tests)
##
## Response: wage
##             Sum Sq    Df  F value    Pr(>F)
## (Intercept) 205837     1 11750.66 < 2.2e-16 ***
## race          3383     1   193.10 < 2.2e-16 ***
## educ         51397     3   978.03 < 2.2e-16 ***
## Residuals   448891 25626
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If we look at the Residuals line the Sum of Squares went from 448727 to 448891, or a difference in 164 which might make sense, but look at what happens to SS race and ss education. This suggests that the interaction term is confounded with race and education. Does this make sense?

Ultimately, what we wanted to know though is that controlling for education is there a difference in wages. To get at this we can look at the pairwise comparisons. Or in otherwords, recall that the model with an interaction term is the same as the multiple means model:

This allows us to answer questions such as: For individuals with a College degree is there a difference in mean weekly wages for blacks and nonblacks.

```
pair.diff<-TukeyHSD(aov(wage~race*educ,data=salary.dat))
pair.diff$`race:educ`[20:25,]
```

```
##                                               diff        lwr        upr
## nonblack:college-nonblack:beyondCollege -2.5656549 -2.8419467 -2.2893630
## black:HS-nonblack:beyondCollege         -5.4570604 -5.9014109 -5.0127099
## nonblack:HS-nonblack:beyondCollege      -4.1424110 -4.4102142 -3.8746079
## nonblack:college-black:college           1.6966109  1.1498656  2.2433561
## black:HS-black:college                  -1.1947947 -1.8429000 -0.5466894
## nonblack:HS-black:college                0.1198547 -0.4226503  0.6623597
##                                                p adj
## nonblack:college-nonblack:beyondCollege 0.000000e+00
## black:HS-nonblack:beyondCollege         0.000000e+00
## nonblack:HS-nonblack:beyondCollege      0.000000e+00
## nonblack:college-black:college          7.427392e-14
## black:HS-black:college                  6.415845e-07
## nonblack:HS-black:college               9.977491e-01
```

Note if we fit the model without an interaction term we cannot address this question directly. We see if we run the pairwise comparisons we get:

```
pair.diff2<-TukeyHSD(aov(wage~race+educ,data=salary.dat))
pair.diff2
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = wage ~ race + educ, data = salary.dat)
##
## $race
##                    diff      lwr      upr p adj
## nonblack-black 1.750093 1.558534 1.941651     0
##
## $educ
##                            diff        lwr       upr p adj
## beyondCollege-belowHS  5.410688  5.0698957  5.751480     0
## college-belowHS        2.828847  2.5296934  3.128001     0
## HS-belowHS             1.289537  0.9961514  1.582923     0
## college-beyondCollege -2.581841 -2.8107596 -2.352922     0
## HS-beyondCollege      -4.121151 -4.3424787 -3.899823     0
## HS-college            -1.539310 -1.6887737 -1.389846     0
```