STAT 411/511 - Final Stats Report

Your name

Date

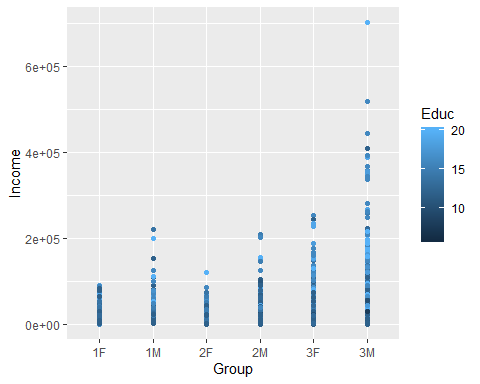
#### Education versus Income

For the final stats report, we will return to the National Longitudinal Youth Survey dataset. See Ch 12, Problems 23-24 (and Display 12.20) for a thorough description of the variables included in this dataset.

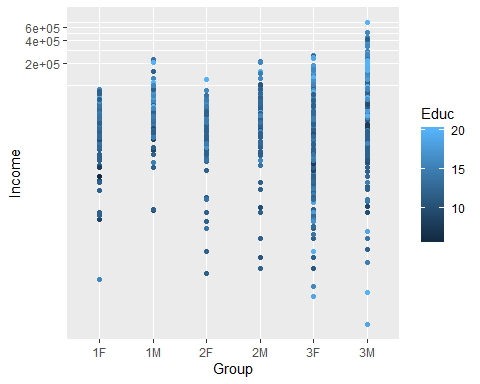
From the Sleuth: “In their 1994 book, The Bell Curve: Intelligence and Class Structure in American Life, psychologist Richard Herstein and political scientist Charles Murray arged that a person’s intelligence is a better predictor of success in life than is education and family’s socioeconomic status. The book was controversial modely for its conclusions about intelligence and racec, but also for the strength of its conclusions drawn from regression analysis on observational data with imperfect measures of intelligence and socioeconomic statis.”

Previously, we used these data to model effects of education level and AFQT score on income in 2005. Now, I would like you to focus on whether we see an effect of race or gender on income levels, after controlling for education level. To accomplish this, we will combine race and gender variables in the income dataset into a single categorical “Group” variable (See R code below).

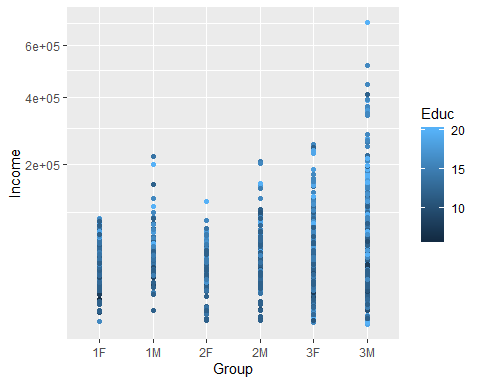
Inc$Income<-Inc$Income2005   
  
####### Just a few plots to see if data needed transformin  
(p1<-ggplot(Inc,aes(x=Group,y=Income,col=Educ,fill=Educ))+geom\_point())



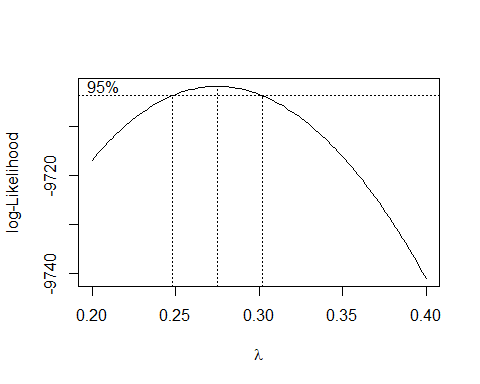
p1+coord\_trans(y="log")



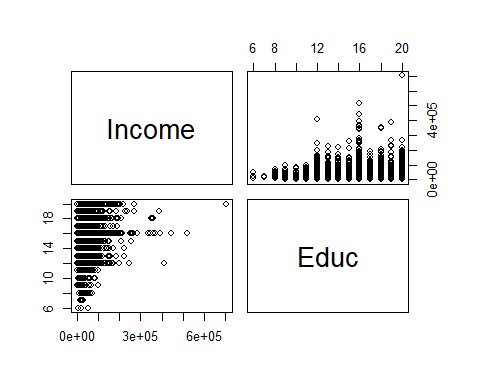
p1+coord\_trans(y='sqrt')



MASS::boxcox(Income~Group, data=Inc,lambda=seq(.2,.4,length.out = 100))



pairs(Inc[,c('Income','Educ')])



### Income def needs transforming, went with a log  
Inc$logInc<-log(Inc$Income)  
lm.Int1<-lm(logInc~Group\*Educ,data = Inc) ## interaction model  
summary(lm.Int1)

##   
## Call:  
## lm(formula = logInc ~ Group \* Educ, data = Inc)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.2040 -0.3299 0.1310 0.5139 2.4935   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.962516 0.468605 16.992 < 2e-16 \*\*\*  
## Group1M 0.946588 0.668322 1.416 0.15679   
## Group2F -0.179672 0.710482 -0.253 0.80038   
## Group2M -0.087785 0.677930 -0.129 0.89698   
## Group3F 0.693144 0.494830 1.401 0.16140   
## Group3M 1.346120 0.489927 2.748 0.00605 \*\*   
## Educ 0.159507 0.036152 4.412 1.07e-05 \*\*\*  
## Group1M:Educ -0.036167 0.050723 -0.713 0.47589   
## Group2F:Educ 0.007381 0.052884 0.140 0.88900   
## Group2M:Educ 0.027722 0.051529 0.538 0.59064   
## Group3F:Educ -0.054329 0.037820 -1.437 0.15098   
## Group3M:Educ -0.052216 0.037534 -1.391 0.16430   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8723 on 2572 degrees of freedom  
## Multiple R-squared: 0.2011, Adjusted R-squared: 0.1976   
## F-statistic: 58.84 on 11 and 2572 DF, p-value: < 2.2e-16

lm.add1<-lm(logInc~Group+Educ,data=Inc)### additive model  
summary(lm.add1)

##   
## Call:  
## lm(formula = logInc ~ Group + Educ, data = Inc)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.2368 -0.3399 0.1307 0.5050 2.5115   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.543776 0.128415 66.532 < 2e-16 \*\*\*  
## Group1M 0.491078 0.136148 3.607 0.000316 \*\*\*  
## Group2F -0.033472 0.122073 -0.274 0.783956   
## Group2M 0.297425 0.124790 2.383 0.017225 \*   
## Group3F -0.008937 0.097819 -0.091 0.927209   
## Group3M 0.674996 0.097540 6.920 5.67e-12 \*\*\*  
## Educ 0.113743 0.006923 16.430 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8729 on 2577 degrees of freedom  
## Multiple R-squared: 0.1984, Adjusted R-squared: 0.1965   
## F-statistic: 106.3 on 6 and 2577 DF, p-value: < 2.2e-16

confint(lm.add1)

## 2.5 % 97.5 %  
## (Intercept) 8.29196854 8.7955832  
## Group1M 0.22410786 0.7580489  
## Group2F -0.27284340 0.2058994  
## Group2M 0.05272669 0.5421238  
## Group3F -0.20074872 0.1828740  
## Group3M 0.48373077 0.8662602  
## Educ 0.10016793 0.1273180

anova(lm.Int1,lm.add1) ## model comparisons, going with additive since they're not sig diff

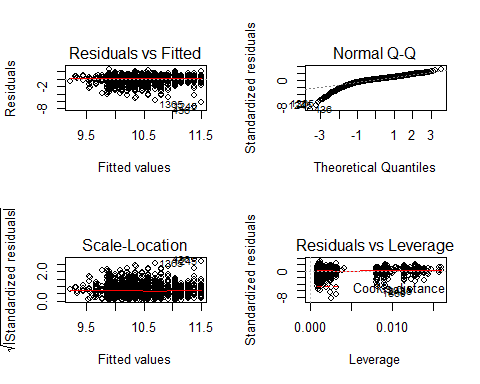
## Analysis of Variance Table  
##   
## Model 1: logInc ~ Group \* Educ  
## Model 2: logInc ~ Group + Educ  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 2572 1957.2   
## 2 2577 1963.8 -5 -6.5291 1.716 0.1275

lm.Male2<-lm(logInc~Educ+I(Group=="1M")+I(Group=="2M")+I(Group=="3M"),data=Inc)  
lm.Male<-lm(logInc~Educ+Group,data=subset(Inc,Group==c("1M","2M","3M")))

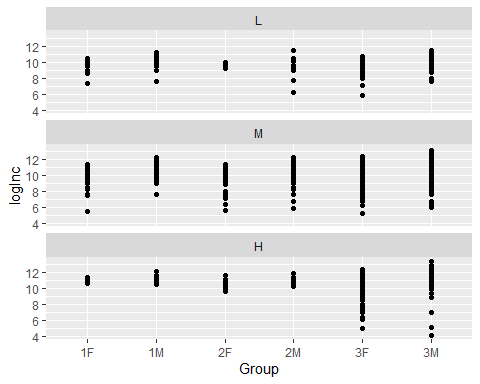
## Warning in `==.default`(Group, c("1M", "2M", "3M")): longer object length  
## is not a multiple of shorter object length

## Warning in is.na(e1) | is.na(e2): longer object length is not a multiple of  
## shorter object length

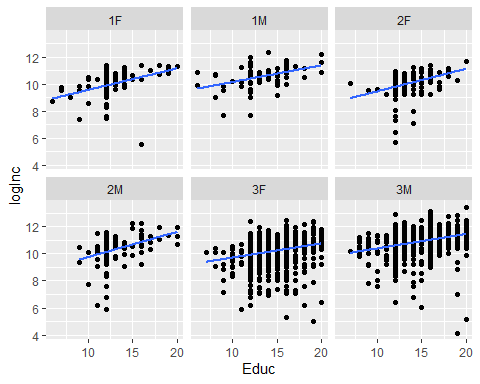
par(mfrow=c(2,2))  
#plot(lm.Int1)  
plot(lm.add1) ## diagnostics



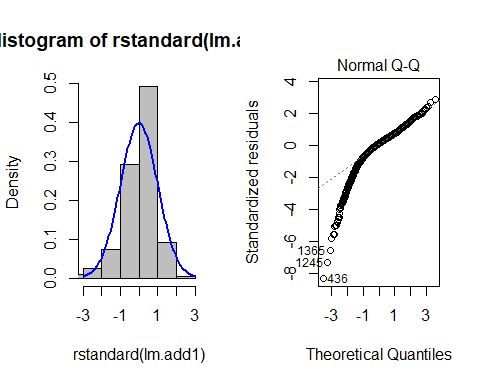
## helps visualize different incomes by group or education  
Inc$ED<-factor(ifelse(Inc$Educ < 12, 'L',  
 ifelse(Inc$Educ <= 16, 'M', 'H')), levels=c('L','M','H'))  
ggplot(Inc,aes(x=Group,y=logInc))+geom\_point()+facet\_wrap(.~ED,nrow=3)+stat\_smooth(method='lm',se=F)



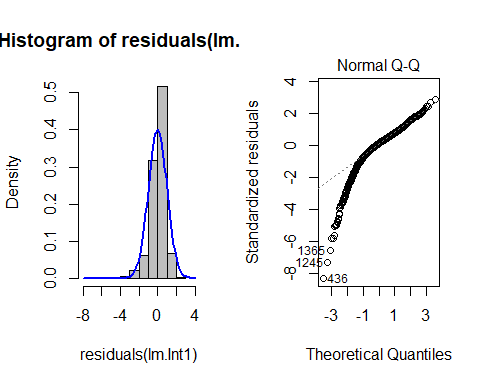
ggplot(Inc,aes(x=Educ,y=logInc))+geom\_point()+facet\_wrap(.~Group,nrow=3,ncol=3)+stat\_smooth(method='lm',se=F)



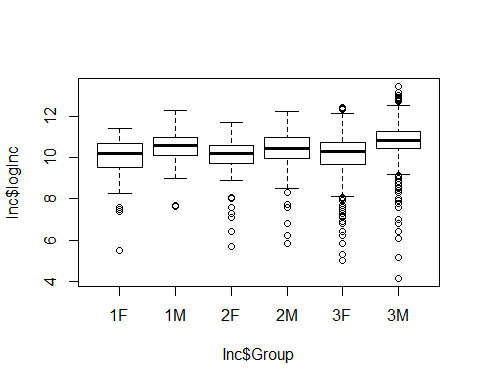
## nvmnd these   
par(mfrow=c(1,2))  
hist(rstandard(lm.add1), freq=F, col='grey', xlim=c(-3,3))  
curve(dnorm, xlim=c(-3,3), add=T, lwd=2, col='blue')  
plot(lm.add1, which=2)



par(mfrow=c(1,2))  
hist(residuals(lm.Int1), freq=F, col='grey', xlim=c(-8,4))  
curve(dnorm, xlim=c(-8,4), add=T, lwd=2, col='blue')  
plot(lm.Int1, which=2)



##  
  
### visualize mean income for different groups  
par(mfrow=c(1,1))  
boxplot(Inc$logInc~Inc$Group)



## income for groups after controlling for education, doesn't show much of a difference  
## think this means that after controlling for education the pattern between income and group  
## doesn't change much  
wch<-c(1,7)  
partial.res<-Inc$logInc-model.matrix(lm.add1)[,wch]%\*%coef(lm.add1)[wch]  
ggpubr::ggarrange(  
 ggplot(Inc,aes(x=Group,y=logInc,col=Educ,fill=Educ))+geom\_point(),  
   
ggplot(data.frame(Inc,partial.res),aes(x=Group,y=partial.res,))+stat\_smooth(method='loess', se=F, size=0.5, linetype='dashed')+  
 geom\_point(),ncol=2)

