The associations of income, gender, and life expectancies between countries.

Aadil Noon

Contents

```
#EXTRA FUNCTIONS
class_diag<-function(probs,truth){

  tab<-table(factor(probs>.5,levels=c("FALSE","TRUE")),truth)
  acc=sum(diag(tab))/sum(tab)
  sens=tab[2,2]/colSums(tab)[2]
  spec=tab[1,1]/colSums(tab)[1]
  ppv=tab[2,2]/rowSums(tab)[2]

  if(is.numeric(truth)==FALSE & is.logical(truth)==FALSE) truth<-as.numeric(truth)-1

  #CALCULATE EXACT AUC
  ord<-order(probs, decreasing=TRUE)
  probs <- probs[ord]; truth <- truth[ord]

  TPR=cumsum(truth)/max(1,sum(truth))
  FPR=cumsum(!truth)/max(1,sum(!truth))

  dup<-c(probs[-1]>=probs[-length(probs)], FALSE)
  TPR<-c(0,TPR[!dup],1); FPR<-c(0,FPR[!dup],1)

  n <- length(TPR)
  auc<- sum( ((TPR[-1]+TPR[-n])/2) * (FPR[-1]-FPR[-n]) )

  data.frame(acc,sens,spec,ppv,auc)
}
```

```
#libraries
library(ggplot2)
library(readxl)
library(dplyr)
library(cluster)
library(graphics)
library(plyr)
library(tidyverse)
library(lmtest)
library(sandwich)
#life expectancies data/wrangling
lifeexpec <- read_xlsx('Life expectancy (3).xlsx')
```

```
lifeexpec<-lifeexpec%>%pivot_wider(names_from = 'Year',values_from = 'Country')
lifeexpec<-lifeexpec%>%select(-c('2000':'2015'))
lifeexpec<-lifeexpec%>%na.omit()
lifeexpec<-lifeexpec%>%rename(c('2016'='Country'))
lifeexpec<-lifeexpec%>%select(-c(contains('birth')))
#sex ratio data/creating a binary
sexratio <- read_xlsx('sexratio.xlsx')
male<-sexratio%>%filter(`greater male sex ratio`>100)%>%mutate(`greater male sex ratio`=1)
female<-sexratio%>%filter(`greater male sex ratio`<100)%>%mutate(`greater male sex ratio`=0)
sexratio<-full_join(female,male)
#joining of life expectancies and sex ratios
lifeexpecsex<-inner_join(lifeexpec,sexratio,by="Country")
#income data
income<-read_xlsx('income.xlsx')
#final join
lifeexpecsexinc<-inner_join(lifeexpecsex,income,by="Country")
data<-lifeexpecsexinc
```

I combined data from the world health association which organized different kinds of life expectancies with data from the UN about gender ratios per country and data from the world bank on average incomes per country. My goal is to explore the assocations of a country's average income, its life expectancies, and its gender ratio.

- **1.

```
#MANOVA
man1<-manova(cbind(`Both Sexes' Life expectancy at age 60 in years`, `Male Life expectancy at age 60 in
summary(man1)
```

```
## Df Pillai approx F num Df den Df Pr(>F)
## IncomeGroup 3 0.76887 17.69 9 462 < 2.2e-16 ***
## Residuals 154
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

```
#significant results and therfore moved onto ANOVAs
summary.aov(man1)
```

```
## Response Both Sexes' Life expectancy at age 60 in years
:
## Df Sum Sq Mean Sq F value Pr(>F)
## IncomeGroup 3 874.31 291.436 78.481 < 2.2e-16 ***
## Residuals 154 571.87 3.713
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Response Male Life expectancy at age 60 in years :
## Df Sum Sq Mean Sq F value Pr(>F)
## IncomeGroup 3 618.87 206.290 52.872 < 2.2e-16 ***
## Residuals 154 600.86 3.902
```

```
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Response Female Life expectancy at age 60 in years :
## Df Sum Sq Mean Sq F value Pr(>F)
## IncomeGroup 3 1163.98 387.99 94.177 < 2.2e-16 ***
## Residuals 154 634.46 4.12
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

```
#All 3 variables were significant and therfore moved onto post-hoc t tests
pairwise.t.test(data$`Both Sexes' Life expectancy at age 60 in years`, data$IncomeGroup, p.adj="none")
```

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data: data$`Both Sexes' Life expectancy at age 60 in
years` and data$IncomeGroup
##
## High income Low income Lower middle income
## Low income < 2e-16 - -
## Lower middle income < 2e-16 0.011 -
## Upper middle income 1.4e-12 3.2e-11 2.1e-06
##
## P value adjustment method: none
```

```
pairwise.t.test(data$`Male Life expectancy at age 60 in years`, data$IncomeGroup, p.adj="none")
```

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data: data$`Male Life expectancy at age 60 in years` and
data$IncomeGroup
##
## High income Low income Lower middle income
## Low income < 2e-16 - -
## Lower middle income < 2e-16 0.08122 -
## Upper middle income 1.8e-11 2.3e-06 0.00084
##
## P value adjustment method: none
```

```
pairwise.t.test(data$`Female Life expectancy at age 60 in years`, data$IncomeGroup, p.adj="none")
```

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data: data$`Female Life expectancy at age 60 in years`
and data$IncomeGroup
##
## High income Low income Lower middle income
```

```
## Low income < 2e-16 - -
## Lower middle income < 2e-16 0.0021 -
## Upper middle income 3.1e-12 5.9e-15 1.6e-08
##
## P value adjustment method: none
```

```
#Performed 1 Manova, 3 Anova, and 12 post hoc t tests
0.05/12
```

```
## [1] 0.004166667
```

```
#to keep the chance of a Type one error at 0.05, the new alpha is 0.004166667
1-.95^12
```

```
## [1] 0.4596399
```

```
# if it was unjusted the chance of a Type one error is just above 0.45
```

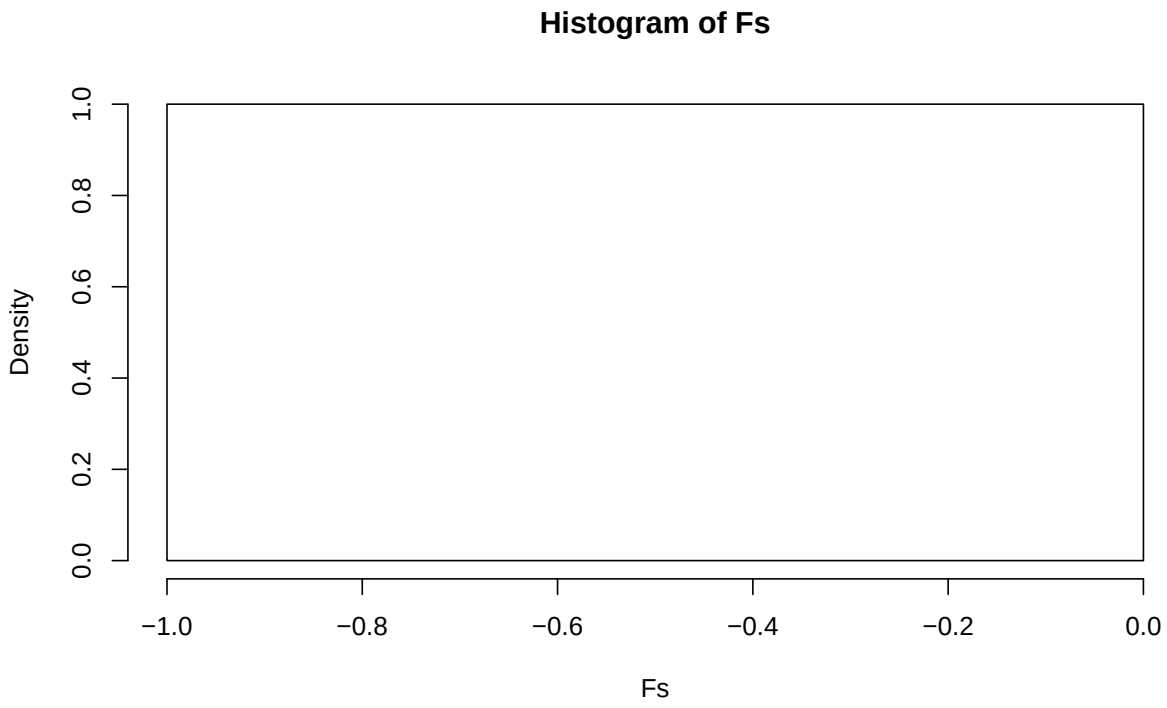
The initial MANOVA test showed that there was signicant mean difference of the three life expectancy variables across the four income groups. The ANOVA tests then showed each variable had a signifcant interaction with the variable of income group. The post hoc tests then showed that for overall life expectancy, male and female life expectancies that there were signifcant differences bewteen low and high income groups along with lower middle class and high income, and Upper Middle income and High income, Upper middle income and low income, and lower middle and upper middle incomes. This last result was much more pronounced in female life expectancy than male. The assumptions of the various test were likely met as the data consited of random samples and independent observations. Futher, the sample groups were large enough to avoid questions of normality and ther was likely equal variance between groups.

- **2.

```
obs_F<-78.481 #obsereved F for both sexes' life expectancy
Fs<-replicate(5000,{
new<-data%>%mutate(`Both Sexes' Life expectancy at age 60 in years`=sample(`Both Sexes' Life expectancy
SSW<- new%>%group_by(IncomeGroup)%>%summarize(SSW=sum((`Both Sexes' Life expectancy at age 60 in years`-
summarize(sum(SSW))%>%pull
SSB<- new%>%mutate(mean=mean(`Both Sexes' Life expectancy at age 60 in years`))%>%group_by(IncomeGroup)%
summarize(SSB=sum((mean-groupmean)^2))%>%summarize(sum(SSB))%>%pull
(SSB/3)/(SSW/154)
})
mean(Fs>obs_F)
```

```
## [1] 0
```

```
hist(Fs, prob=T); abline(v = obs_F, col="red",add=T)
```



The null hypothesis is that life expectancy of both sexes does not differ significantly between income groups. My alternative hypothesis is that life expectancy of both sexes does differ significantly between income groups. Because I got a p value of 0, I am very confident in rejecting the null hypothesis.

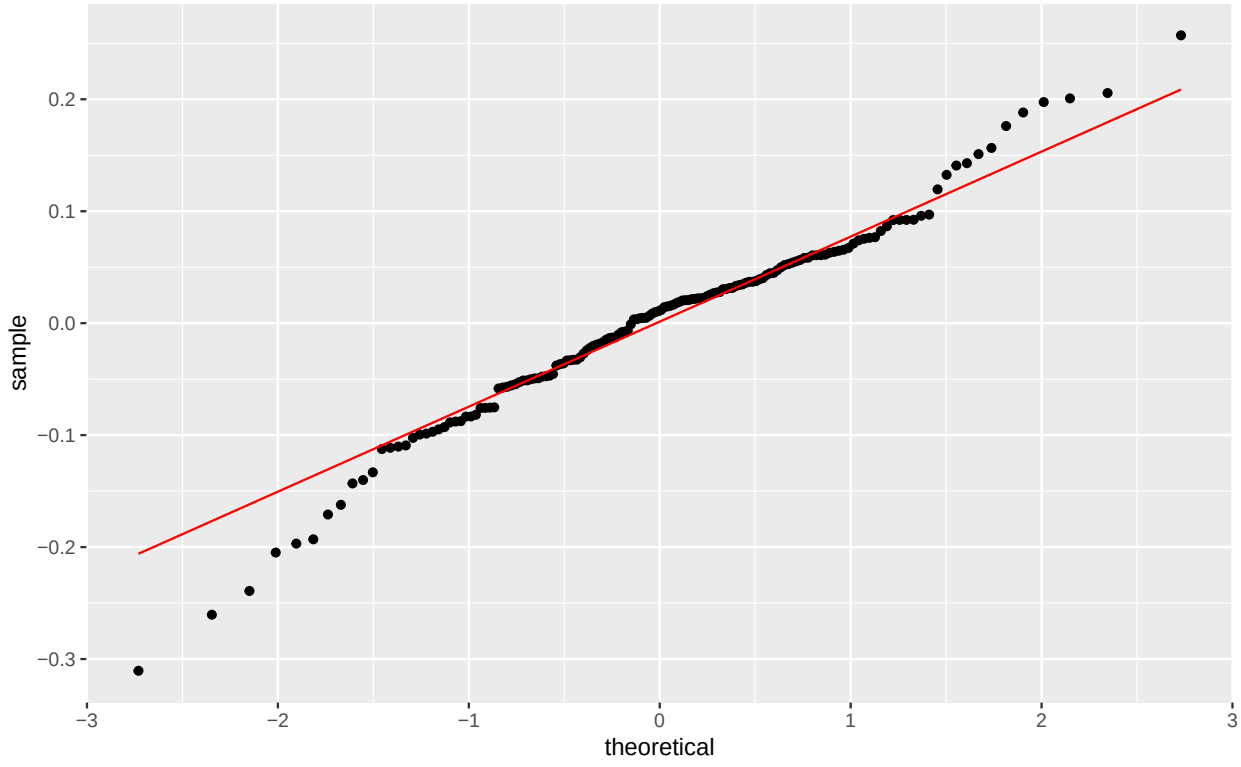
- **3.

```
#Centering
data$male_c<-data$`Male Life expectancy at age 60 in years` -mean(data$`Male Life expectancy at age 60 i
data$female_c<-data$`Female Life expectancy at age 60 in years` -mean(data$`Female Life expectancy at ag
data$both_c<-data$`Both Sexes' Life expectancy at age 60 in years` -mean(data$`Both Sexes' Life expectan
#regression
fit<-lm(both_c~male_c*female_c,data=data)
summary(fit)
```

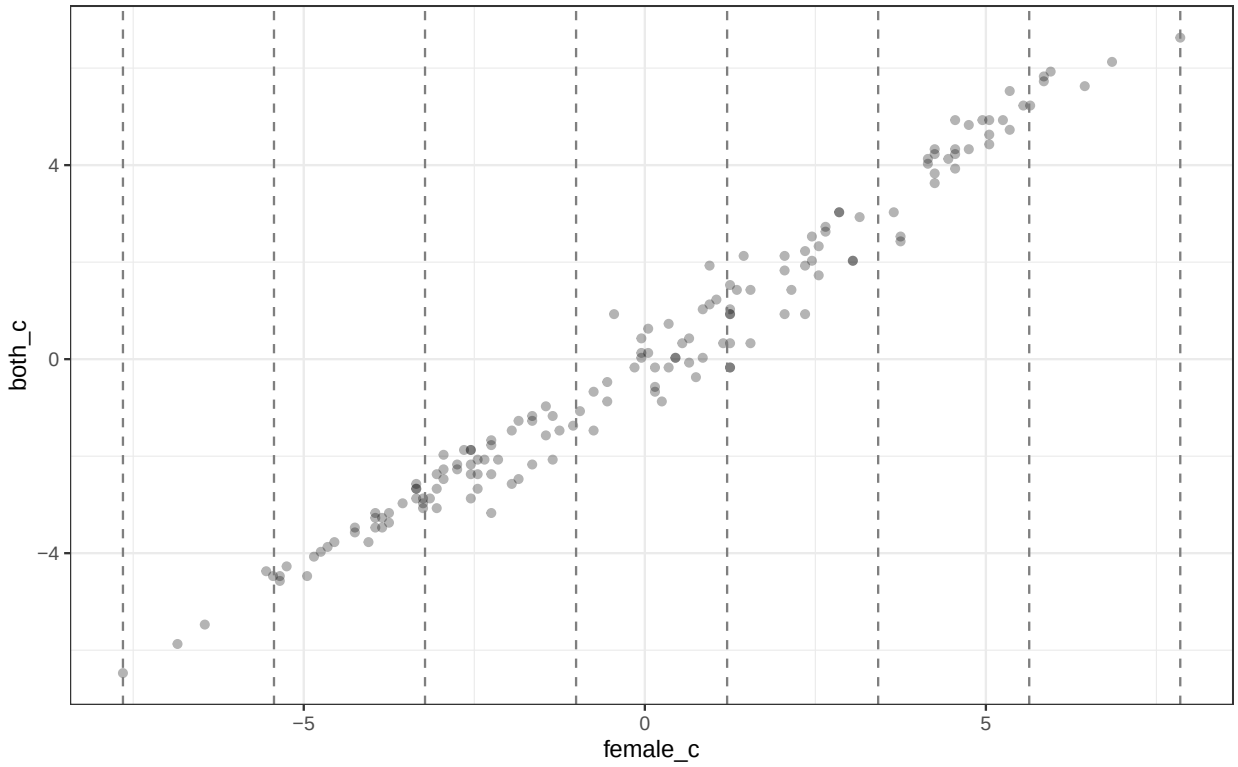
```
##
## Call:
## lm(formula = both_c ~ male_c * female_c, data = data)
##
## Residuals:
## Min 1Q Median 3Q Max
## -0.31046 -0.04984 0.01123 0.05260 0.25711
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.015837 0.010008 -1.582 0.1156
## male_c 0.427722 0.007076 60.447 <2e-16 ***
## female_c 0.559999 0.005597 100.055 <2e-16 ***
```

```
## male_c:female_c 0.001836 0.000813 2.259 0.0253 *
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 0.08976 on 154 degrees of
freedom
## Multiple R-squared: 0.9991, Adjusted R-squared: 0.9991
## F-statistic: 5.978e+04 on 3 and 154 DF, p-value: <
2.2e-16
```

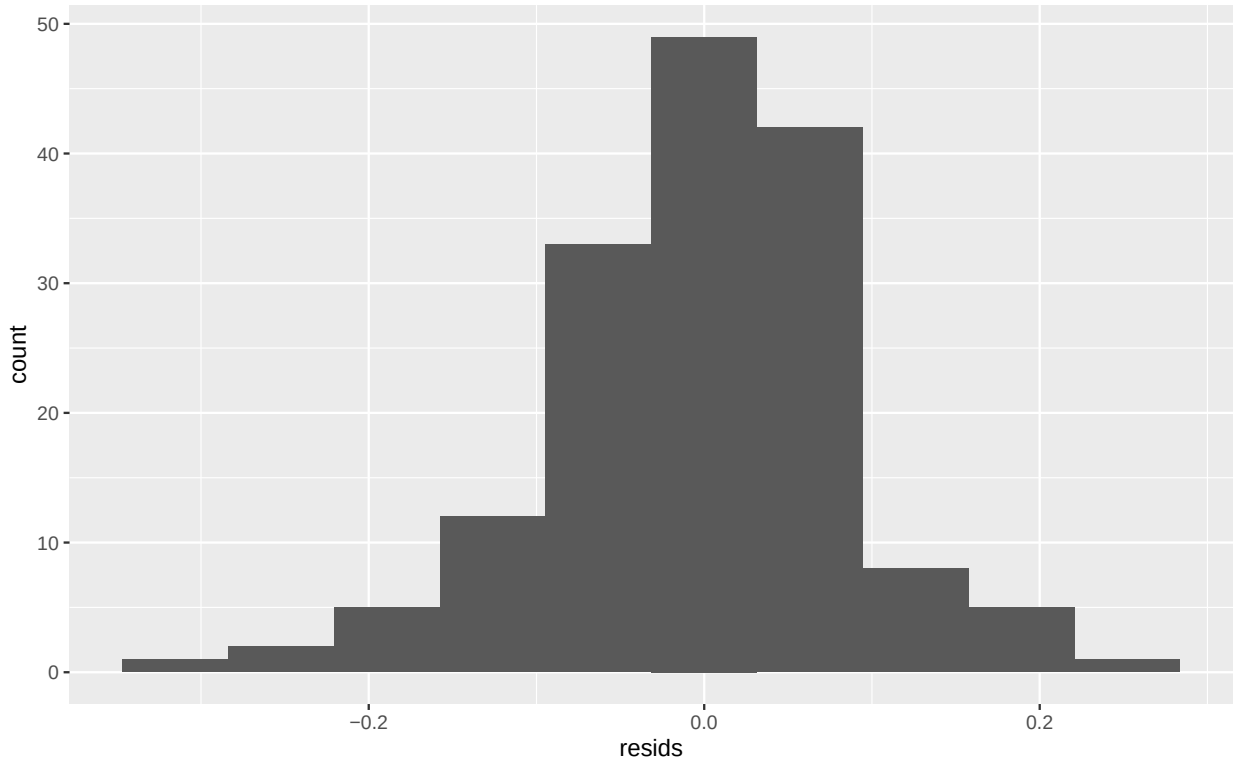
```
#graph
resids<-fit$residuals; fitvals<-fit$fitted.values
ggplot()+geom_qq(aes(sample=resids))+geom_qq_line(aes(sample=resids), color='red')
```



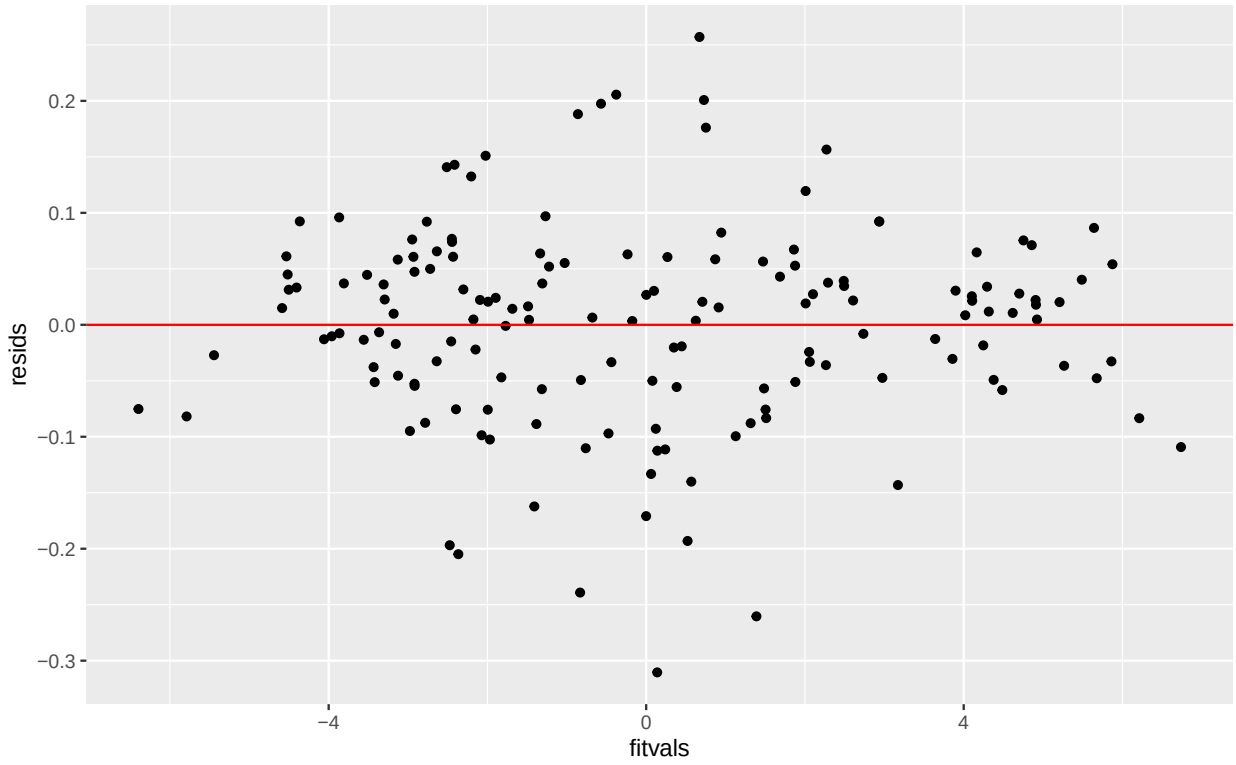

```
#assumptions
#linear
breaks <- seq(min(data$female_c), max(data$female_c), len=8)
ggplot(data, aes(female_c,both_c)) +
geom_point(alpha=.3) +
theme_bw()+
geom_vline(xintercept=breaks, lty=2,color='gray50')
```




```
#normality
ggplot()+geom_histogram(aes(resids),bins=10)
```



```r
#homoskedasticity
resids<-fit$residuals; fitvals<-fit$fitted.values
ggplot()+geom_point(aes(fitvals,resids))+geom_hline(yintercept=0, col="red")
```




```r
#Robust SE
coeftest(fit, vcov = vcovHC(fit))[,1:2]
```

```
##                    Estimate   Std. Error
## (Intercept)     -0.015837289 0.0115602418
## male_c           0.427722224 0.0111152872
## female_c         0.559999119 0.0087335888
## male_c:female_c  0.001836149 0.0008136394
```

```r
#proportion explained
(sum((data$both_c-mean(data$both_c))^2)-sum(fit$residuals^2))/sum((data$both_c-mean(data$both_c))^2)
```

```
## [1] 0.999142
```

Based on the coefficents, there is a slight interaction between life expectancy between the sexes and much larger interactions with the each of them and the overall life expectancy.More robust standard errors did not impact the results, indicating a strong model. My model accounts for nearly all off the variation (99.9142%).

- **4.

```
samp_distn<-replicate(5000, {
boot_dat <- sample_frac(data, replace=T)
fit <- lm(both_c~male_c*female_c, data=boot_dat)
coef(fit)
})
summary(fit)
```

```
##
## Call:
## lm(formula = both_c ~ male_c * female_c, data = data)
##
## Residuals:
## Min 1Q Median 3Q Max
## -0.31046 -0.04984 0.01123 0.05260 0.25711
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.015837 0.010008 -1.582 0.1156
## male_c 0.427722 0.007076 60.447 <2e-16 ***
## female_c 0.559999 0.005597 100.055 <2e-16 ***
## male_c:female_c 0.001836 0.000813 2.259 0.0253 *
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 0.08976 on 154 degrees of
freedom
## Multiple R-squared: 0.9991, Adjusted R-squared: 0.9991
## F-statistic: 5.978e+04 on 3 and 154 DF, p-value: <
2.2e-16
```

Bootstrapped SEs produced the same cooefficets and p-values along with SE values.

- **5. (40 pts)** Perform a logistic regression predicting a binary categorical variable (if you don't have one, make/get one) from at least two explanatory variables (interaction not necessary).

```
data3<-data[,-c(4,7,8,9)]
fit<-glm(data3$'greater male sex ratio'~`Male Life expectancy at age 60 in years`+`Female Life expectanc
coeftest(fit)
```

```
##
## z test of coefficients:
##
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.094091 1.145946 0.0821 0.934561
## `Male Life expectancy at age 60 in years` 0.630912
0.193841 3.2548 0.001135 **
## `Female Life expectancy at age 60 in years` -0.583529
0.159646 -3.6551 0.000257 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

```r
#confusion matrix
probs<-predict(fit,type="response")
table(predict=as.numeric(probs>.5),truth=data$'greater male sex ratio')%>%addmargins
```

```
##        truth
## predict   0   1 Sum
##     0    81  40 121
##     1    18  19  37
##     Sum  99  59 158
```

```r
#Sensitivty=
19/59
```

```
## [1] 0.3220339
```

```r
#Specficity=
81/99
```

```
## [1] 0.8181818
```

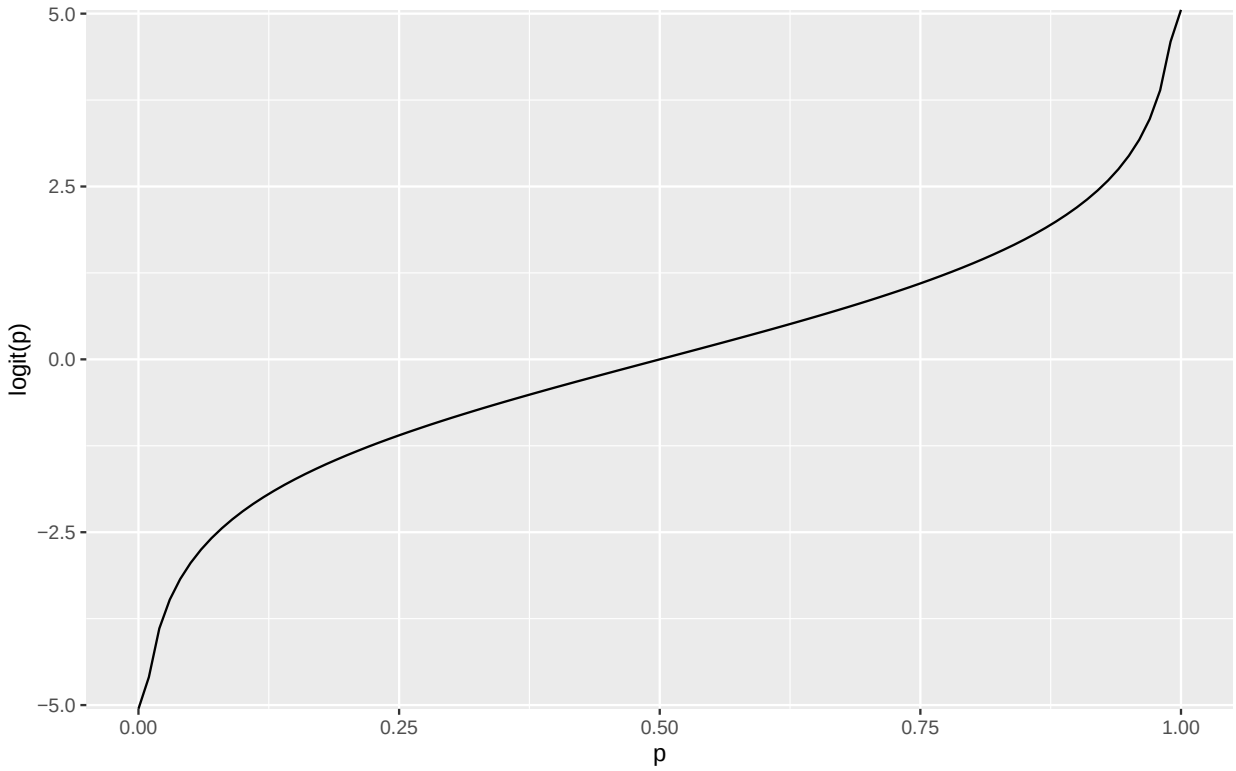
```r
#Precision=
19/37
```

```
## [1] 0.5135135
```

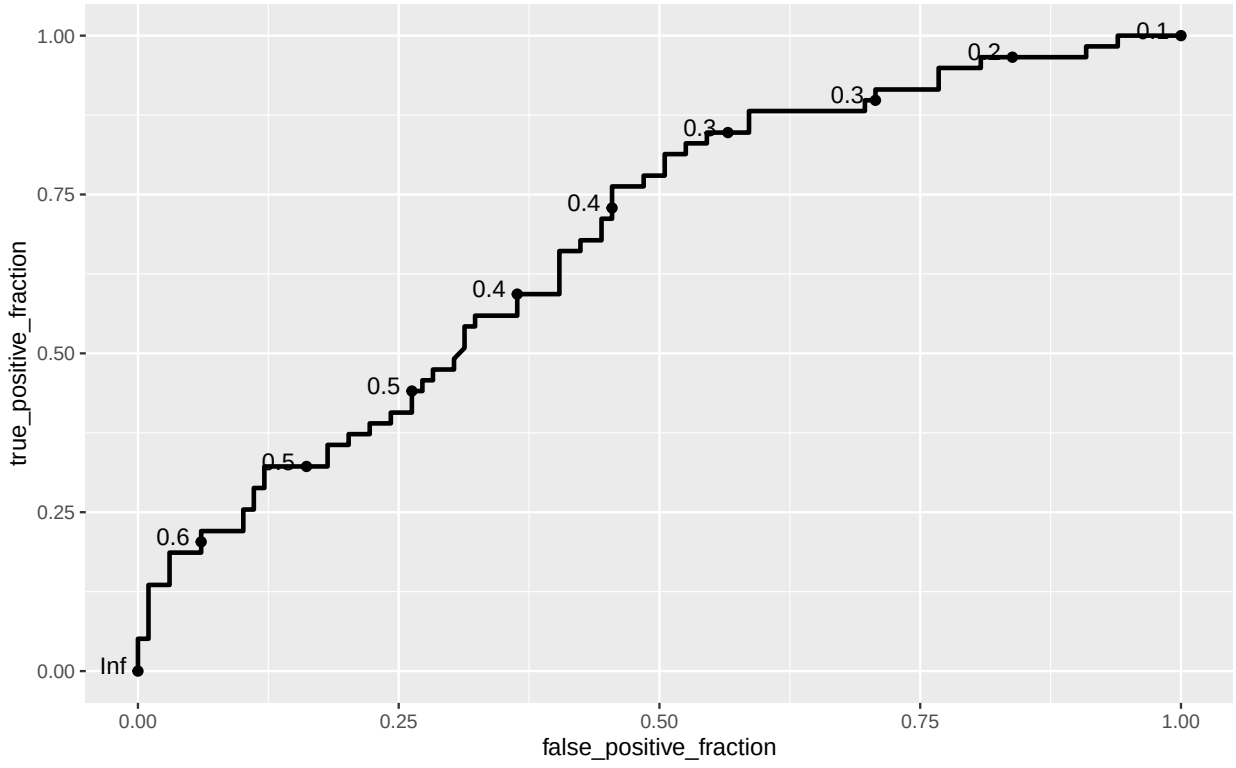
```r
#accuracy=
(81+19)/158
```

```
## [1] 0.6329114
```

```r
#logit
odds<-function(p)p/(1-p)
logit<-function(p)log(odds(p))
p<-data$'greater male sex ratio'
ggplot()+ stat_function(aes(p),fun=logit,geom="line")+ ylab("logit(p)")+xlab("p")
```




```
#ROC curve and AUC
roc1<-ggplot()+geom_roc(aes(d=p,m=probs))
roc1
```



```
#AUC=
calc_auc(roc1)
```

```
##   PANEL group       AUC
## 1     1    -1 0.6787365
```

```
#10-fold
set.seed(1234)
k=10
data1<-data3%>%sample_frac()
folds <- ntile(1:nrow(data1),n=10)

#diags<-NULL
#for(i in 1:k){
#train<-data1[folds!=i,]
#test<-data1[folds==i,]

#truth<-test$'greater male sex ratio'

#fit <- glm('greater male sex ratio'~'Male Life expectancy at age 60 in years'+'Female Life expectancy

#probs<- predict(fit, newdata=test, type="response")

#diags<-rbind(diags,class_diag(probs,truth))
#}
#mean(diags$auc)
#summary(diags)
```

There is a significant positive association of Male life expectancy with a country with more men than women. There is also a significant negative association of Female life expectancy and a larger male sex ratio in a country.My model is not very sensitive but is relatively Specific and is medicoraly precise and accurate.

```
- Perform 10-fold (or repeated random sub-sampling) CV and report average out-of-sample Accuracy, Sensi
```

- **6. Perform 10-fold CV using this model: if response in binary, compare model's out-of-sample accuracy to that of your logistic regression in part 5; if response is numeric, compare the residual standard error (at the bottom of the summary output, aka RMSE): lower is better fit!

```
library(glmnet)
fit<-lm(`Both Sexes' Life expectancy at age 60 in years`~.,data=data3)
set.seed(1234)
predic<-model.matrix(fit)
x<-predic[,-1]
y<-as.matrix(data$`Both Sexes' Life expectancy at age 60 in years`)
x<-scale(x)
cv<-cv.glmnet(x,y)
lasso<-glmnet(x,y,lambda=cv$lambda.1se)
coef(lasso)
```

```
## 7 x 1 sparse Matrix of class "dgCMatrix"
##                                                      s0
```

```
## (Intercept)                                  19.771519
## `Male Life expectancy at age 60 in years`    1.184498
## `Female Life expectancy at age 60 in years`  1.851758
## `greater male sex ratio`                     .
## IncomeGroupLow income                        .
## IncomeGroupLower middle income               .
## IncomeGroupUpper middle income               .
```

```
#10-fold
set.seed(1234)
k=10
data1<-data[sample(nrow(data)),]
folds<-cut(seq(1:nrow(data)),breaks=10,labels=F)

fit<-glm(`Both Sexes' Life expectancy at age 60 in years`~`Female Life expectancy at age 60 in years`+`

diags<-NULL
for(i in 1:k){
train<-data1[folds!=i,]
test<-data1[folds==i,]

truth<-test$'greater male sex ratio'

probs<- predict(fit, newdata=test, type="response")

diags<-rbind(diags,class_diag(probs,truth))
}
RSS <- c(crossprod(fit$residuals))
MSE <- RSS / length(fit$residuals)
#RMSE=
sqrt(MSE)
```

```
## [1] 0.09007567
```

The variables that the LASSO left in were female and male life expectancies.The model has a low RMSE at 0.09007567. ...