Take Home Final (S-520)

FNU Anirudh

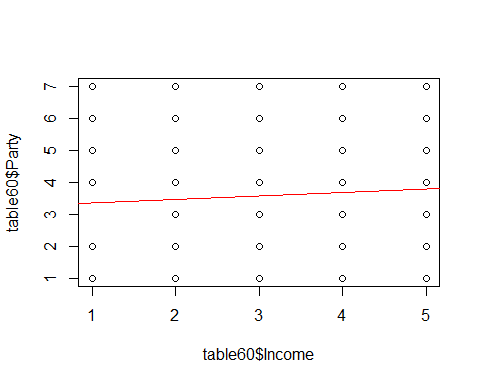
December 12, 2015

# Solution 1

table60<- read.table('C:/Stats/TAKEHOME FINAL/election1960.txt',header=TRUE)  
fit1<- lm(Party~Income,data = table60)  
summary(fit1)

##   
## Call:  
## lm(formula = Party ~ Income, data = table60)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.7865 -1.6824 -0.6824 2.3176 3.6298   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.26610 0.20787 15.71 <2e-16 \*\*\*  
## Income 0.10408 0.06386 1.63 0.103   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.263 on 1001 degrees of freedom  
## Multiple R-squared: 0.002647, Adjusted R-squared: 0.00165   
## F-statistic: 2.656 on 1 and 1001 DF, p-value: 0.1035

plot(table60$Income,table60$Party)  
abline(fit1,col="red")



slope = cor(table60$Income,table60$Party) \* sd(table60$Party) / sd(table60$Income)   
intercept = mean(table60$Party) - slope \* mean(table60$Income)

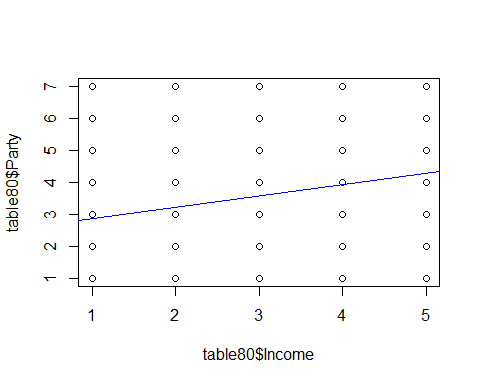
The equation of regression line is y=slope1\* x + Intercept1 i.e. Party=0.104\* Income + 3.266

# Solution 2

table80<- read.table('C:/Stats/TAKEHOME FINAL/election1980.txt',header=TRUE)  
fit2<- lm(Party~Income,data = table80)  
summary(fit2)

##   
## Call:  
## lm(formula = Party ~ Income, data = table80)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.291 -1.881 -0.586 2.061 4.119   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.52803 0.16971 14.896 < 2e-16 \*\*\*  
## Income 0.35265 0.05479 6.436 1.85e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.037 on 1058 degrees of freedom  
## Multiple R-squared: 0.03768, Adjusted R-squared: 0.03677   
## F-statistic: 41.43 on 1 and 1058 DF, p-value: 1.852e-10

slope2 = cor(table80$Income,table80$Party) \* sd(table80$Party) / sd(table80$Income)   
intercept2 = mean(table80$Party) - slope2 \* mean(table80$Income)  
plot(table80$Income,table80$Party)  
abline(fit2,col="blue")



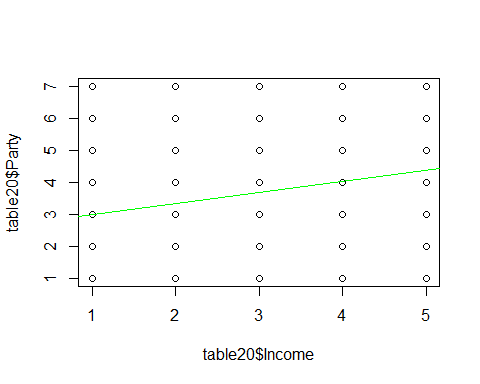
The equation of regression line is y=slope2\* x + Intercept2 i.e. Party=0.352\* Income + 2.528

# Solution 3

table20<- read.table('C:/Stats/TAKEHOME FINAL/election2000.txt',header=TRUE)  
fit3<- lm(Party~Income,data = table20)  
summary(fit3)

##   
## Call:  
## lm(formula = Party ~ Income, data = table20)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.3710 -1.9999 -0.3426 1.9718 4.0001   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.65710 0.16983 15.646 < 2e-16 \*\*\*  
## Income 0.34277 0.05495 6.238 6.17e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.134 on 1174 degrees of freedom  
## Multiple R-squared: 0.03208, Adjusted R-squared: 0.03126   
## F-statistic: 38.91 on 1 and 1174 DF, p-value: 6.17e-10

slope3 = cor(table20$Income,table20$Party) \* sd(table20$Party) / sd(table20$Income)   
intercept3 = mean(table20$Party) - slope3 \* mean(table20$Income)  
plot(table20$Income,table20$Party)  
abline(fit3,col="green")

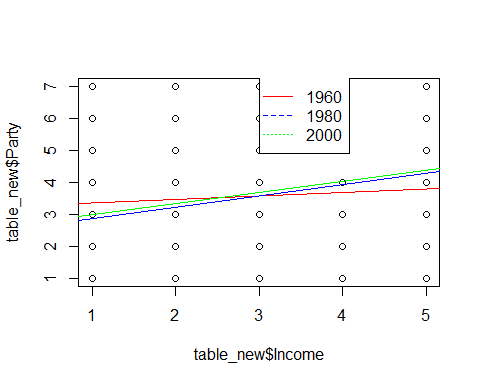


The equation of regression line is y=slope3\* x + Intercept3

i.e. Party=0.342\* Income + 2.657

# Solution 4

table60$Year<-rep("1960",nrow(table60))  
table80$Year<-rep("1980",nrow(table80))  
table20$Year<-rep("2000",nrow(table20))  
table\_new<-rbind(table60,table80,table20)  
plot(table\_new$Income,table\_new$Party)  
abline(fit1,col="red")  
abline(fit2,col="blue")  
abline(fit3,col="green")  
leg.line<- c("1960","1980","2000")  
legend(list(x=3,y=7.3),legend=leg.line,col=c("red","blue","green"),lty=c(1,2,3),merge=TRUE)



# Solution 5

* In 1960, Most of the High Income people were in between Independent Democrat and true independent, even low class preferred the same.
* In 2000, Low Income people are either True independent or weak democrat whereas High Income people are Independent democrat or true Independent.

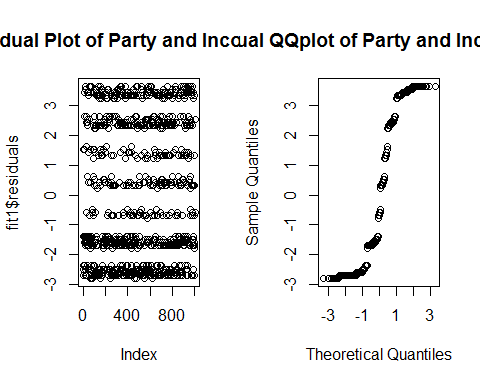
To summarize, People in 2000 are more inclined towards republicans compared to 1960's or earlier.

# Solution 6

summary(fit1)$coeff

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 3.2660965 0.20787168 15.712080 6.859904e-50  
## Income 0.1040787 0.06385869 1.629828 1.034525e-01

par(mfrow=c(1,2))   
plot(fit1$residuals,main="Residual Plot of Party and Income 1960")  
qqnorm(fit1$residuals,main="Residual QQplot of Party and Income 1960")



Slope=0.104 has confidence interval (-0.02,0.23)

We should not take it's meaning literally because:-

* As seen from Residual Plots there is no linearity between Party and Income. Linear Regression depends on assumptions like Linearity, Homoscedascity, Independence and Normality of errors.
* Linear Regression is not a good fit for categorical data,Both Party and Income are categorical data which should be fit using different model.

# Solution 7

table60<- read.table('C:/Stats/TAKEHOME FINAL/election1960.txt',header=TRUE)  
table80<- read.table('C:/Stats/TAKEHOME FINAL/election1980.txt',header=TRUE)  
table20<- read.table('C:/Stats/TAKEHOME FINAL/election2000.txt',header=TRUE)  
# Party Total in 1960  
p160<- sum((table60$Party==1))  
p260<- sum((table60$Party==2))  
p360<- sum((table60$Party==3))  
p460<- sum((table60$Party==4))  
p560<- sum((table60$Party==5))  
p660<- sum((table60$Party==6))  
p760<- sum((table60$Party==7))  
  
# Party Total in 1980  
  
p180<- sum((table80$Party==1))  
p280<- sum((table80$Party==2))  
p380<- sum((table80$Party==3))  
p480<- sum((table80$Party==4))  
p580<- sum((table80$Party==5))  
p680<- sum((table80$Party==6))  
p780<- sum((table80$Party==7))  
  
# Party Total in 2000  
  
p120<- sum((table20$Party==1))  
p220<- sum((table20$Party==2))  
p320<- sum((table20$Party==3))  
p420<- sum((table20$Party==4))  
p520<- sum((table20$Party==5))  
p620<- sum((table20$Party==6))  
p720<- sum((table20$Party==7))  
  
observed<- matrix(c(p160,p260,p360,p460,p560,p660,p760,p180,p280,p380  
 ,p480,p580,p680,p780,p120,p220,p320,p420,p520,p620  
 ,p720),ncol=3,nrow = 7)  
colnames(observed)<- c("1960","1980","2000")  
rownames(observed)<- c("1","2","3","4","5","6","7")  
tt=as.table(observed,header=TRUE)  
# Put my results in table  
observed=read.table("C:/Stats/Assignment 11/Observed.txt")  
observed

## X1960 X1980 X2000  
## 1 242 217 270  
## 2 241 246 203  
## 3 56 106 156  
## 4 89 103 85  
## 5 65 122 143  
## 6 148 155 148  
## 7 162 111 171

p60=sum(observed$X1960)/sum(observed)  
p80=sum(observed$X1980)/sum(observed)  
p20=sum(observed$X2000)/sum(observed)  
p1= sum(observed[1,])  
p2= sum(observed[2,])  
p3= sum(observed[3,])  
p4= sum(observed[4,])  
p5= sum(observed[5,])  
p6= sum(observed[6,])  
p7= sum(observed[7,])  
p11= p1\*p60  
p21= p2\*p60  
p31= p3\*p60  
p41= p4\*p60  
p51= p5\*p60  
p61= p6\*p60  
p71= p7\*p60  
#  
p12= p1\*p80  
p22= p2\*p80  
p32= p3\*p80  
p42= p4\*p80  
p52= p5\*p80  
p62= p6\*p80  
p72= p7\*p80  
#  
p13=p1\*p20  
p23=p2\*p20  
p33=p3\*p20  
p43=p4\*p20  
p53=p5\*p20  
p63=p6\*p20  
p73=p7\*p20  
expected<- matrix(c(p11,p21,p31,p41,p51,p61,p71,p12,p22,p32,p42,p52,p62  
 ,p72,p13,p23,p33,p43,p53,p63,p73),ncol=3,nrow=7)  
expected

## [,1] [,2] [,3]  
## [1,] 225.74467 238.57363 264.6817  
## [2,] 213.66780 225.81044 250.5218  
## [3,] 98.47299 104.06916 115.4579  
## [4,] 85.77678 90.65144 100.5718  
## [5,] 102.18895 107.99630 119.8148  
## [6,] 139.65823 147.59494 163.7468  
## [7,] 137.49058 145.30411 161.2053

df=6\*2  
X2 = sum((observed - expected)^2 / expected)  
1 - pchisq(X2, df=df)

## [1] 5.817569e-14

Since p-value is so low i.e. p-value<0.05, we reject our Null Hypothesis i.e Distribution of Party is independent of year and we conclude that Distribution of Party is dependent on year

# Solution 8

H0:- Population average age is same between 3 years (Null Hypothesis)

H1:- Population average age is not same between 3 years (Alternate Hypothesis)

n1=nrow(table60)  
n2=nrow(table80)  
n3=nrow(table20)  
N=n1+n2+n3  
meana=mean(table60$Age)  
meanb=mean(table80$Age)  
meanc=mean(table20$Age)  
grandmean=mean(table\_new$Age)  
SSB = n1\*(meana-grandmean)^2 + n2\*(meanb-grandmean)^2 + n3\*(meanc-grandmean)^2   
between.df = 2   
between.meansquare = SSB/2  
SSW = sum( (table60$Age-meana)^2 ) + sum( (table80$Age-meanb)^2 ) + sum( (table20$Age-meanc)^2 )  
within.df = N - 3   
within.meansquare = SSW/within.df  
# F-test   
F = between.meansquare/within.meansquare  
P=1 - pf(F, df1=between.df, df2=within.df)  
P

## [1] 0.003352559

fit<- lm(table\_new$Age~table\_new$Year)  
  
# Alternatively to check our solution, we can use ANOVA  
anova(fit)

## Analysis of Variance Table  
##   
## Response: table\_new$Age  
## Df Sum Sq Mean Sq F value Pr(>F)   
## table\_new$Year 2 3033 1516.32 5.7081 0.003353 \*\*  
## Residuals 3236 859625 265.64   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

p-value=0.003353<0.05 for 95% Confidence Interval hence we reject our Null Hypothesis i.e. Population average age is same for 3 years. We conclude that Population average age differs between 3 years.

# Solution 9

H0:- Population proportion of women remained same in 2000. (Null Hypothesis)

H1:- Population proportion of women were not same in 2000 (Alternate Hypothesis)

female60<- nrow(table60[table60$Sex==2,])  
femaleratio60<- female60/nrow(table60)  
maleratio60<- 1-femaleratio60  
  
female20<-nrow(table20[table20$Sex==2,])  
male20<-nrow(table20[table20$Sex==1,])  
observed<- c(female20,male20)  
expfem20<- femaleratio60 \*nrow(table20)  
expmal20<- maleratio60 \*nrow(table20)  
expected<-c(expfem20,expmal20)  
  
# Pearson's chi-squared  
X2 = sum((observed - expected)^2 / expected)  
1 - pchisq(X2, df=2-0-1)

## [1] 0.07546333

# Cross Check values with LR chi-squared test  
G2 = 2 \* sum(observed \* log(observed/expected))  
1 - pchisq(G2, df=2-0-1)

## [1] 0.07516072

p-value comes to 0.075 using both Pearson's or Chi-Squared test hence we cannot reject our Null Hypothesis since 0.075 > 0.05 for 95% Confidence Interval

Hence we conclude that Population proportion of women is same in 2000 as it was in 1960.