HW6

Shaojie Zhang (shaojiez)

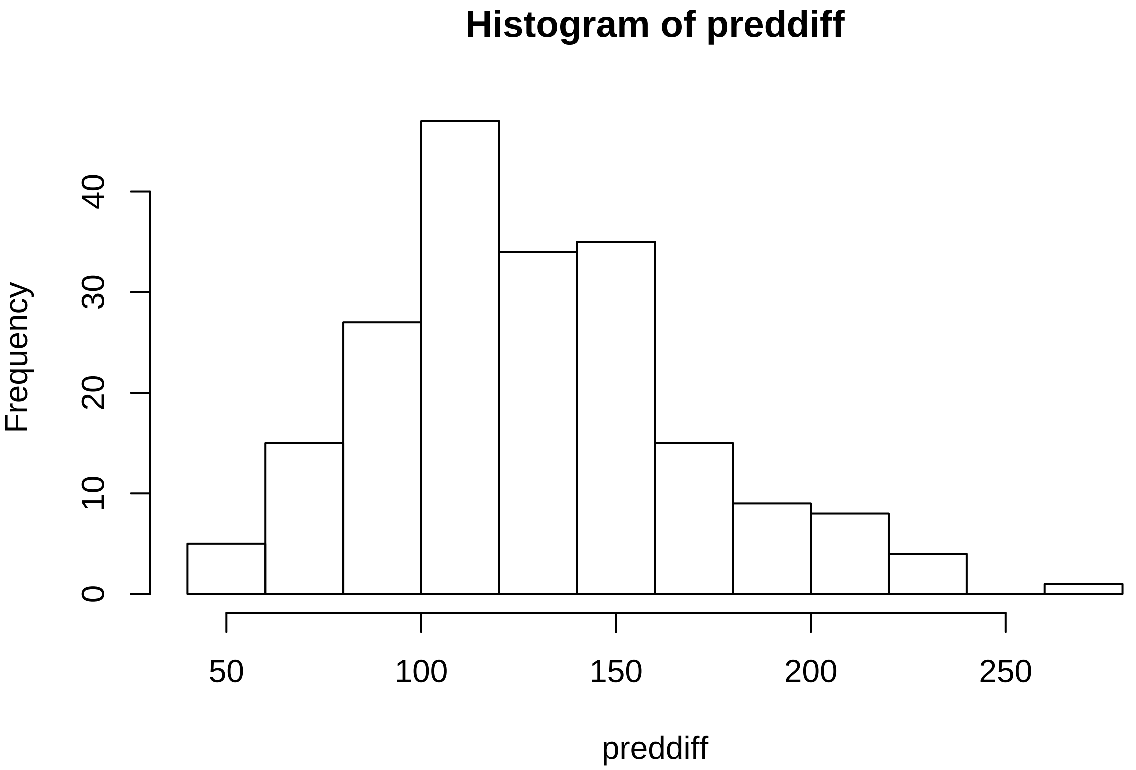
02/28/2019

Problem 1

```
housetrain=read.csv("housetrain.csv",header=T)
housetest=read.csv("housetest.csv",header=T)
housedata=rbind(housetrain,housetest)
```

a

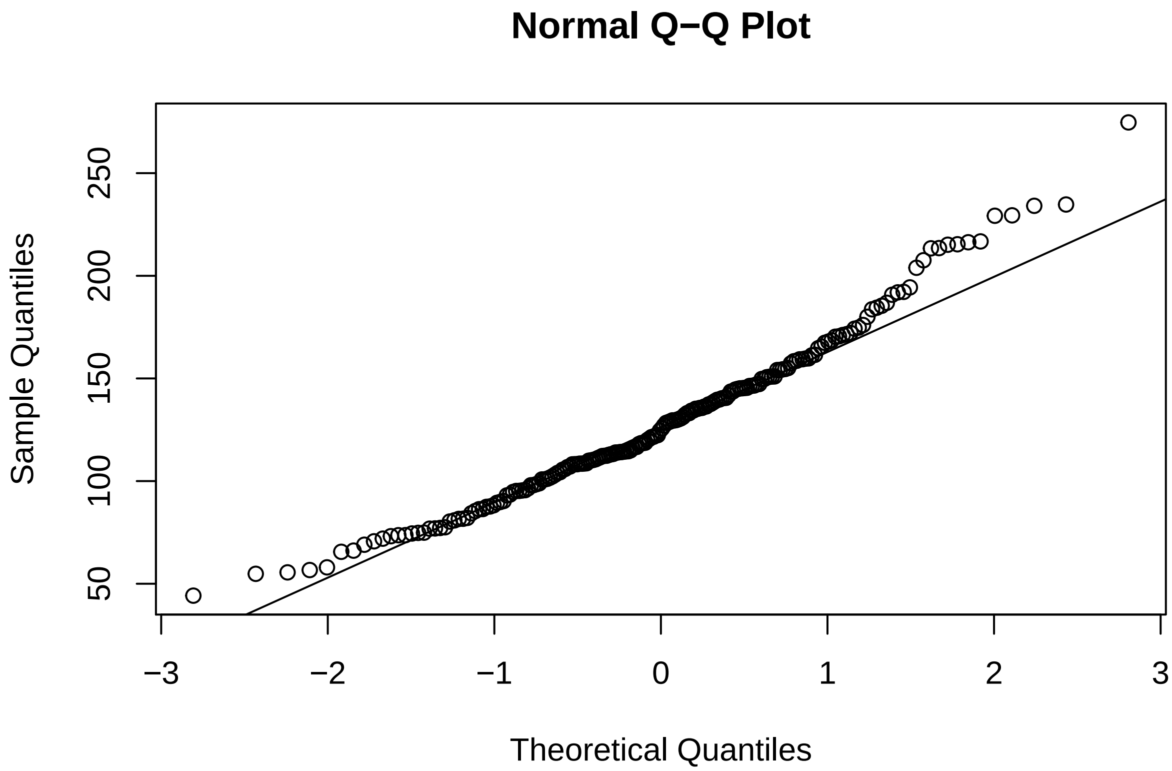
```
B=200
original=rep(1:5,10605/5)
preddiff=NULL

for(b in 1:B){
  new=sample(original)
  boot=sample(10605,10605,replace=T)
  temp=housedata[boot,]
  prederr2=NULL
  prederr3=NULL

  for(j in 1:5){
   traind=temp[new!=j,]
   testd=temp[new==j,]
   model2=lm(Median_house_value ~ Mean_household_income,data=traind)
   prederr2[j]=mean((predict(model2,newdata=testd) - testd$Median_house_value)^2)
   model3=lm(Median_house_value ~ Mean_household_income + Median_household_income,data=traind)
   prederr3[j]=mean((predict(model3,newdata=testd)-testd$Median_house_value)^2)
  }

  preddiff[b]=mean(prederr2-prederr3)
}

hist(preddiff)
```



Since the prediction differences are all positive, so that we conclude model3 is better than model2.

b

```
qqnorm(preddiff)
qqline(preddiff)
```



```r
t.test(preddiff)
```

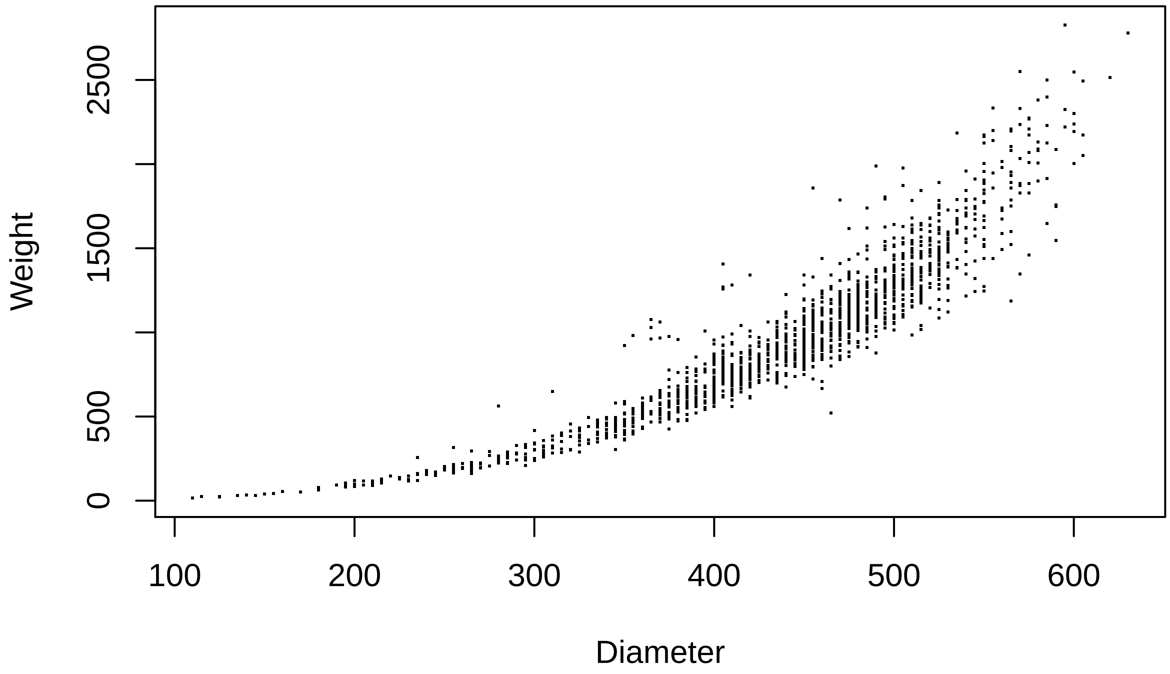
```
##
##  One Sample t-test
##
## data:  preddiff
## t = 45.04, df = 199, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  123.5360 134.8487
## sample estimates:
## mean of x
##  129.1923
```

From the t test, we should reject the null hypothesis. The qq plot is straight so model3 is better.

Problem2

a

```r
abalone=read.csv("fishdata.csv",header=T)
plot(abalone$Diameter,abalone$Weight,pch=".",xlab="Diameter",ylab="Weight")
```



Since the trend of the curve is very obvious so that a linear regression will not serve as a good fit here.

b

```r
library(np)
```

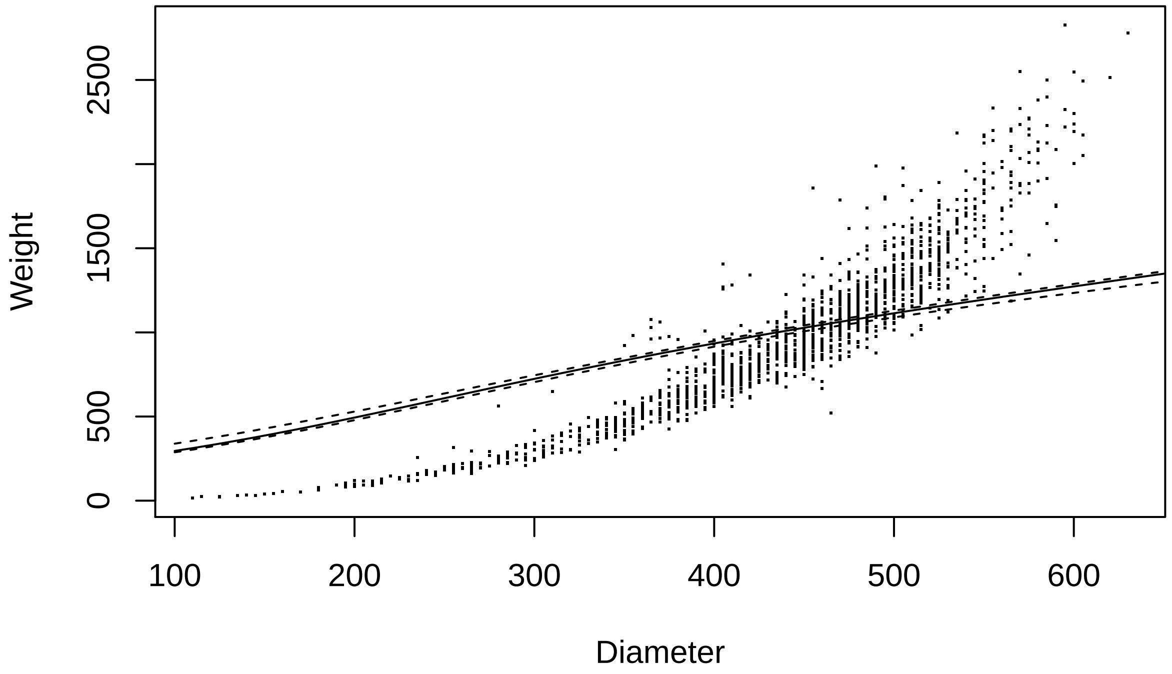
```
## Warning: package 'np' was built under R version 3.4.4
```

```r
n = nrow(abalone)
band=sd(abalone$Weight)/(n^0.2)
kernel1=npreg(Weight~Diameter, data=abalone, bws = band)
```

c

```r
x0=100+5*c(0:110)
x = as.data.frame(x0)
names(x) = c("Diameter")

bootfish=NULL
for(b in 1:1000){
  rows=sample(nrow(abalone),replace=T)
  newfish=abalone[rows,]
  newfit=npreg(Weight~Diameter, bws = band, data = newfish, newdata=x, residuals=T)
  bootfish=rbind(bootfish, newfit$mean)
}

bootquant=apply(bootfish,2,quantile,prob=c(0.025,0.975))
```

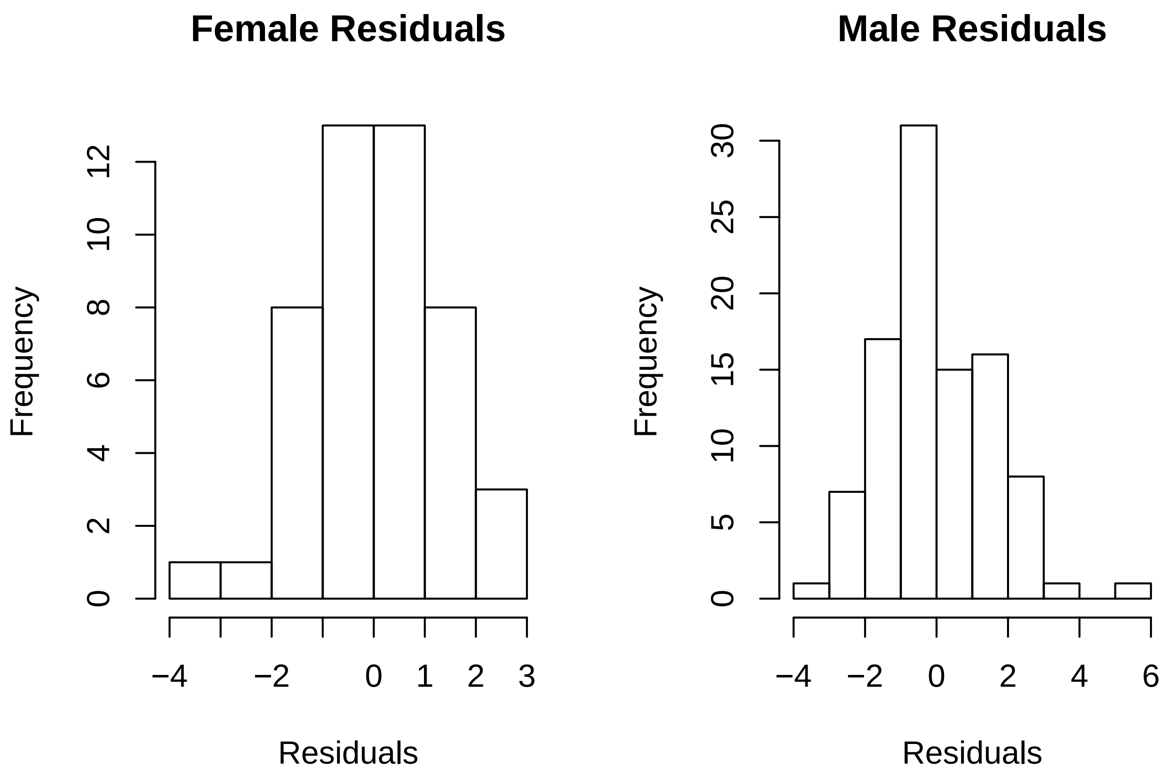
```r
plot(abalone$Diameter,abalone$Weight,pch=".",xlab="Diameter",ylab="Weight")
lines(x0,newfit$mean)
lines(x0,2*newfit$mean-bootquant[1,],lty=2)
lines(x0,2*newfit$mean-bootquant[2,],lty=2)
```



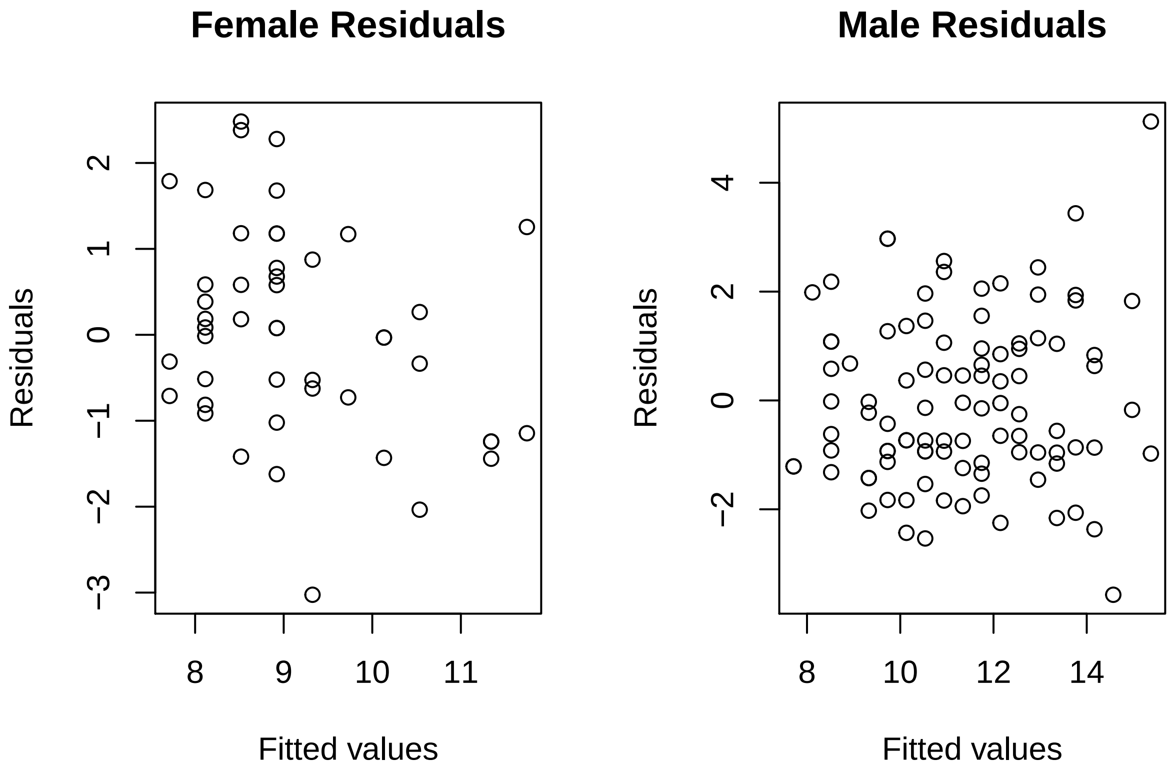
Problem3

a

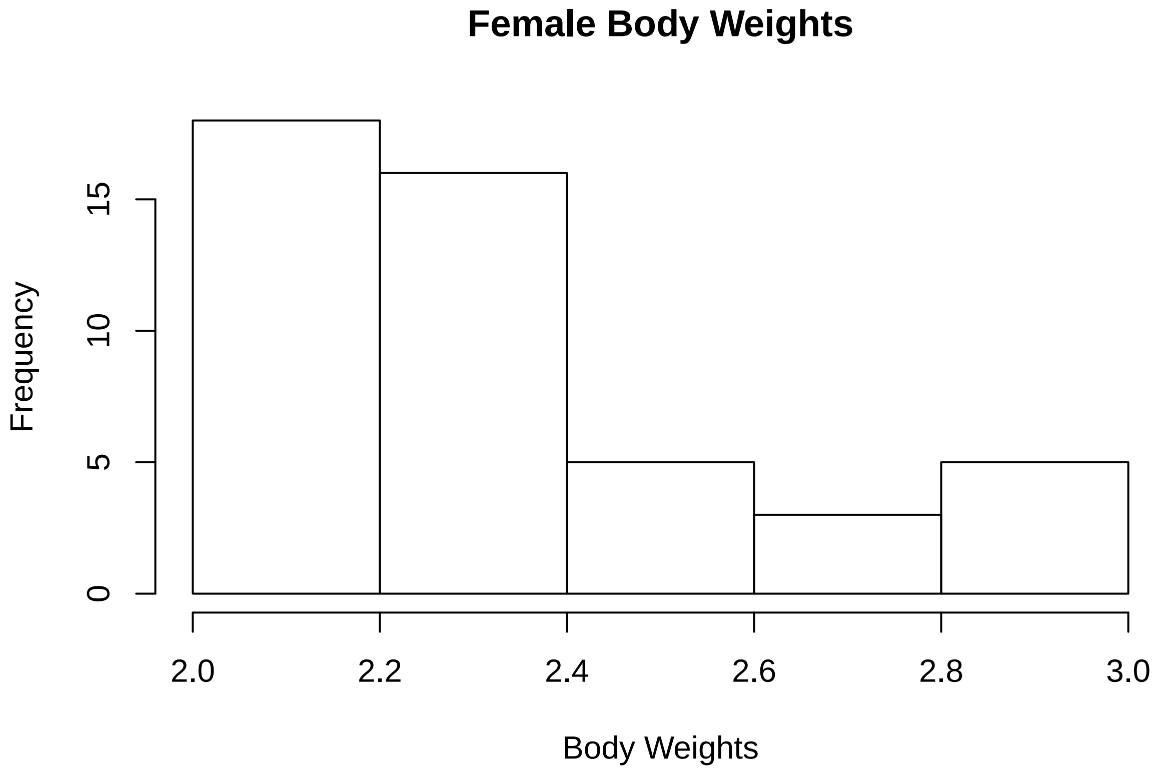
```
library(MASS)
data(cats)

lm1=lm(Hwt~Bwt,data=cats)

par(mfrow=c(1,2))
hist(lm1$resid[cats$Sex=="F"],xlab="Residuals",main="Female Residuals")
hist(lm1$resid[cats$Sex=="M"],xlab="Residuals",main="Male Residuals")
```



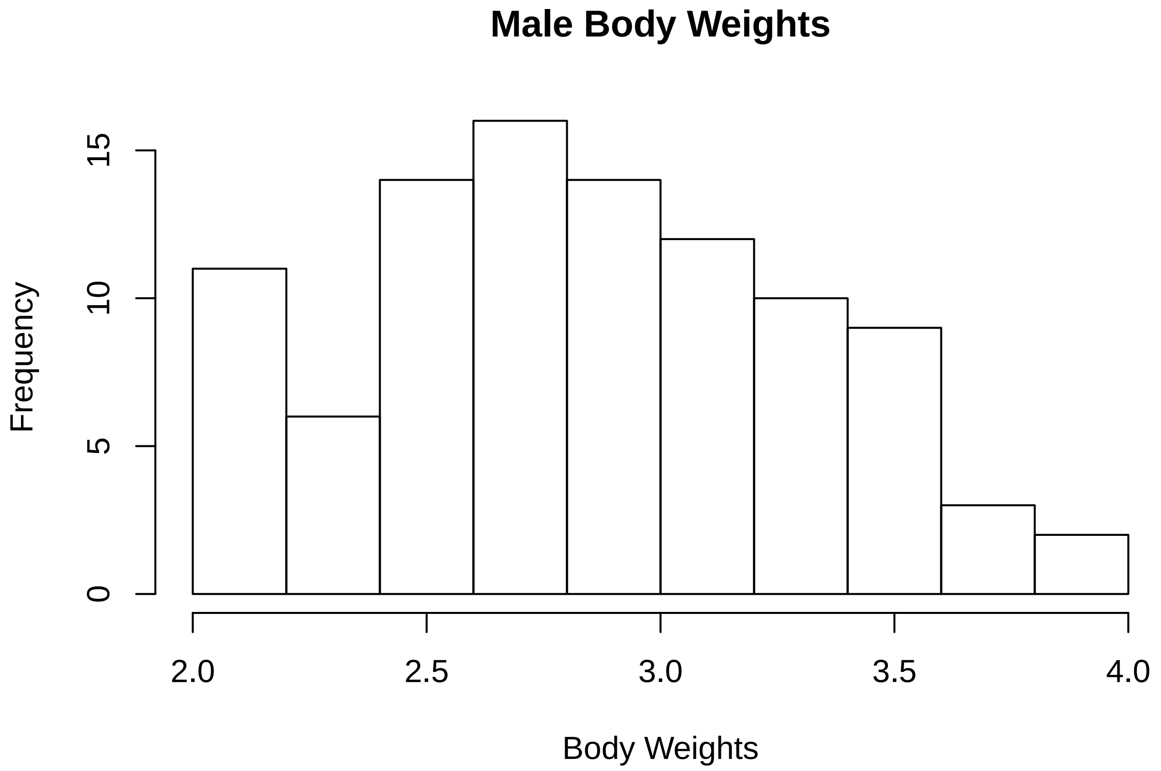
```r
plot(lm1$fitted[cats$Sex=="F"],lm1$resid[cats$Sex=="F"], xlab="Fitted values",ylab="Residuals",main="Fe
plot(lm1$fitted[cats$Sex=="M"],lm1$resid[cats$Sex=="M"], xlab="Fitted values",
     ylab="Residuals", main="Male Residuals")
```



```r
hist(cats$Bwt[cats$Sex=="F"],xlab="Body Weights",main="Female Body Weights")
```



```r
hist(cats$Bwt[cats$Sex=="M"],xlab="Body Weights",main="Male Body Weights")
```



```
par(mfrow=c(1,1))
summary(cats$Bwt[cats$Sex=="F"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    2.15    2.30    2.36    2.50    3.00
```

```
summary(cats$Bwt[cats$Sex=="M"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0     2.5     2.9     2.9     3.2     3.9
```

After a quick EDA on the data, from the hist we see that the male residuals are more spread out, and from the scatter plot we see that female frequency decrease as the residual increases. This suggests that the slope is wrong. And from the hist we see that the male residuals look like a gaussian distribution. Female body weights are smaller than the male body weights.

b

```
lm2=lm(Hwt~Bwt*Sex,data=cats)
summary(lm2)
```

```
##
## Call:
## lm(formula = Hwt ~ Bwt * Sex, data = cats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7728 -1.0118 -0.1196  0.9272  4.8646
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.9813     1.8428   1.618 0.107960
## Bwt           2.6364     0.7759   3.398 0.000885 ***
## SexM         -4.1654     2.0618  -2.020 0.045258 *
## Bwt:SexM      1.6763     0.8373   2.002 0.047225 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.442 on 140 degrees of freedom
## Multiple R-squared:  0.6566, Adjusted R-squared:  0.6493
## F-statistic: 89.24 on 3 and 140 DF,  p-value: < 2.2e-16
```

```
lmF=lm(Hwt~Bwt,data=cats[cats$Sex=="F",])
lmM=lm(Hwt~Bwt,data=cats[cats$Sex=="M",])
summary(lmF)
```

```
##
## Call:
## lm(formula = Hwt ~ Bwt, data = cats[cats$Sex == "F", ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.00871 -0.68599 -0.04506  0.79583  2.21858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.9813     1.4855   2.007 0.050785 .
## Bwt           2.6364     0.6254   4.215 0.000119 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.162 on 45 degrees of freedom
## Multiple R-squared:  0.2831, Adjusted R-squared:  0.2671
## F-statistic: 17.77 on 1 and 45 DF,  p-value: 0.0001186
```

```
summary(lmM)
```

```
##
## Call:
## lm(formula = Hwt ~ Bwt, data = cats[cats$Sex == "M", ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7728 -1.0478 -0.2976  0.9835  4.8646
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.1841     0.9983  -1.186    0.239
## Bwt           4.3127     0.3399  12.688   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.557 on 95 degrees of freedom
## Multiple R-squared:  0.6289, Adjusted R-squared:  0.625
## F-statistic:   161 on 1 and 95 DF,  p-value: < 2.2e-16
```

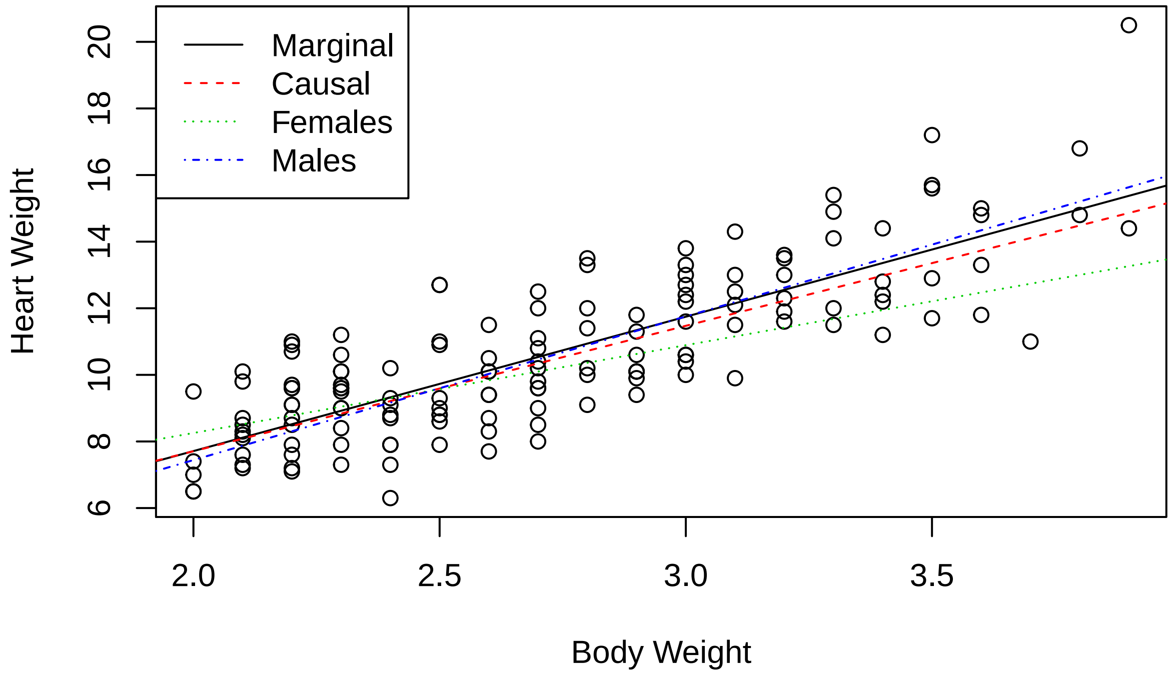
We need: thetahat*rhat(x;F)*p(F) + rhat(x;M)*p(M)

```
means=c(mean(cats$Sex=="F"),mean(cats$Sex=="M"))
thetahat=means[1]*lmF$coef+means[2]*lmM$coef
thetahat
```

```
## (Intercept)         Bwt
##   0.1754524   3.7655646
```

So the line is thetahat(x) = 3.766x + 0.175

```
plot(cats$Bwt,cats$Hwt,xlab="Body Weight",ylab="Heart Weight")
abline(lm1$coef,lty=1,col=1)
abline(thetahat,lty=2,col=2)
abline(lmF$coef,lty=3,col=3)
abline(lmM$coef,lty=4,col=4)
legend("topleft",legend=c("Marginal","Causal","Females","Males "),lty=c(1,2,3,4),col=c(1,2,3,4))
```



The marginal line is useful when we want to predict heart weight for a random cat drawn from the population of the sample. The causal line is useful when we want to predict the heart weight for a random cat drawn from a population where 1/3 of the cats are female. The female/male line is useful when we want to predict the heart weight for a female/male cat.

c

```
datf=(cats$Sex=="F")
datm=(cats$Sex=="M")

B=10000
Tstar=NULL
for(bb in 1:B){
  mnoise=lmM$resid[sample(sum(datm),replace=T)]
  fnoise=lmF$resid[sample(sum(datf),replace=T)]

  newym = data.frame(Bwt=cats$Bwt[cats$Sex=="M"], Hwt=lm1$coef[1]+lm1$coef[2]*cats$Bwt[cats$Sex=="M"]+mn
  newyf=data.frame(Bwt=cats$Bwt[cats$Sex=="F"],
Hwt=lm1$coef[1]+lm1$coef[2]*cats$Bwt[cats$Sex=="F"]+fnoise)

  bootfitM=lm(Hwt~Bwt,data=newym)$coef
  bootfitF=lm(Hwt~Bwt,data=newyf)$coef

  Tstar[bb]=(bootfitF[1]-bootfitM[1])^2+(bootfitF[2]-bootfitM[2])^2
}

tobs=(lmF$coef[1]-lmM$coef[1])^2+(lmF$coef[2]-lmM$coef[2])^2
tobs
```

```
## (Intercept)
##    20.16042
```

```r
pvalue=mean(Tstar>=tobs)
pvalue
```

```
## [1] 0.0163
```

So from the pvalue and tobs we can see that the probability of being in the part above the T=20.16 is 0.017. So we should reject the null hypothesis.