PS2

January 18, 2018

Author:Vatsal Jatakia,Saniya Ambavanekar,Kushal Giri

Submit exactly TWO files: (i) a PDF/HTML file with your write-up and graphs and (ii) a .r/.txt/.Rmd file with code to reproduce your graphs. Miguel will randomly check some of you to ensure your graphs can be reproduced from your code; if not, you'll be penalized.

Groups of 3 have been randomly assigned for this problem set (see "PS2 Groups.") Submit one set of answers per group.

The data set NHANES in the R package of the same name contains data on a representative sample of Americans. Variables include:

For this problem we'll restrict our attention to adults. You can use subset() to create a data frame contain only the people sample aged 18 and up:

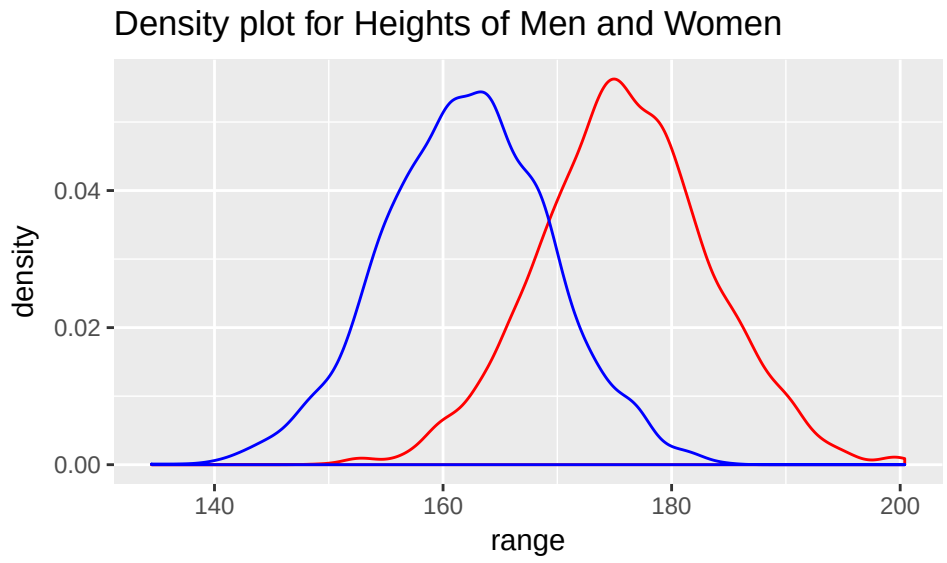
Adults = subset(NHANES, Age >= 18)

Note: There are a few missing values; deal with these as best you can.

```
library(ggplot2)
library(NHANES)
adults = subset(NHANES, Age >= 18)
men_ht = adults$Height[adults$Gender == 'male' ]
men_ht = na.omit(men_ht)
wom_ht = adults$Height[adults$Gender == 'female']
wom_ht = na.omit(wom_ht)
```

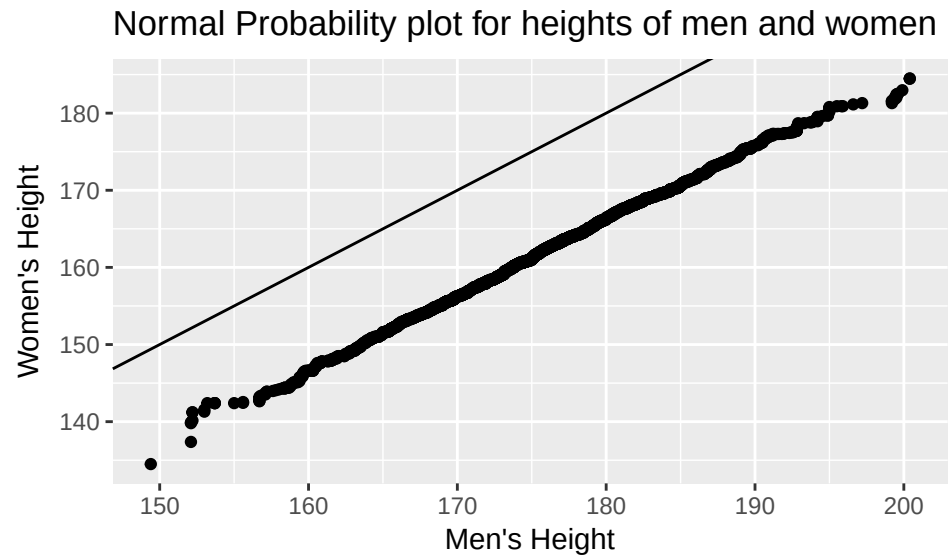
Q1. Submit 2 or 3 graphs (no more or we'll take off points) that let you compare the distributions of adult women's heights and adult men's heights. Does it look like the difference between the distributions is well-approximated by an additive shift or a multiplicative shift, or is it something more complicated? If it's a shift, describe that shift quantitatively; if it's something more complicated, describe the difference in words.

```
ggplot() + geom_density(aes(x=men_ht),
colour="red", data=as.data.frame(men_ht)) +
geom_density(aes(x=wom_ht), colour="blue",                                    data=as.data
  labs(x = "range")
```



Here, red curve corresponds to male and blue curve corresponds to female. We can see that the densities of the men and women have almost similar curve with some distance(shift) which can be observed better observed in the following qqplot.

```
df = as.data.frame(qqplot(men_ht, wom_ht,plot.it = F))
ggplot(df, aes(x = x, y=y),na.rm = T)+geom_point()+
geom_abline()+labs(x = "Men's Height",y= "Women's Height")+
ggtitle("Normal Probability plot for heights of men and women")
```



Looking at the above plot, it is safe to say that there is an additive shift. The shift is calculated as:

```
mean(men_ht)-mean(wom_ht)
```

```
## [1] 13.81445
```

Hence, the shift can be quantitatively described as **13.81** which is also defined as the perpendicular distance between the curve and the standard line. There is a negative shift which means that the rate of increase in x(men's height) is greater than y(women's height).

Q2. Submit 2 or 3 graphs that let you compare the distributions of adult women's weights and adult men's weights. Does it look like the difference between the distributions is well-approximated by an additive shift or

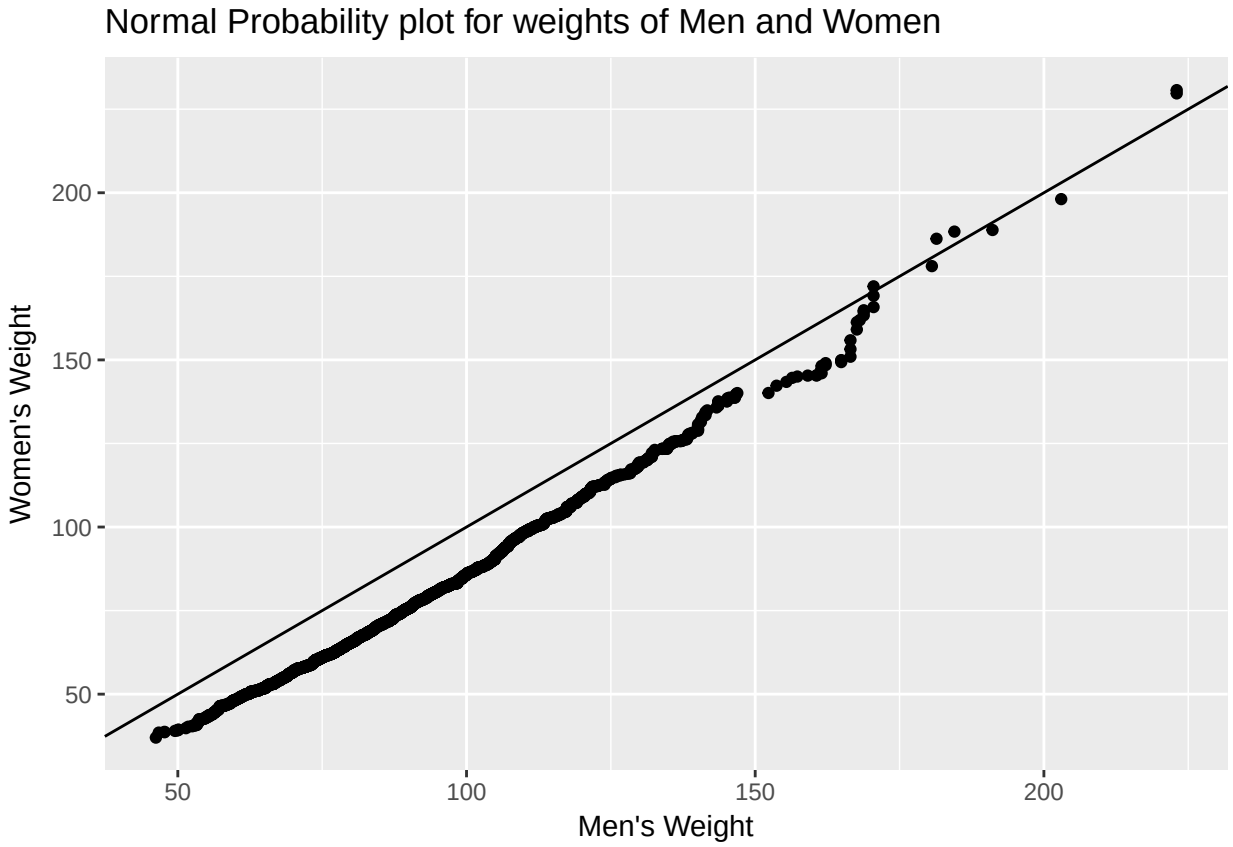
a multiplicative shift, or is it something more complicated? If it's a shift, describe that shift quantitatively; if it's something more complicated, describe the difference in words.

```
men_wt<-adults$Weight[adults$Gender=='male']
men_wt<-na.omit(men_wt)

wom_wt<-adults$Weight[adults$Gender=='female']
wom_wt<-na.omit(wom_wt)
```

QQ plot for Weight of men Vs Weight of women

```
df1 = as.data.frame(qqplot(men_wt, wom_wt,plot.it = F))
ggplot(df1, aes(x = x, y=y),na.rm = T)+geom_point()+
geom_abline()+labs(x="Men's Weight", y = "Women's Weight")+
  ggtitle("Normal Probability plot for weights of Men and Women")
```



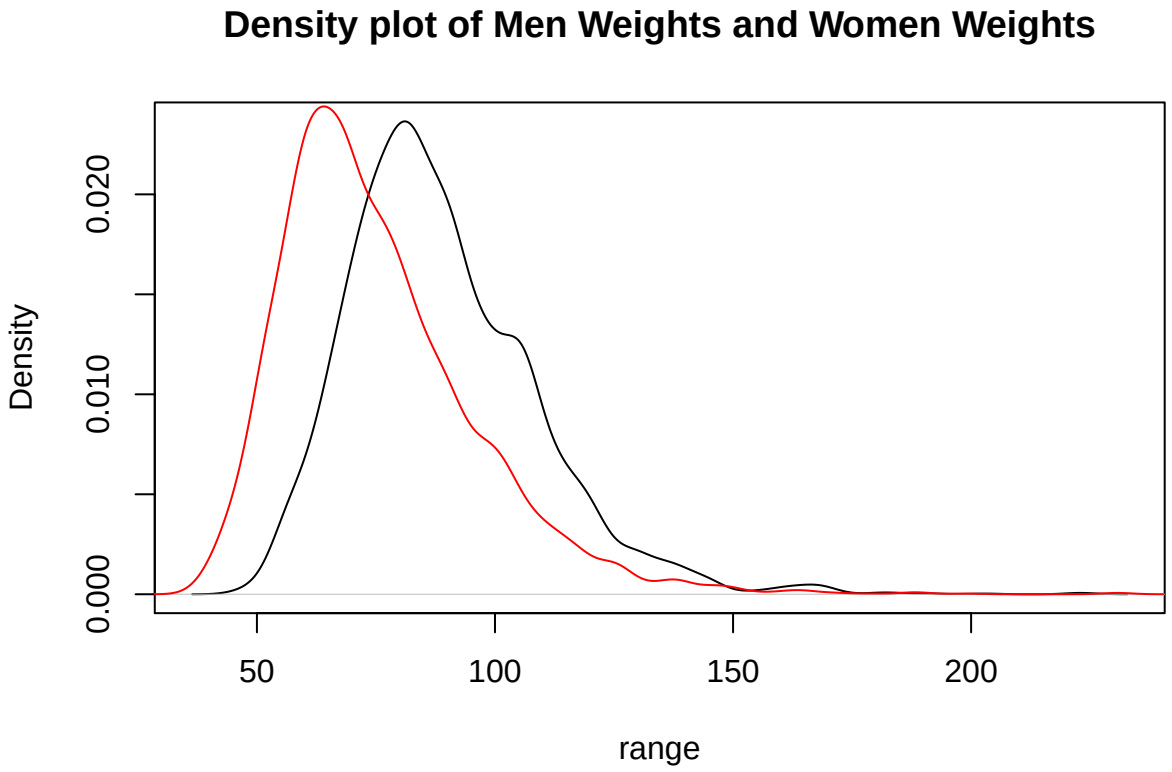
From the observation of the QQ plot, we can observe a weak additive shift. The shift is observed for a major part of the curve except at the top extreme where it coincides with the line.The shift can be quantitatively defined as:

```
mean(men_wt)-mean(wom_wt)
```

```
## [1] 13.50522
```

The shift can be further verified by the density plot.

```
plot(density(men_wt),
     main = "Density plot of Men Weights and Women Weights",
     xlab = "range")
lines(density(wom_wt),col="red")
```

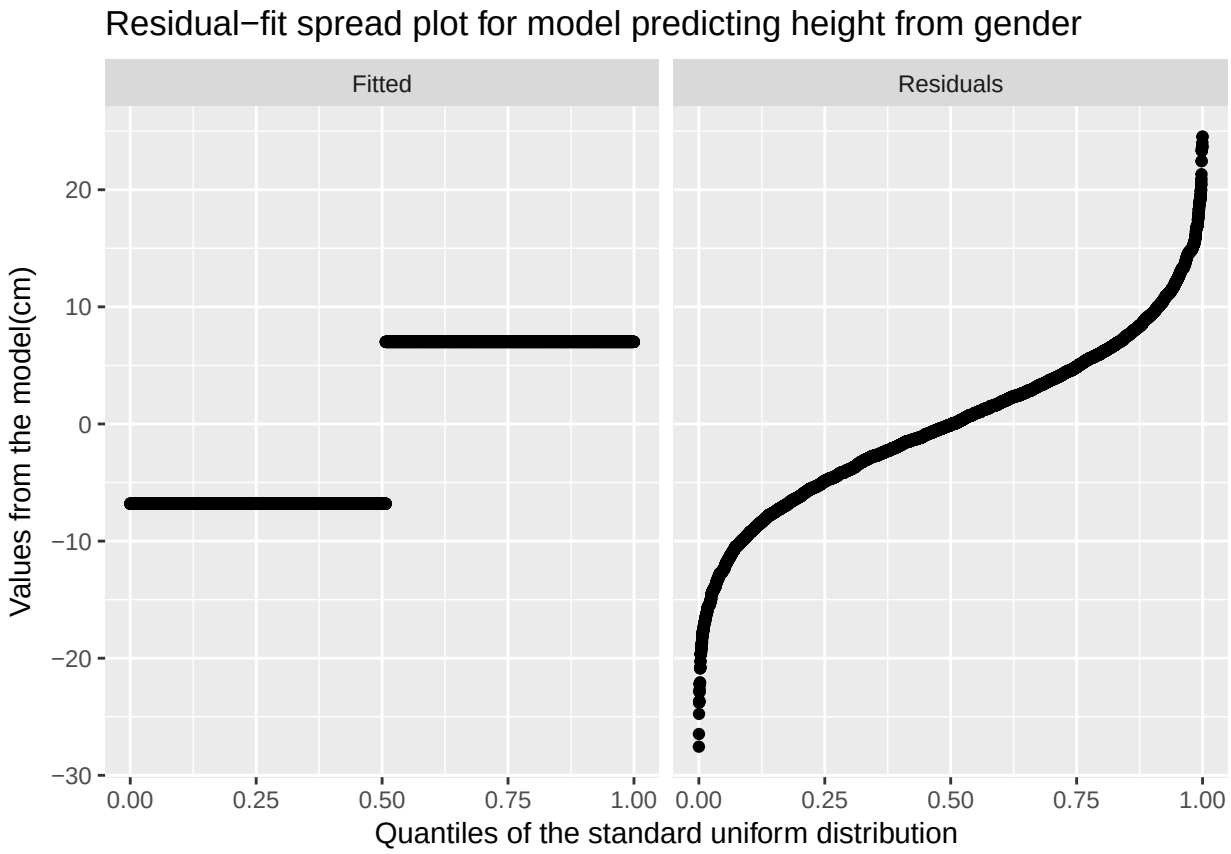


Here the red curve corresponds to women and the black curve corresponds to men.

From the density plots we can observe there is a negative shift in case of the weights of the men and women. The rate at which x increases is higher than y and hence the weights of men are higher than women for most parts of the curve.

Q3.Consider the following model for the dependence of height on gender: lm(Height ~ Gender, data = Adults) Reproduce this residual-fit spread plot:

```
library(tidyr)
Adults = subset(NHANES, Age >= 18)
dataset=na.omit(data.frame(height=Adults$Height,gender=Adults$Gender))
model=lm(height ~gender,data=dataset)
adult.fitted=sort(fitted.values(model)-mean(fitted.values(model)))
adult.residuals=sort(residuals(model))
n = length(adult.residuals)
f.value = (0.5:(n - 0.5))/n
adult.fit = data.frame(f.value, Fitted = adult.fitted, Residuals = adult.residuals)
adult.fit.long = adult.fit %>% gather(type, value, Fitted:Residuals)
ggplot(adult.fit.long, aes(sample = value)) +
stat_qq(distribution = "qunif") +facet_grid(~type)+
labs(title="Residual-fit spread plot for model predicting height from gender",
x = "Quantiles of the standard uniform distribution", y = "Values from the model(cm)")
```




```
Rsquare=var(adult.fitted)/var(dataset$height)
```

The coefficient of determination value is 0.4671;which implies that the model captures approximately 47% of the variance of the data.The remaining 53% variance lies in the residuals of the model.

Also,the left side of the plot(centered fitted values) is shorter than the right side(residual values).So,the spread of the residuals is large relative to the spread of the fitted values. Also,the distribution of the residual plot doesn't seem to be normally distributed;which implies that the model might not fit the data properly.