S670 PS2

Natasha Randall, Jeevan Rachepalli, Leo Huang

January 19, 2018

Clean and Summarize the Data

```
library(NHANES)

#Return the subset of the dataset with adults (18 years old or older).
adults = subset(NHANES, Age >= 18)
dim(adults)
```

```
## [1] 7481   76
```

```
#Extract the 4 columns needed for the subsequent analysis.
adult = adults[,c("Gender","Age","Height","Weight")]

#View summaries of relevant columns of data:
head(adult)
```

```
##   Gender Age Height Weight
## 1   male  34  164.7   87.4
## 2   male  34  164.7   87.4
## 3   male  34  164.7   87.4
## 5 female  49  168.4   86.7
## 8 female  45  166.7   75.7
## 9 female  45  166.7   75.7
```

```
str(adult)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    7481 obs. of  4 variables:
##  $ Gender: Factor w/ 2 levels "female","male": 2 2 2 1 1 1 1 2 2 2 ...
##  $ Age   : int  34 34 34 49 45 45 45 66 58 54 ...
##  $ Height: num  165 165 165 168 167 ...
##  $ Weight: num  87.4 87.4 87.4 86.7 75.7 75.7 75.7 68 78.4 74.7 ...
```

```
summary(adult)
```

```
##     Gender           Age            Height          Weight      
##  female:3795   Min.   :18.00   Min.   :134.5   Min.   : 37.0  
##  male  :3686   1st Qu.:31.00   1st Qu.:161.5   1st Qu.: 66.8  
##                Median :45.00   Median :168.8   Median : 79.3  
##                Mean   :46.23   Mean   :168.9   Mean   : 82.0  
##                3rd Qu.:59.00   3rd Qu.:176.1   3rd Qu.: 93.9  
##                Max.   :80.00   Max.   :200.4   Max.   :230.7  
##                                NA's   :57      NA's   :61
```

```
#Count of rows with missing data.
sum(!complete.cases(adult))
```

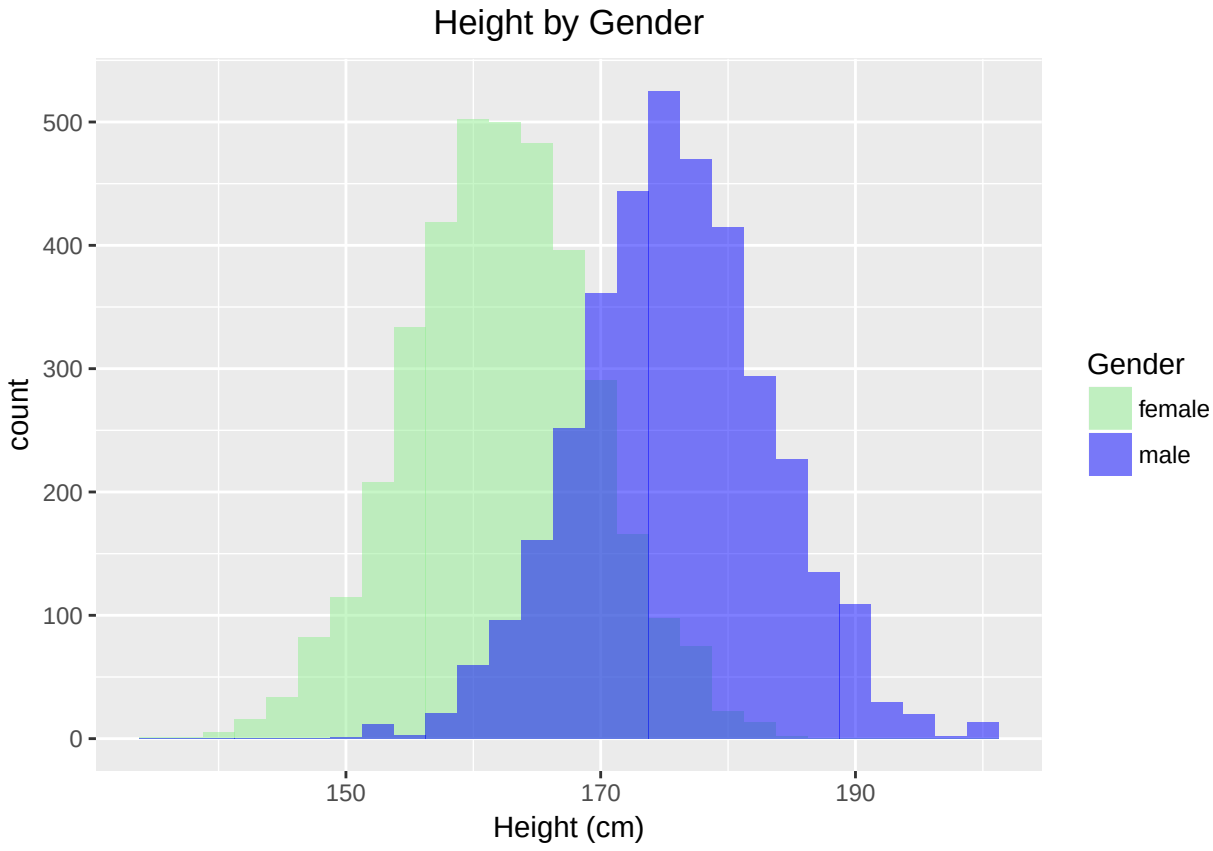
```
## [1] 67
```

```r
# Less than 1% of missing data so we are going to drop those rows
adult = adult[complete.cases(adult),]
dim(adult)
```

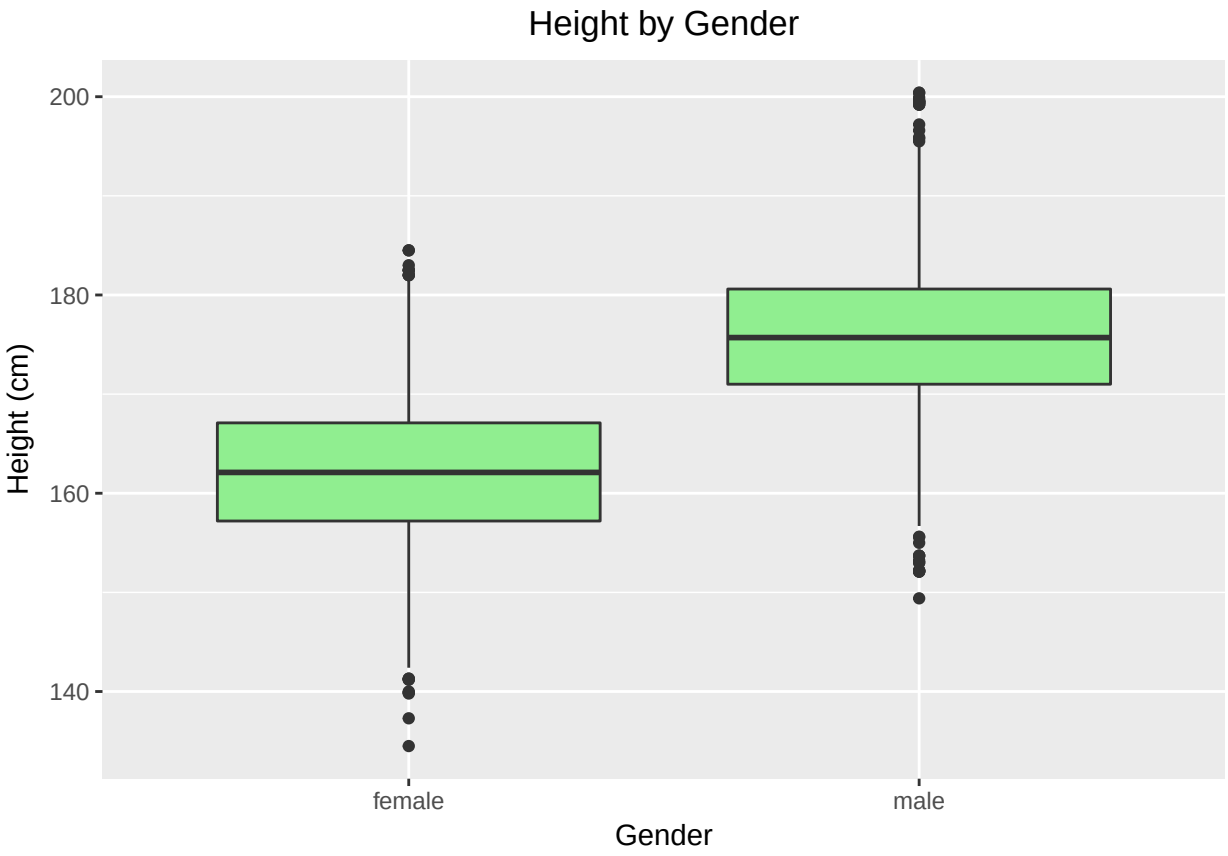
```
## [1] 7414    4
```

Q1

```r
library(ggplot2)

#Create an overlapping density plot of Height by Gender (Female vs Male).
ggplot(adult,aes(x=Height,fill=Gender)) +
  geom_histogram(position="identity", alpha=0.5, binwidth = 2.5) +
  ggtitle("Height by Gender") +
  xlab("Height (cm)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_fill_manual(values = c("female" = "light green", "male" = "blue"))
```

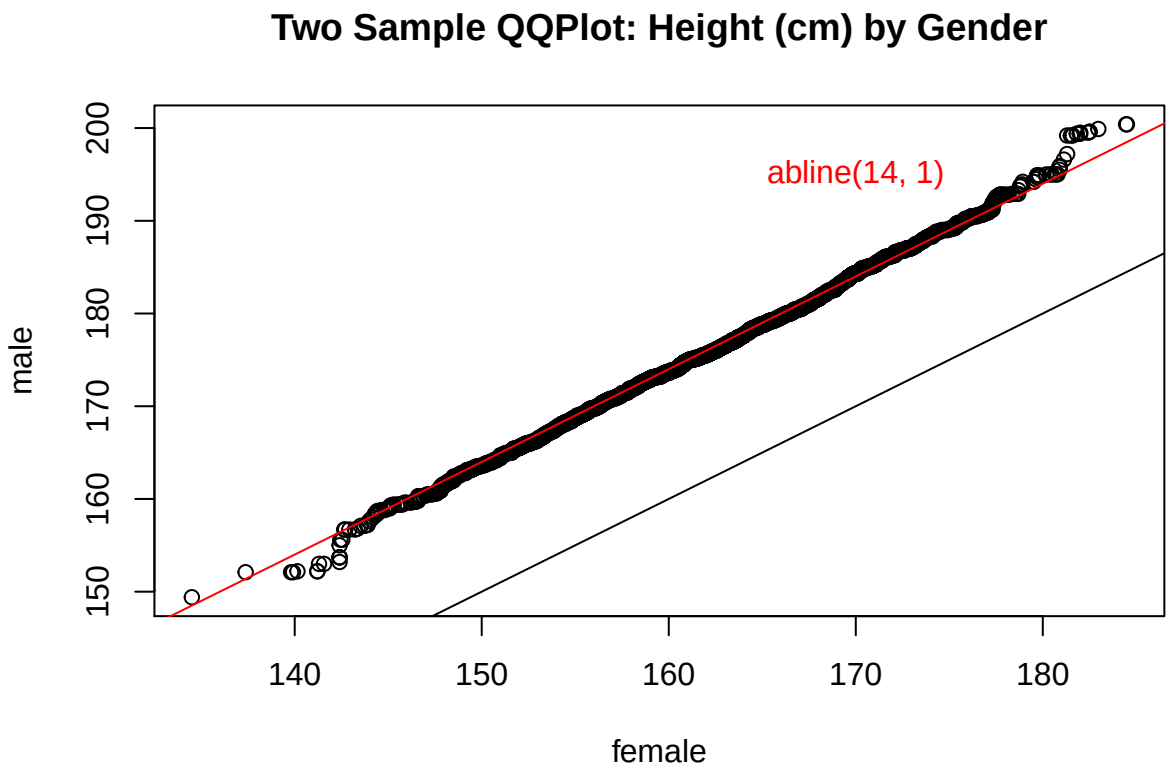



```r
#Create a boxplot plot by Gender.
ggplot(adult,aes(x=Gender, y= Height)) +
  geom_boxplot(fill="light green") +
  ggtitle("Height by Gender") +
  ylab("Height (cm)") +
  theme(plot.title = element_text(hjust = 0.5))
```



Based on the graphs above, the heights of both groups (male and female) is nearly normal and have similar variance.

```
#Plot two sample QQplot.
male = adult$Height[adult$Gender == "male"]
female = adult$Height[adult$Gender == "female"]
qqplot(female, male, main="Two Sample QQPlot: Height (cm) by Gender")
abline(0, 1)
abline(14, 1, col='red')
text(170, 195, labels="abline(14, 1)", col="red")
```

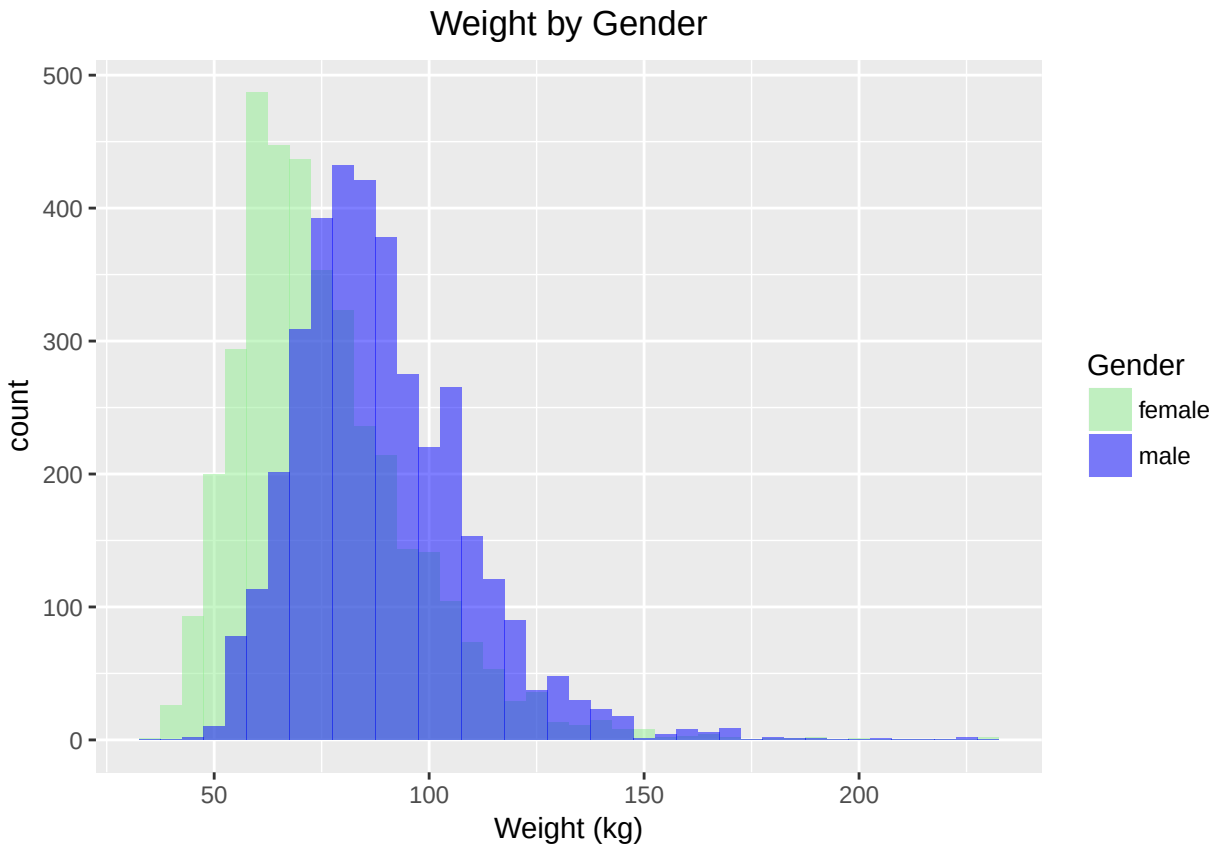



```
#Same two sample QQplot plotted with ggplot
# adult.df = as.data.frame(qqplot(female,male,plot.it = FALSE))
# ggplot(adult.df,aes(x = x, y = y)) + geom_point() + geom_abline() +
#   geom_abline(intercept=14,slope=1,col='red')
```

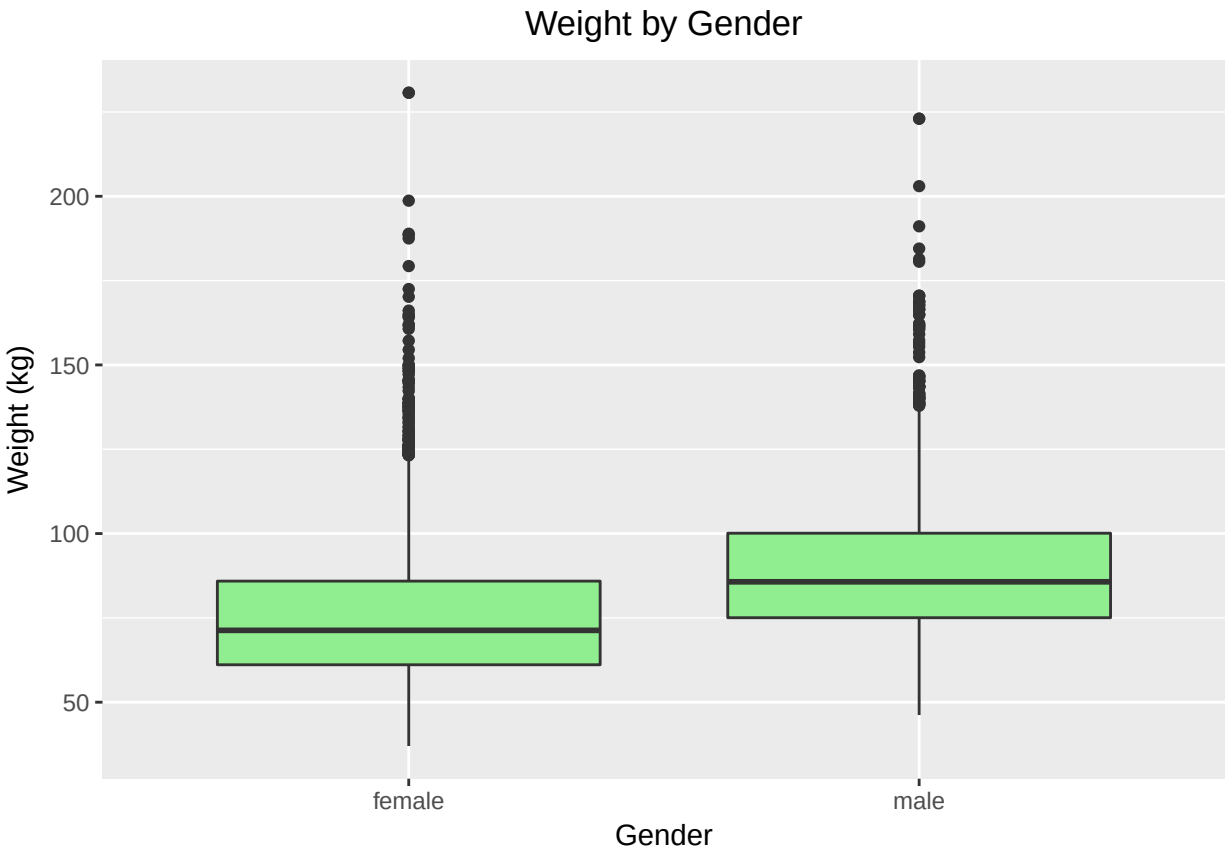
As can be seen from the two sample qqplot, the distributions are well-approximated by an additive shift (an increase in intercept (mean) of approximately 14 cm in males compare with females). It should be noted that there is very slight deviation from this additive shift in the tails (the extreme values).

Q2

```
#Create an overlapping density plot of Height by Gender (Female vs Male).
ggplot(adult, aes(x=Weight, fill=Gender)) +
  geom_histogram(position="identity", alpha=0.5, binwidth = 5) +
  ggtitle("Weight by Gender") +
  xlab("Weight (kg)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_fill_manual(values = c("female" = "light green", "male" = "blue"))
```

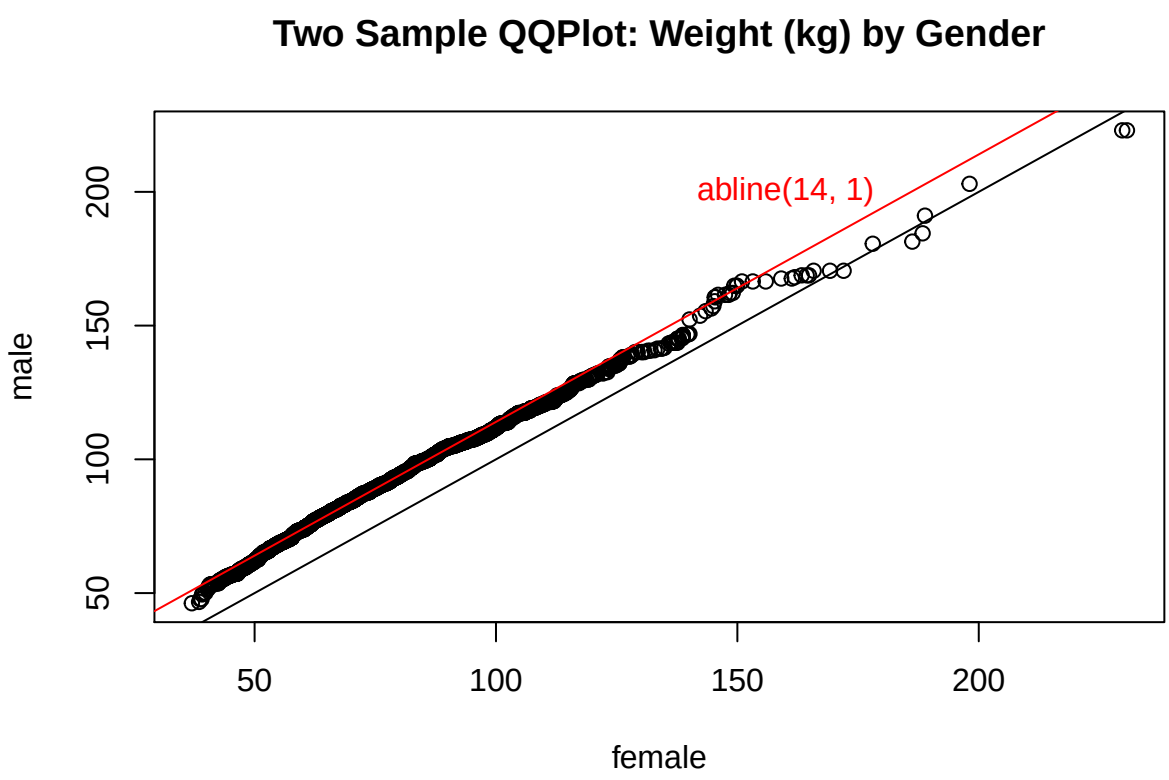



```
#Create a boxplot plot by Gender.
ggplot(adult,aes(x=Gender, y=Weight)) +
  geom_boxplot(fill="light green") +
  ggtitle("Weight by Gender") +
  ylab("Weight (kg)") +
  theme(plot.title = element_text(hjust = 0.5))
```



The two graphs above show that the distributions of weight by gender (female vs male) are both positively skewed. The variance of both groups appear similar.

```
#Plot two sample QQplot.
male = (adult$Weight[adult$Gender == "male"])
female = (adult$Weight[adult$Gender == "female"])
qqplot(female,male, main="Two Sample QQPlot: Weight (kg) by Gender")
abline(0, 1)
abline(14, 1, col="red")
text(160, 200, labels="abline(14, 1)", col="red")
```




```
#Same two sample QQplot plotted with ggplot
# adult.df = as.data.frame(qqplot(female,male,plot.it = FALSE))
# ggplot(adult.df,aes(x = x, y = y)) + geom_point() + geom_abline() +
#   geom_abline(intercept=14,slope=1,col='red') + xlab("Female weights in Kilograms") +
#   ylab("Male weights in Kilograms") + ggtitle("Male weight vs Female Weight")
```

As can be seen from the two sample qqplot, the distributions are well-approximately by an additive shift for around 80% of the distributions (an increase in intercept (mean) of approximately 14 kg in males compared with females). However, for higher weights (above approx. the 80th percentile), this is no longer true. Weights are much lower for men than an additive shift would predict. [It actually appears that weights are about the same for female and males in this part of the distributions.]

Q3

```
library(tidyr)

#Linear model of height predicted by gender.
height.lm = lm(Height ~ Gender, data = adult)
summary(height.lm)
```

```
##
## Call:
## lm(formula = Height ~ Gender, data = adult)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -27.5629  -4.8629  -0.0629   4.8404  24.5404
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 162.0629     0.1203 1347.36   <2e-16 ***
## Gendermale   13.7968     0.1714   80.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.378 on 7412 degrees of freedom
## Multiple R-squared:  0.4664, Adjusted R-squared:  0.4663
## F-statistic:  6479 on 1 and 7412 DF,  p-value: < 2.2e-16
```

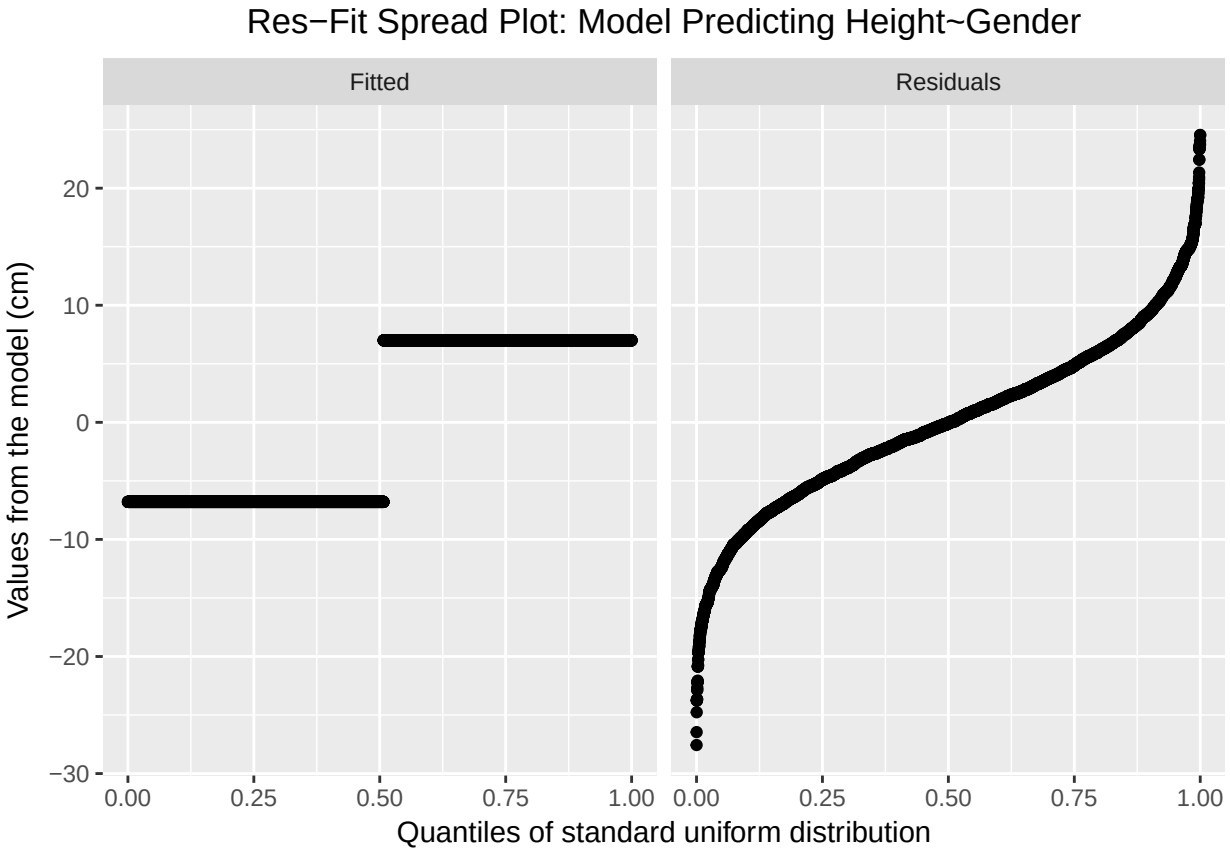
```
#Fitted and residual vales from model.
adult.fitted = fitted(height.lm) - mean(fitted(height.lm))
adult.res = residuals(height.lm)

#Create a data frame of fitted and residual values from lm model.
adult.lmdf = data.frame(Fitted=adult.fitted, Residuals=adult.res)

#Must transform from wide format to long format to make residual-fit spread plot.
adult.gather = gather(adult.lmdf, key=type, value=value, Fitted:Residuals)
head(adult.gather)
```

```
##     type     value
## 1 Fitted  7.002587
## 2 Fitted  7.002587
## 3 Fitted  7.002587
## 4 Fitted -6.794165
## 5 Fitted -6.794165
## 6 Fitted -6.794165
```

```
tail(adult.gather)
```

```
##            type      value
## 14823 Residuals  1.4403725
## 14824 Residuals  1.4403725
## 14825 Residuals  1.4403725
## 14826 Residuals -0.0596275
## 14827 Residuals -7.0596275
## 14828 Residuals -7.0596275
```

```
#Create residual-fit spread plot.
ggplot(adult.gather, aes(sample=value)) +
  stat_qq(distribution="qunif") +
  facet_grid(~type) +
  xlab("Quantiles of standard uniform distribution") +
  ylab("Values from the model (cm)") +
  ggtitle("Res-Fit Spread Plot: Model Predicting Height~Gender") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
var(adult.fitted)/(var(adult.fitted)+var(adult.res))
```

```
## [1] 0.4664202
```

This plot implies that it is not a good model as we can see the spread of the fitted values from the model is narrow and the spread of the residuals is broad. This means that a decent chunk of variation remains in the residuals instead of the fitted values in the model. In addition, there are only two values available in the fitted values for the prediction. This limitation comes from the fact that we are predicting heights only based on gender observations. These findings can also be verified by the R^2 value (i.e. only 0.466) of the model. [Note that R^2 should be interpreted based on context/field, as this would be considered quite high in the behavioral sciences.]