Rahul Khandekar

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*. #0

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------

## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(tidyr)
library(cluster)
library(ggplot2)
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
library(vegan)
```

```
## Loading required package: permute

## Loading required package: lattice

## This is vegan 2.5-6
```

```r
library(tidyverse)
library(dplyr)
select<-dplyr::select
data_1<-read.csv("data 1.csv")
data_2<-read.csv("data 2.csv")
data_2 %>% inner_join(data_1, by="ï..ID") %>% na.omit() %>% glimpse()
```

```
## Rows: 202
## Columns: 14
## $ ï..ID  <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1...
## $ ssf    <dbl> 109.1, 102.8, 104.6, 126.4, 80.3, 75.2, 87.2, 97.9, 75.1, 65...
## $ pcBfat <dbl> 19.75, 21.30, 19.88, 23.66, 17.64, 15.58, 19.99, 22.43, 17.9...
## $ lbm    <dbl> 63.32, 58.55, 55.36, 57.18, 53.20, 53.77, 60.17, 48.33, 54.5...
## $ ht     <dbl> 195.9, 189.7, 177.8, 185.0, 184.6, 174.0, 186.2, 173.8, 171....
## $ wt     <dbl> 78.9, 74.4, 69.1, 74.9, 64.6, 63.7, 75.2, 62.3, 66.5, 62.9, ...
## $ sex    <chr> "f", "f", "f", "f", "f", "f", "f", "f", "f", "f", "f", "f", ...
## $ sport  <chr> "B_Ball", "B_Ball", "B_Ball", "B_Ball", "B_Ball", "B_Ball", ...
## $ rcc    <dbl> 3.96, 4.41, 4.14, 4.11, 4.45, 4.10, 4.31, 4.42, 4.30, 4.51, ...
## $ wcc    <dbl> 7.5, 8.3, 5.0, 5.3, 6.8, 4.4, 5.3, 5.7, 8.9, 4.4, 5.3, 7.3, ...
## $ hc     <dbl> 37.5, 38.2, 36.4, 37.3, 41.5, 37.4, 39.6, 39.9, 41.1, 41.6, ...
## $ hg     <dbl> 12.3, 12.7, 11.6, 12.6, 14.0, 12.5, 12.8, 13.2, 13.5, 12.7, ...
## $ ferr   <int> 60, 68, 21, 69, 29, 42, 73, 44, 41, 44, 38, 26, 30, 48, 30, ...
## $ bmi    <dbl> 20.56, 20.67, 21.86, 21.88, 18.96, 21.04, 21.69, 20.62, 22.6...
```

```r
data_3 <- data_2 %>% inner_join(data_1, by="ï..ID") %>% na.omit() %>% glimpse()
```

```
## Rows: 202
## Columns: 14
## $ ï..ID  <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1...
## $ ssf    <dbl> 109.1, 102.8, 104.6, 126.4, 80.3, 75.2, 87.2, 97.9, 75.1, 65...
## $ pcBfat <dbl> 19.75, 21.30, 19.88, 23.66, 17.64, 15.58, 19.99, 22.43, 17.9...
## $ lbm    <dbl> 63.32, 58.55, 55.36, 57.18, 53.20, 53.77, 60.17, 48.33, 54.5...
## $ ht     <dbl> 195.9, 189.7, 177.8, 185.0, 184.6, 174.0, 186.2, 173.8, 171....
## $ wt     <dbl> 78.9, 74.4, 69.1, 74.9, 64.6, 63.7, 75.2, 62.3, 66.5, 62.9, ...
## $ sex    <chr> "f", "f", "f", "f", "f", "f", "f", "f", "f", "f", "f", "f", ...
## $ sport  <chr> "B_Ball", "B_Ball", "B_Ball", "B_Ball", "B_Ball", "B_Ball", ...
## $ rcc    <dbl> 3.96, 4.41, 4.14, 4.11, 4.45, 4.10, 4.31, 4.42, 4.30, 4.51, ...
## $ wcc    <dbl> 7.5, 8.3, 5.0, 5.3, 6.8, 4.4, 5.3, 5.7, 8.9, 4.4, 5.3, 7.3, ...
## $ hc     <dbl> 37.5, 38.2, 36.4, 37.3, 41.5, 37.4, 39.6, 39.9, 41.1, 41.6, ...
## $ hg     <dbl> 12.3, 12.7, 11.6, 12.6, 14.0, 12.5, 12.8, 13.2, 13.5, 12.7, ...
## $ ferr   <int> 60, 68, 21, 69, 29, 42, 73, 44, 41, 44, 38, 26, 30, 48, 30, ...
## $ bmi    <dbl> 20.56, 20.67, 21.86, 21.88, 18.96, 21.04, 21.69, 20.62, 22.6...
```

```r
nrow(data_3)
```

```
## [1] 202
```

```r
ncol(data_3)
```

```
## [1] 14
```

I used the "ais" dataset, which has physical characteristics of Australian athletes and divides them on the sport they play. I broke the dataset into two separate datasets, and then I used an "inner join" function to merge them together. There are 12 numerical variables (physical characteristics) and 2 categorical variables (sport type and sex). The "sex" variable was treated as a binary variable with males and females as the two options which can be used. BMI was used as a numerical variable, and this refers to an athlete's body mass index. "Ht" refers to the height of an athlete, and "wt" refers to the weight of an athlete. The categorical variable of "sport" indicates which sport an athlete competed in.

#1

```
gg3<-ggplot(data_3, aes(x = ht, y = wt)) + geom_point(alpha = .5) + geom_density_2d(h=2) + coord_fixed()
manva<-manova(cbind(ht,wt)~sport, data=data_3)
summary(manva)
```

```
##            Df  Pillai approx F num Df den Df    Pr(>F)
## sport       9 0.78279   13.719     18    384 < 2.2e-16 ***
## Residuals 192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary.aov(manva)
```

```
##  Response ht :
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## sport         9  7592.1  843.57   14.14 < 2.2e-16 ***
## Residuals   192 11454.7   59.66
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response wt :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## sport         9  17121 1902.27  16.711 < 2.2e-16 ***
## Residuals   192  21856  113.83
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
manva6<-manova(cbind(ht,wt)~sex, data=data_3)
summary(manva6)
```

```
##            Df  Pillai approx F num Df den Df    Pr(>F)
## sex         1 0.34538   52.497      2    199 < 2.2e-16 ***
## Residuals 200
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary.aov(manva6)
```

```
##  Response ht :
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## sex           1  6012.4  6012.4  92.254 < 2.2e-16 ***
## Residuals   200 13034.4    65.2
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response wt :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## sex           1  11638 11638.0  85.141 < 2.2e-16 ***
## Residuals   200  27338   136.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
data_3%>% group_by(sport)%>% summarise(mean(ht),mean(wt))
```

```
## # A tibble: 10 x 3
##    sport   `mean(ht)` `mean(wt)`
##    <chr>        <dbl>      <dbl>
##  1 B_Ball        189.       79.8
##  2 Field         181.       90.0
##  3 Gym           153.       43.6
##  4 Netball       176.       69.6
##  5 Row           182.       78.5
##  6 Swim          181.       75.1
##  7 T_400m        175.       64.0
##  8 T_Sprnt       176.       71.5
##  9 Tennis        174.       64.5
## 10 W_Polo        188.       86.7
```

```
data_3%>% group_by(sex)%>% summarise(mean(ht),mean(wt))
```

```
## # A tibble: 2 x 3
##   sex   `mean(ht)` `mean(wt)`
##   <chr>      <dbl>      <dbl>
## 1 f           175.       67.3
## 2 m           186.       82.5
```

```
1-(.95)^3
```

```
## [1] 0.142625
```

```
.05/3
```

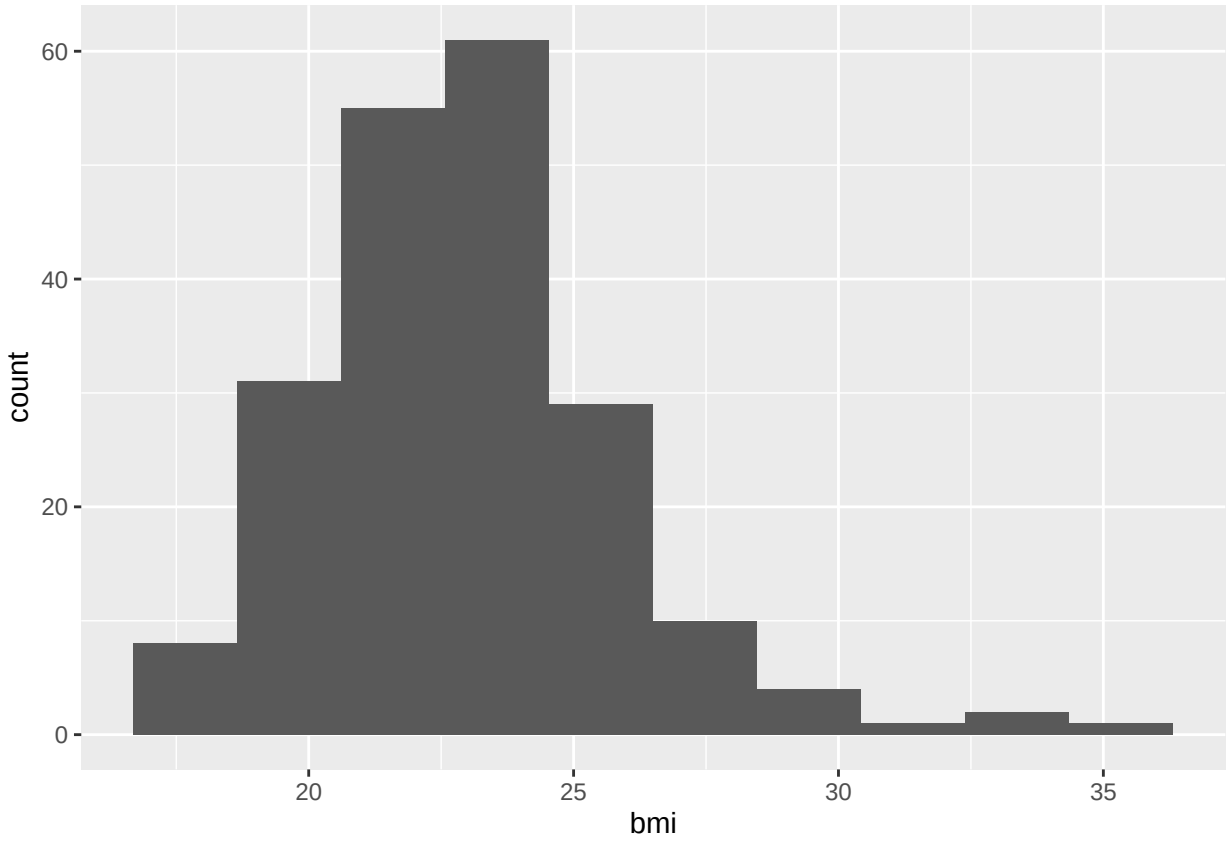
```
## [1] 0.01666667
```

After running a MANOVA test across two numeric variables, height "ht" and weight "wt", there was a significant difference found across the categorical variables of sport "sport" and gender "sex". The null hypothesis is that there is not a significant difference for height or weight for the categorical variables. The alternate hypothesis is that there is a significant difference between the variables for height and weight. By running a MANOVA test, we assume that samples are observed independently, there are no multivariate outliers, there is a sample size of at least 25, and there is no collinearity. The assumptions were probably not all met, as there are outliers since it is a comprehensive record of all Australian athletes. The summary statistics show that there is a significant difference between at least one of the groups. Because of this, a post hoc test was run to see which groups differed and if multiple groups did, how many of them differed. The p

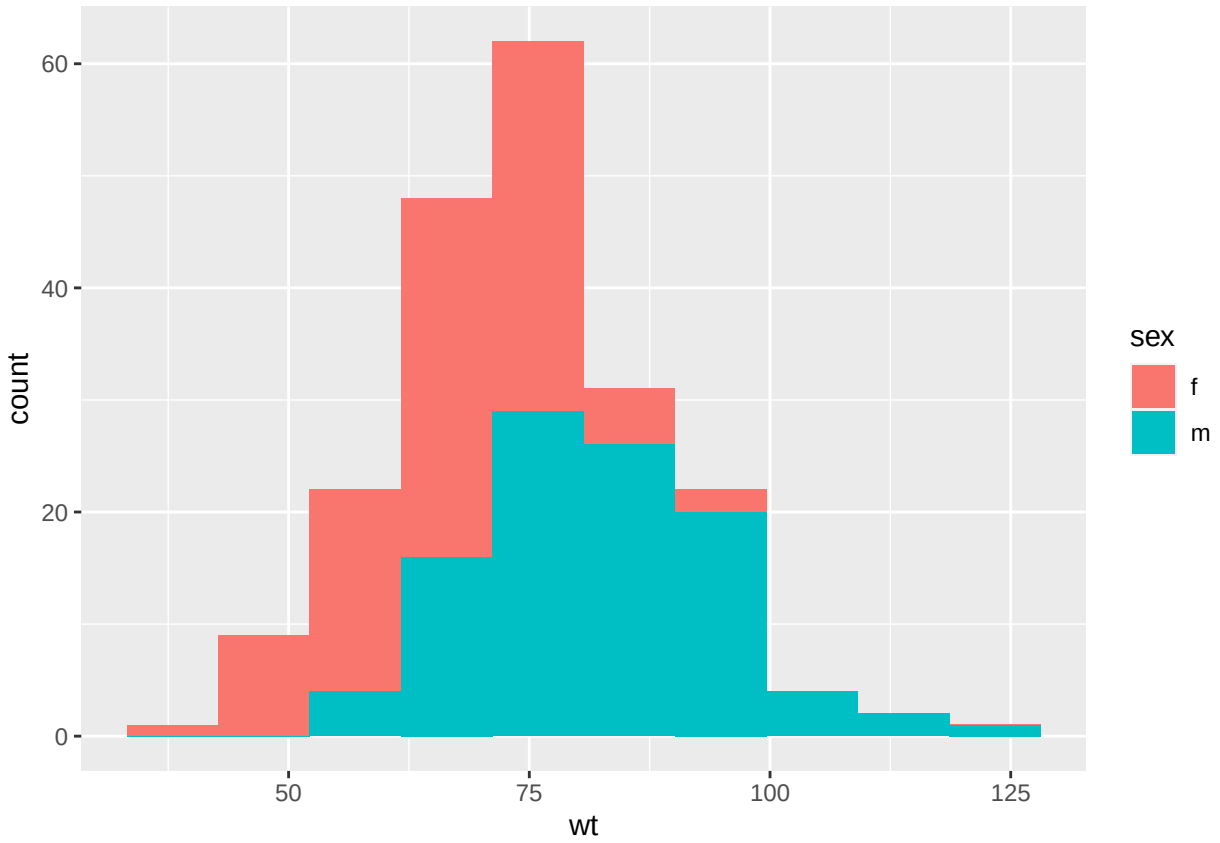
value of (<2.2e-16) shows there is a significant difference. The height "ht" and weight "wt" numeric variables both had significant differences across the gender "sex" binomial variable. Therefore, the null hypothesis that there is no significant difference across genders for height or weight is rejected. The same test was run but with sport as the categorical variable, and the null hypothesis was rejected once again with a p values of <2.2e-16. After this, the probability of a type one error was found. Because three tests were run, the equation of 1-(.95)^3 gave the probability of a type one error as 0.142625. Neither of the ANOVA tests support the null hypothesis as they both gave p values below 2.2e-16. I then did a bonferroni correction and this equation is given by 0.05/3= 0.01666667. Both of the tests still reject the null hypothesis as p < 2.2e-16 is less than 0.017. Therefore, the MANOVA and ANOVA tests gave the same result of significant differences between height and weight across sexes.

#2

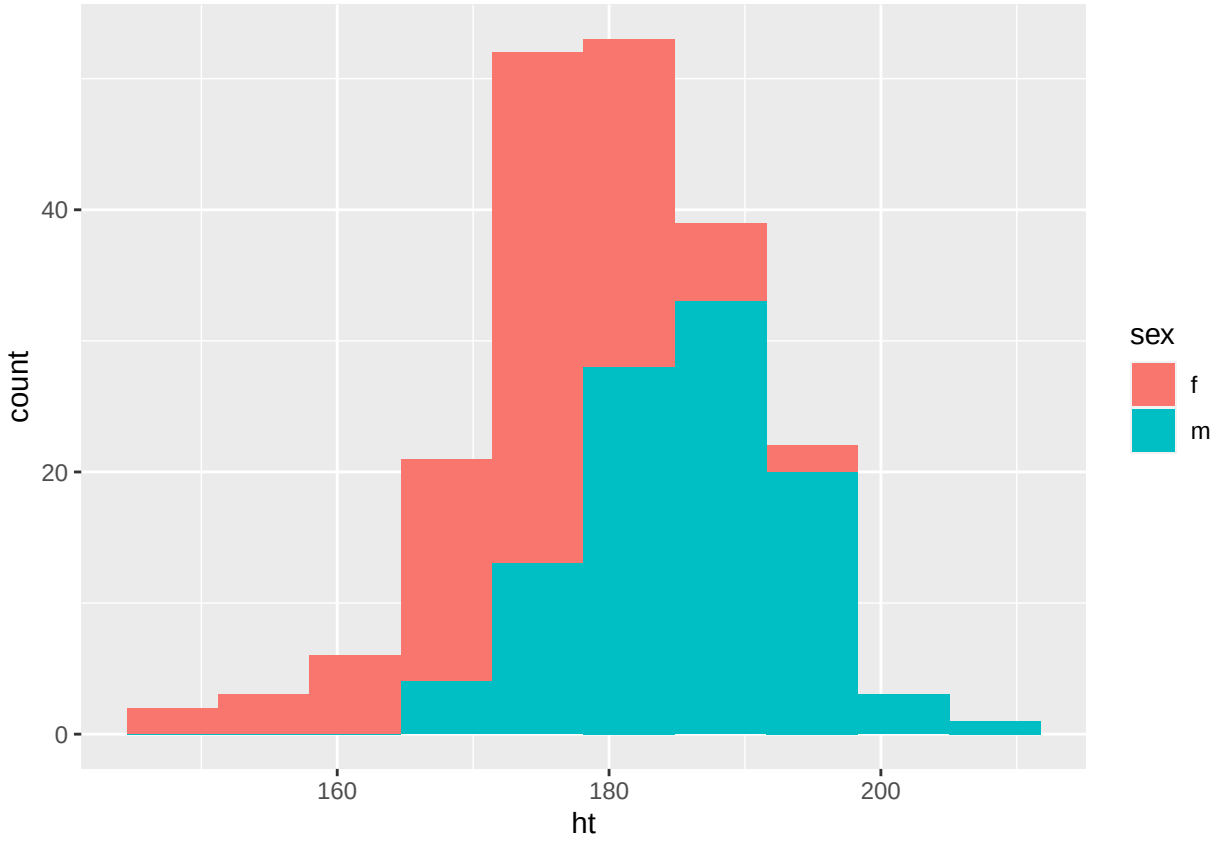
```
ggplot(data_3,aes(bmi,fill= wt+ht+(wt*ht)))+geom_histogram(bins=10)
```




```
ggplot(data_3,aes(wt,fill= sex))+geom_histogram(bins=10)
```




```
ggplot(data_3,aes(ht,fill= sex))+geom_histogram(bins=10)
```




```
d1 <- data_3%>%select(ht,wt)%>%dist()
library(vegan)
adonis(d1~sex,data=data_3)
```

```
##
## Call:
## adonis(formula = d1 ~ sex, data = data_3)
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##            Df SumsOfSqs MeanSqs F.Model     R2 Pr(>F)
## sex         1     17650 17650.4  87.438 0.3042  0.001 ***
## Residuals 200     40373   201.9         0.6958
## Total     201     58023                 1.0000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearMod <- lm(ht ~ wt, data=data_3)
r1<-lm(bmi~ht, data=data_3)$residuals
r2<-lm(bmi~wt,data=data_3)$residuals
coef(lm(r1~r2))
```

```
##   (Intercept)            r2
```

```
## -3.382826e-17  1.383471e+00
```

```
coef(lm(bmi~ht*wt,data=data_3))
```

```
##  (Intercept)           ht           wt        ht:wt
## 19.175085998 -0.106036730  0.653526612 -0.001920094
```

```
linearMod
```

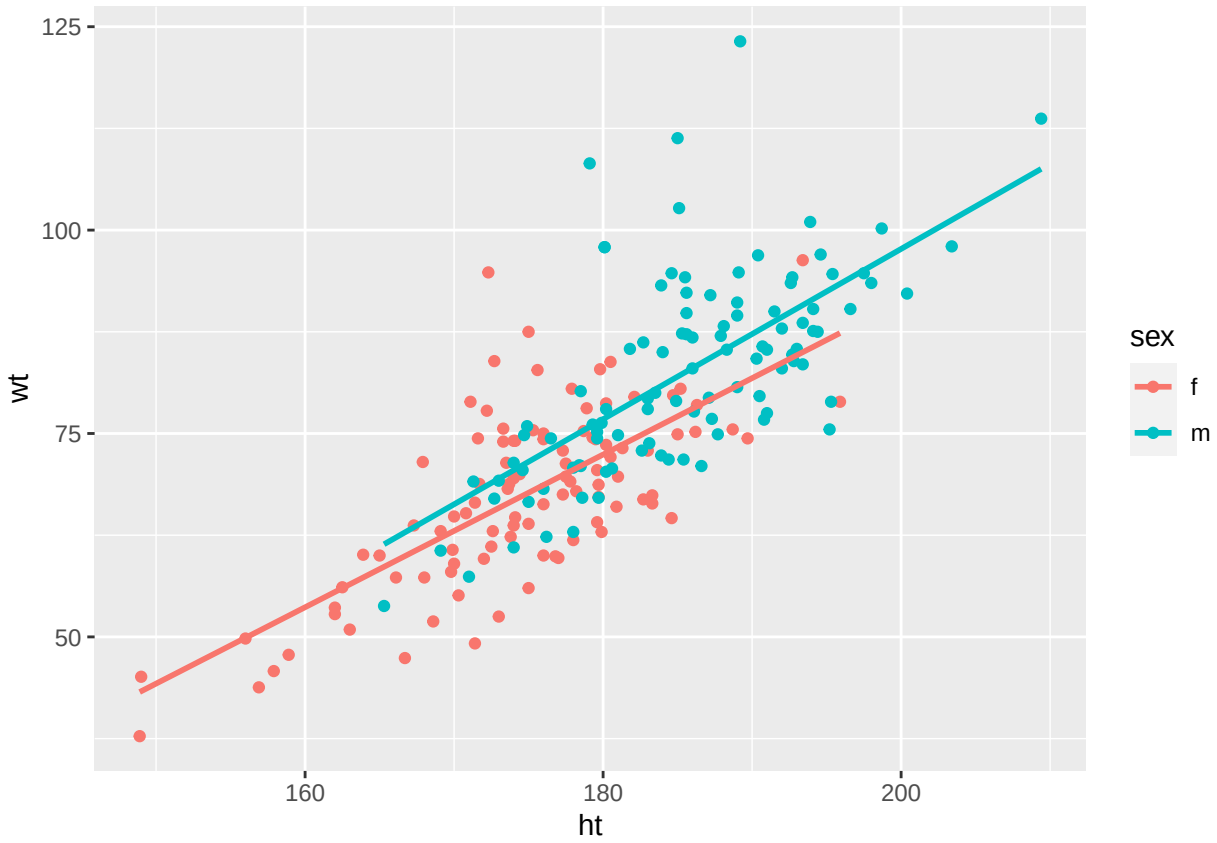
```
##
## Call:
## lm(formula = ht ~ wt, data = data_3)
##
## Coefficients:
## (Intercept)           wt
##    139.1560       0.5459
```

```
summary(linearMod)
```

```
##
## Call:
## lm(formula = ht ~ wt, data = data_3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.1240  -3.9482   0.3684   4.5999  14.8274
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 139.15599    2.35526   59.08   <2e-16 ***
## wt            0.54592    0.03088   17.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.096 on 200 degrees of freedom
## Multiple R-squared:  0.6099, Adjusted R-squared:  0.6079
## F-statistic: 312.6 on 1 and 200 DF,  p-value: < 2.2e-16
```

```
ggplot(data=data_3,aes(x=ht,y=wt,color=sex))+geom_point()+geom_smooth(method="lm",se=FALSE)
```

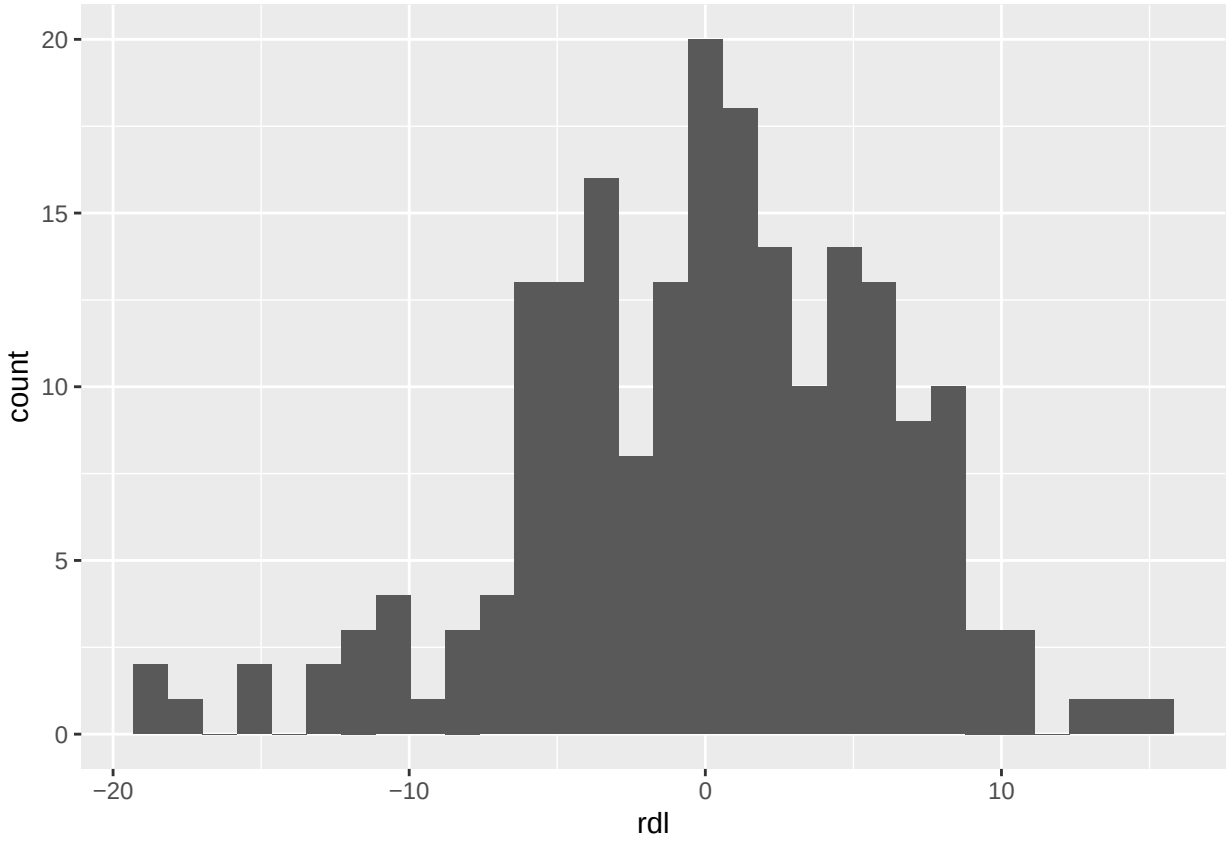
```
## `geom_smooth()` using formula 'y ~ x'
```



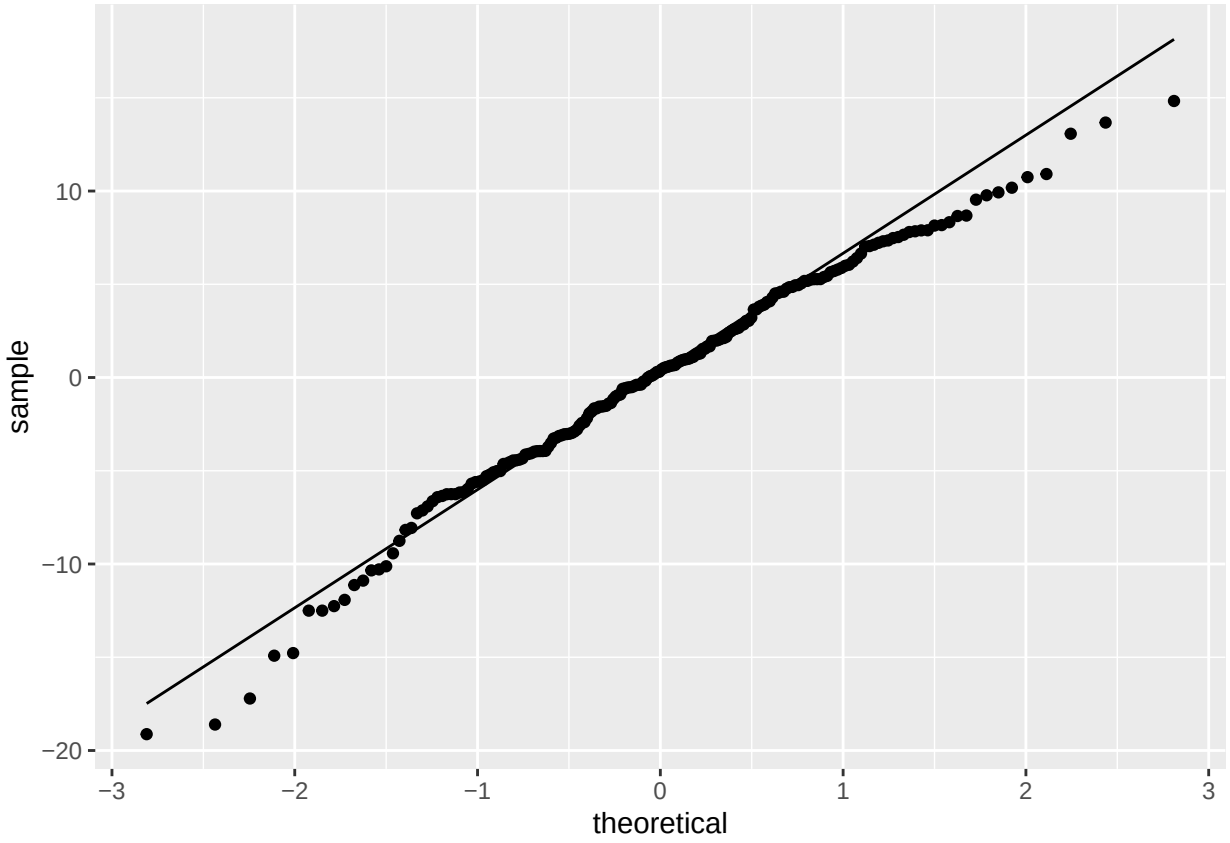

```
library(lmtest)
library(sandwich)
bptest(linearMod)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  linearMod
## BP = 16.97, df = 1, p-value = 3.797e-05
```

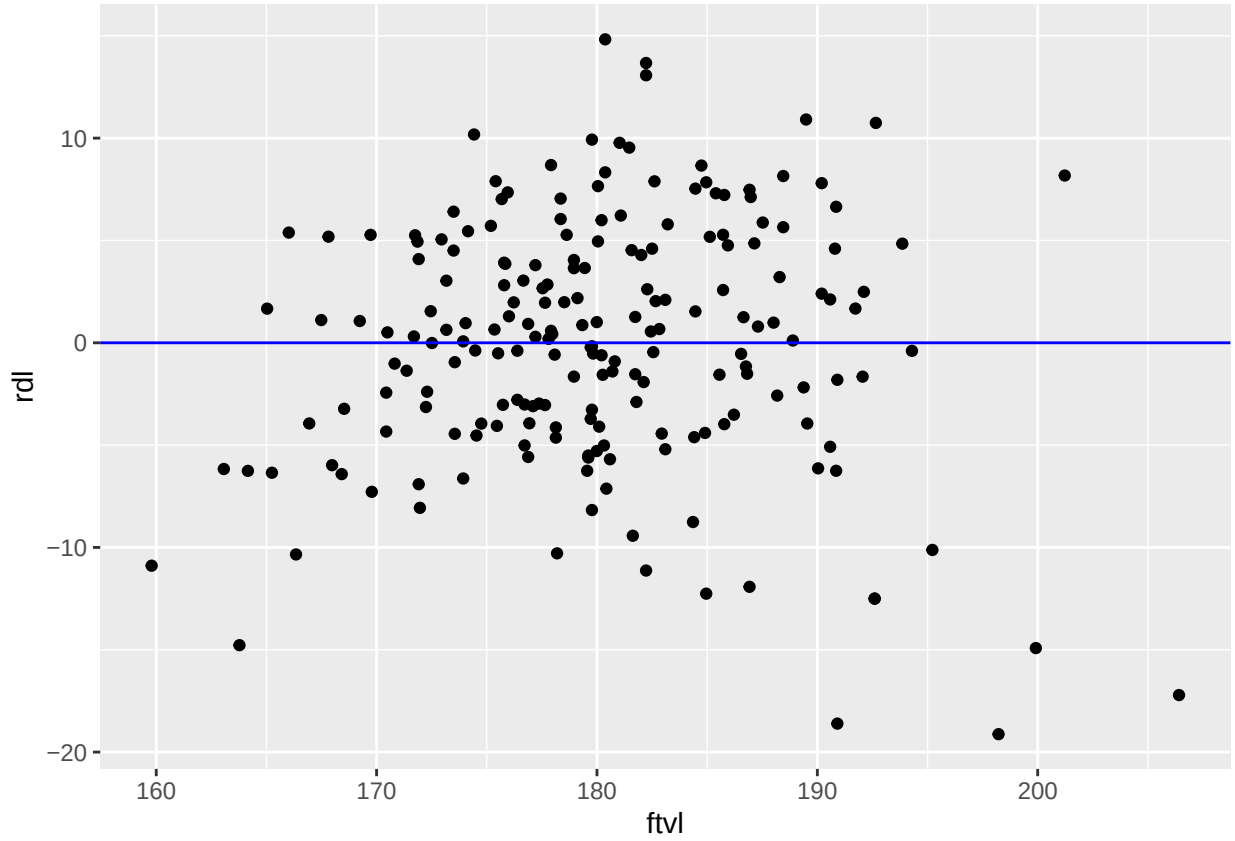
```
ftvl <-linearMod$fitted.values
rdl <-linearMod$residuals
ggplot()+geom_histogram(aes(rdl),bins = 30)
```




```
ggplot() +geom_qq(aes(sample=rdl)) + geom_qq_line(aes(sample=rdl))
```




```
ggplot()+geom_point(aes(ftvl,rdl))+geom_hline(yintercept = 0, color='blue')
```




```
summary(linearMod)$coef[,1:2]
```

```
##                Estimate Std. Error
## (Intercept) 139.1559917 2.35526315
## wt            0.5459153 0.03087522
```

The null hypothesis is that there is not a significant difference between the two genders for height and weight, and the alternate hypothesis is that there is a significant difference between the two sexes for height and weight. I ran a permanova test to analyze the multivariate data, and with a p value of 0.001, the null hypothesis is rejected.

#3

```
library(lmtest)
library(sandwich)
data_3$ht<-data_3$ht-mean(data_3$ht)
data_3$wt<-data_3$wt-mean(data_3$wt)
newfit <-lm(bmi ~ ht*wt,data = data_3)
summary(newfit)
```

```
##
## Call:
## lm(formula = bmi ~ ht * wt, data = data_3)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.45110 -0.05219 -0.02294  0.04805  0.43361
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.316e+01  9.219e-03 2512.12   <2e-16 ***
## ht          -2.501e-01  1.319e-03 -189.62   <2e-16 ***
## wt           3.077e-01  9.167e-04  335.66   <2e-16 ***
## ht:wt       -1.920e-03  4.425e-05  -43.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.113 on 198 degrees of freedom
## Multiple R-squared:  0.9985, Adjusted R-squared:  0.9984
## F-statistic: 4.294e+04 on 3 and 198 DF,  p-value: < 2.2e-16
```

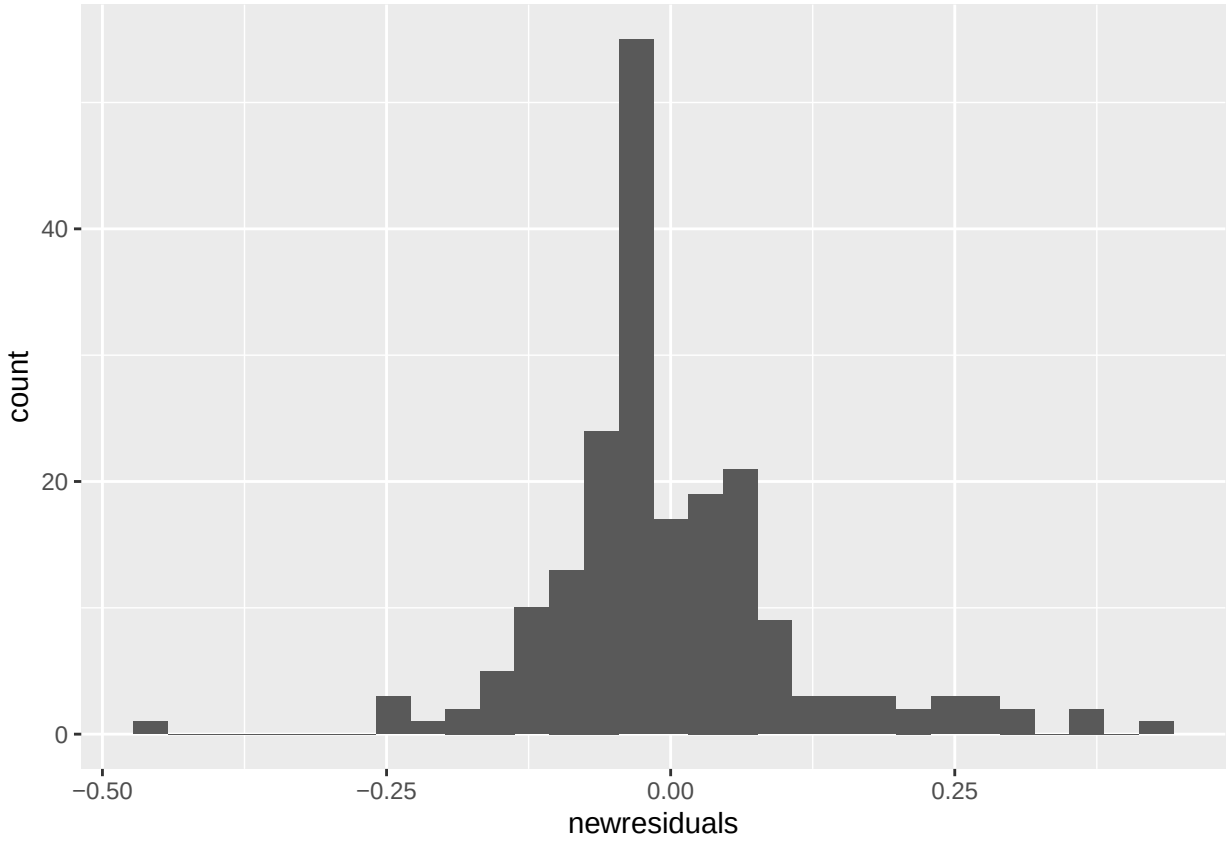
```
bptest(newfit)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  newfit
## BP = 57.96, df = 3, p-value = 1.603e-12
```

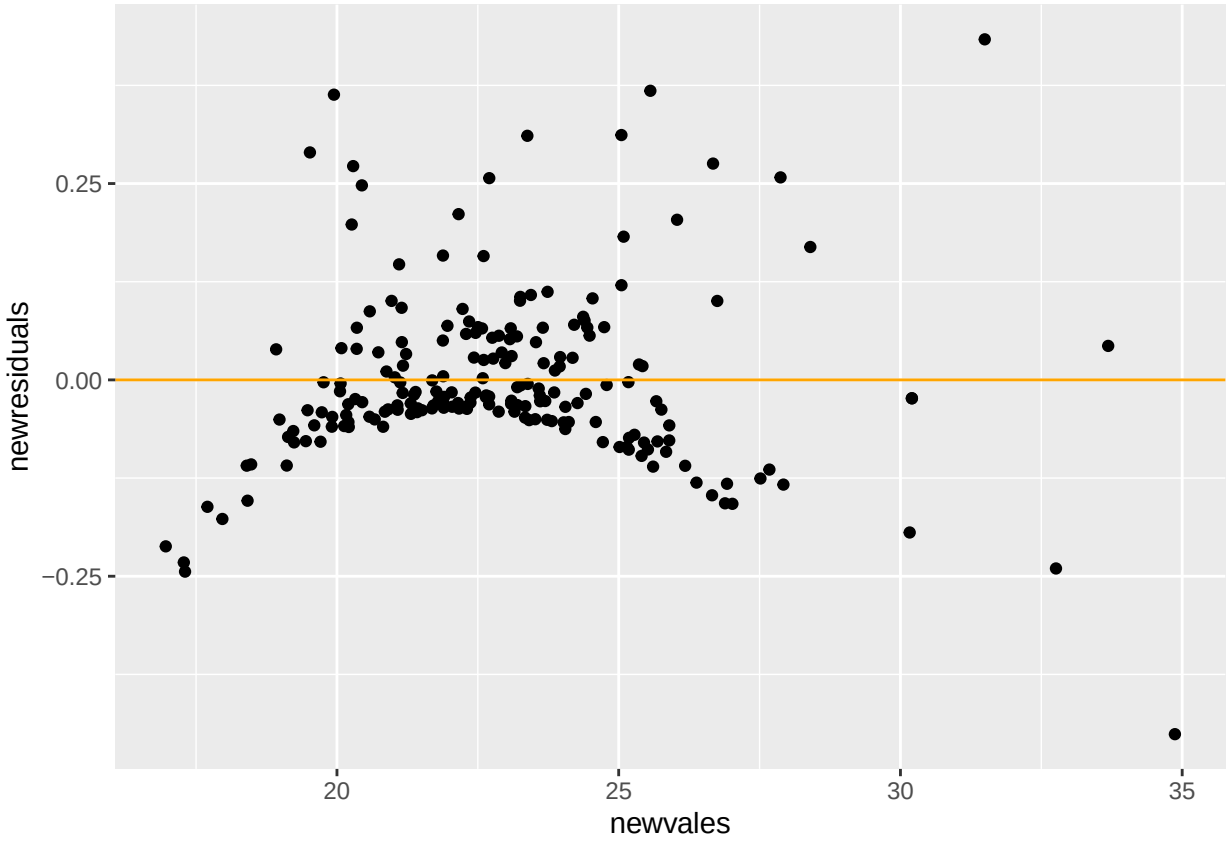
```
newvales <-newfit$fitted.values
newresiduals <- newfit$residuals
summary(linearMod)$coef[,1:2]
```

```
##                Estimate Std. Error
## (Intercept) 139.1559917 2.35526315
## wt            0.5459153 0.03087522
```

```
ggplot()+geom_histogram(aes(newresiduals),bins = 30)
```




```
ggplot()+geom_point(aes(newvales,newresiduals))+geom_hline(yintercept = 0, color='orange')
```




```
coeftest(newfit,vcov =vcovHC(newfit))[,1:2]
```

```
##                 Estimate   Std. Error
## (Intercept) 23.158144225 0.0113314757
## ht          -0.250058988 0.0026453640
## wt           0.307710079 0.0019213430
## ht:wt       -0.001920094 0.0001185701
```

```
summary(newfit)
```

```
##
## Call:
## lm(formula = bmi ~ ht * wt, data = data_3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45110 -0.05219 -0.02294  0.04805  0.43361
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.316e+01  9.219e-03 2512.12   <2e-16 ***
## ht          -2.501e-01  1.319e-03 -189.62   <2e-16 ***
## wt           3.077e-01  9.167e-04  335.66   <2e-16 ***
## ht:wt       -1.920e-03  4.425e-05  -43.39   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.113 on 198 degrees of freedom
## Multiple R-squared:  0.9985, Adjusted R-squared:  0.9984
## F-statistic: 4.294e+04 on 3 and 198 DF,  p-value: < 2.2e-16
```

```
summary(newfit)
```

```
##
## Call:
## lm(formula = bmi ~ ht * wt, data = data_3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45110 -0.05219 -0.02294  0.04805  0.43361
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.316e+01  9.219e-03 2512.12   <2e-16 ***
## ht          -2.501e-01  1.319e-03 -189.62   <2e-16 ***
## wt           3.077e-01  9.167e-04  335.66   <2e-16 ***
## ht:wt       -1.920e-03  4.425e-05  -43.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.113 on 198 degrees of freedom
## Multiple R-squared:  0.9985, Adjusted R-squared:  0.9984
## F-statistic: 4.294e+04 on 3 and 198 DF,  p-value: < 2.2e-16
```

After mean centering the numeric variables and finding the coefficients, the data came out with a y intercept of 2.316e+01. This means that when the weight and height of an athlete are 0, the expected bmi is 2.316e+01. For every unit increase in weight, the bmi also increases by 3.077e-01 units. When the height increases by a unit, the bmi actually decreases by 2.501e-01 units. This could be because of associations and not necessarily causations. Additionally, it was also found that the bmi of an athlete decreases by 1.920e-03 units for the interaction between height and weight. Pertaining to the assumptions of linearity, normality, and homoskedasticity, linearity and homoskedasticity are not met due to the variance in the graphs above. Normality looks to be met due to the bell curve shape of the graph. The Bptest function further supports that homoskedasticity is not met. The "coeftest" function was run to find standard errors, and the error is larger with this function. The errors are (int= 0.0113, ht= 0.003, wt=0.002, and ht*wt=0.000119). Height explains 0.250 variation, weight explains 0.308 of the variation, and their interaction only explains 0.00192 of the variation.

#4

```
sample1 <-replicate(5000, {
  btstrp <-data_3[sample(nrow(data_3),replace = TRUE),]
  fitline <-lm(bmi ~ ht*wt, data=btstrp)
  coef(fitline)
})
sample1 %>%t%>%as.data.frame() %>% summarize_all(sd)
```

```
##   (Intercept)          ht         wt        ht:wt
## 1  0.01009274 0.002413993 0.00172919 9.955491e-05
```

The bootstrapped standard errors are (intercept=0.01010906, ht=0.002410436, wt=0.00174591, ht*wt=9.92128e-05). They are similar to the standard errors found previously, but the standard errors are lower for the interaction between height and weight but greater for the intercept, the height, and the weight compared to the previous model.

#5

```
diags<-NULL
data_3$sex1=ifelse(data_3$sex=='female', 1, 0)
logitMod <- glm(sex1 ~ ht + wt, data=data_3, family=binomial(link="logit"))
```

```
## Warning: glm.fit: algorithm did not converge
```

```
summary(logitMod)
```

```
##
## Call:
## glm(formula = sex1 ~ ht + wt, family = binomial(link = "logit"),
##     data = data_3)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -2.409e-06  -2.409e-06  -2.409e-06  -2.409e-06  -2.409e-06
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.657e+01  2.506e+04  -0.001    0.999
## ht           1.833e-15  4.131e+03   0.000    1.000
## wt          -1.052e-15  2.888e+03   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 0.0000e+00  on 201  degrees of freedom
## Residual deviance: 1.1719e-09  on 199  degrees of freedom
## AIC: 6
##
## Number of Fisher Scoring iterations: 25
```

```
coeftest(logitMod)
```

```
##
## z test of coefficients:
##
##                Estimate  Std. Error z value Pr(>|z|)
## (Intercept) -2.6566e+01  2.5057e+04 -0.0011   0.9992
## ht           1.8329e-15  4.1312e+03  0.0000   1.0000
## wt          -1.0518e-15  2.8879e+03  0.0000   1.0000
```

```
exp(coef(logitMod))
```

```
##  (Intercept)           ht           wt
## 2.900701e-12 1.000000e+00 1.000000e+00
```

```
guess1 <-predict(logitMod,data=data_3,type="response")
data_3$predict <-predict(logitMod,data=data_3,type="response")
table(predict=as.numeric(data_3$predict>.5),truth=data_3$sex)%>%addmargins
```

```
##        truth
## predict   f   m Sum
##     0   100 102 202
##     Sum 100 102 202
```

```
#confusion matrix above
#sensitivity
76/102
```

```
## [1] 0.745098
```

```
#accuracy
(81+76)/202
```

```
## [1] 0.7772277
```

```
#specifity
81/107
```

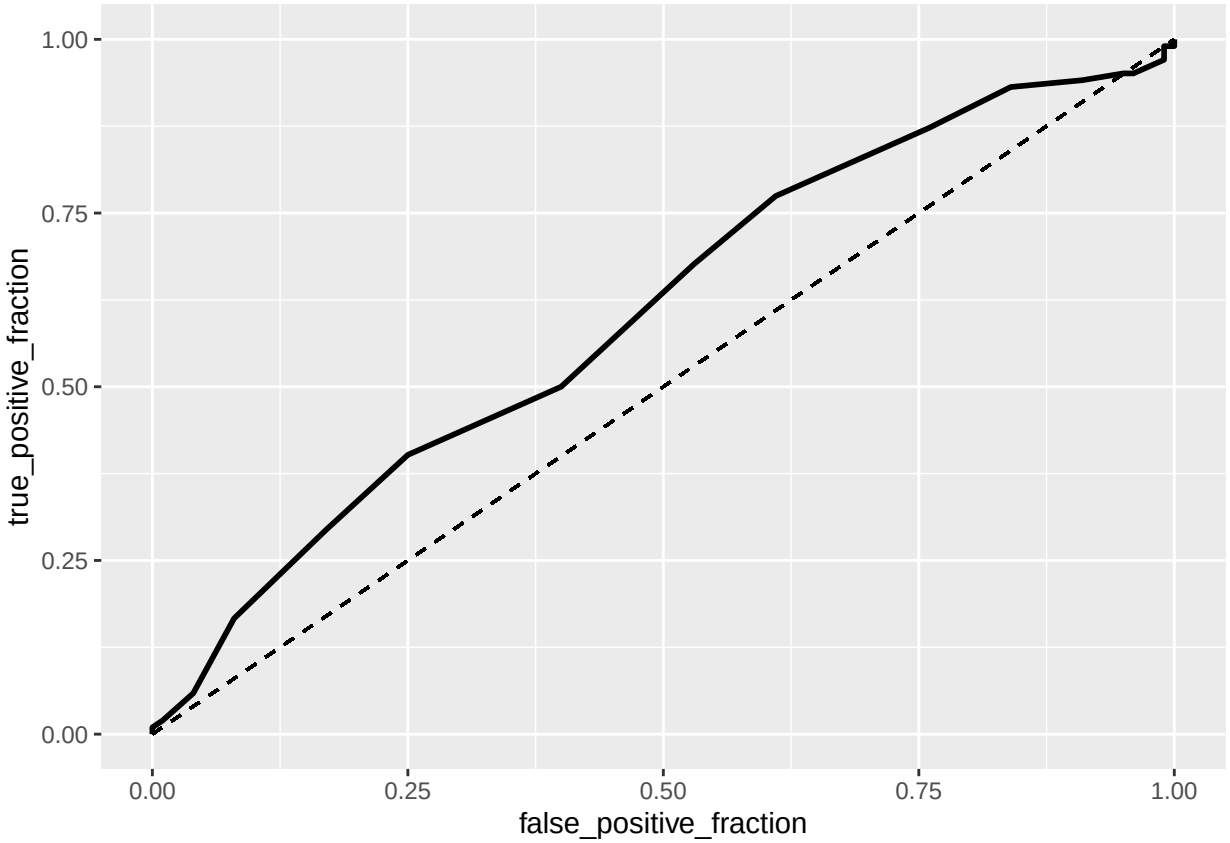
```
## [1] 0.7570093
```

```
#PPV
76/95
```

```
## [1] 0.8
```

```
library(plotROC)
library(ROCR)
ROCcurves<- ggplot(data_3) + geom_roc(aes(d=bmi,m=predict),n.cuts = 0)
data_4 <- data_3 %>% mutate(prob=predict(logitMod,type="response"), prediction = ifelse(prob>.5,1,0))
class <- data_4 %>% transmute(prob,predict,truth=sex)
ROC <-ggplot(class)+geom_roc(aes(d=truth,m=prob),n.cuts = 0)+geom_segment(aes(x=0,xend=1,y=0,yend=1),lty
ROC
```

```
## Warning in verify_d(data$d): D not labeled 0/1, assuming f = 0 and m = 1!
```




```
calc_auc(ROC)
```

```
## Warning in verify_d(data$d): D not labeled 0/1, assuming f = 0 and m = 1!
```

```
##   PANEL group       AUC
## 1     1    -1 0.6038725
```

The binary categorical variable is gender or "sex". For the coefficient estimates, for every unit increase in height, the probability of an Australian athlete being male increases by 0.12276095. For every increase in weight, the probability of an athlete being male increases by 0.06215562. The sensitivity from the model was 0.745098, the accuracy is 0.7772277, and the specificity is 0.7570093. Specificity is the proportion of the negatives that are correctly identified, sensitivity is the proportion of positives that are correctly identified, and accuracy is the proportion of actual elements that are the same as the estimated ones. The PPV gives the percent of estimated "male" counts that are actually male, and this value is 0.8. The AUC value was 0.850098. An ROC curve was generated, and after a repeated random sub sampling CV was unable to be performed.

#6

I had issues on my computer and couldn't get it fixed due to shops being closed.

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.