Project 2 (Modeling)

Jaqueline Ma

Introduction

For this project, I chose the 'studentdata' dataset from the LearnBayes package in Rstudio. In this dataset, the variables consisted of Student (which is the student number), Height, Gender, Shoes (the number of pairs of shoes owned), Number (number chosen between 1 and 10), Dvds (number of movie dvds owned), ToSleep (time the person went to sleep the previous night -hours past midnight), WakeUp (time the person woke up the next morning), Haircut (cost of last haircut including tip), Job (number of hours working on a job per week), and Drink (usual drink at suppertime among milk, water and pop). There are 657 observations and the variable of gender was dummy coded under a new variable called y. The NAs and the Student variable were also omitted to allow for easier data analysis, leading to a total of 559 observations.

```
library(LearnBayes)
head(studentdata)
```

```
## Student Height Gender Shoes Number Dvds ToSleep WakeUp
Haircut Job Drink
## 1 1 67 female 10 5 10 -2.5 5.5 60 30.0 water
## 2 2 64 female 20 7 5 1.5 8.0 0 20.0 pop
## 3 3 61 female 12 2 6 -1.5 7.5 48 0.0 milk
## 4 4 61 female 3 6 40 2.0 8.5 10 0.0 water
## 5 5 70 male 4 5 6 0.0 9.0 15 17.5 pop
## 6 6 63 female NA 3 5 1.0 8.5 25 0.0 water
```

```
students <- studentdata%>%mutate(y = ifelse(Gender == "male", 1, 0))%>%na.omit
students$Student <- NULL
students%>%count
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   559
```

MANOVA Test

For this data, a MANOVA test was conducted to see if any of the response variables differ by levels of a categorical explanatory variable. The null hypothesis would be that the means of the response variables are equal whereas the alternative hypothesis would be that the means of the response variables are significantly different. I tested the response variables of Height, and Shoes against the variable of Drink (null = for the DVs of Height, and Shoes, means for each Drink are equal; alternative = for at least one DV, at least one Drink mean is different).

After running the MANOVA, I can reject the null hypothesis and conclude that there is a mean difference among the three levels of Drink for at least one of the dependent variables, *Pillai trace* = 0.024356, *psuedo F* = 3.4273, $p = 0.008546$

Since the overall MANOVA was significant, I performed univariate ANOVAs for each variable as follow-up tests. The result of these ANOVAs was that only the univariate ANOVA for Shoes was significant; for Shoes, at least one Drink differs (the other variable was not significant), $F = 5.4281$, $p = 0.004627$.

Post hoc analysis was performed conducting pairwise comparisons to determine which Drink differed by Shoes. Only the Drinks of milk and water were found to differ significantly from each other after adjusting for multiple comparisons (bonferroni alpha = 0.05/10 = 0.005)

There were a total of 6 tests performed (1 MANOVA, 2 ANOVAS and 3 post hoc/t-tests). Based on the number of tests, the probability of at least 1 type I error would be 0.2649081 or 26.5%. The significance level would then be adjusted to 0.008333333.

Some assumptions of this test include: random sample, multivariate normality of dependent variables, homogeneity of within group covariance matrices and linear relationships among dependent variables. The multivariate normality was estimated by making multivariate plots of response variables for each group. Examination of these density plots revealed a departure from multivariate normality. Random sample might also be violated since it is not explicit on where this student data came from. In addition, examination of covariance matrices for each group revealed differenes and a lack of homogeneity. It is possible that MANOVA would not be an appropriate analysis technique.

```
man1 <- manova(cbind(Height, Shoes)~Drink, data = students)
summary(man1)
```

```
## Df Pillai approx F num Df den Df Pr(>F)
## Drink 2 0.024356 3.4273 4 1112 0.008546 **
## Residuals 556
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

```
summary.aov(man1)
```

```
## Response Height :
## Df Sum Sq Mean Sq F value Pr(>F)
## Drink 2 45.7 22.865 1.2943 0.2749
## Residuals 556 9822.0 17.665
##
## Response Shoes :
## Df Sum Sq Mean Sq F value Pr(>F)
## Drink 2 2065 1032.57 5.4281 0.004627 **
## Residuals 556 105767 190.23
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

```
students%>%group_by(Drink)%>%summarize(mean(Height),mean(Shoes))
```

```
## # A tibble: 3 x 3
##   Drink `mean(Height)` `mean(Shoes)`
##   <fct>          <dbl>         <dbl>
## 1 milk            66.5          12.5
## 2 pop             67.2          14.2
## 3 water           66.6          17.3
```

```
pairwise.t.test(students$Shoes, students$Drink, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  students$Shoes and students$Drink
##
##       milk   pop
```

```
## pop   0.3562 -
## water 0.0034 0.0244
##
## P value adjustment method: none
```

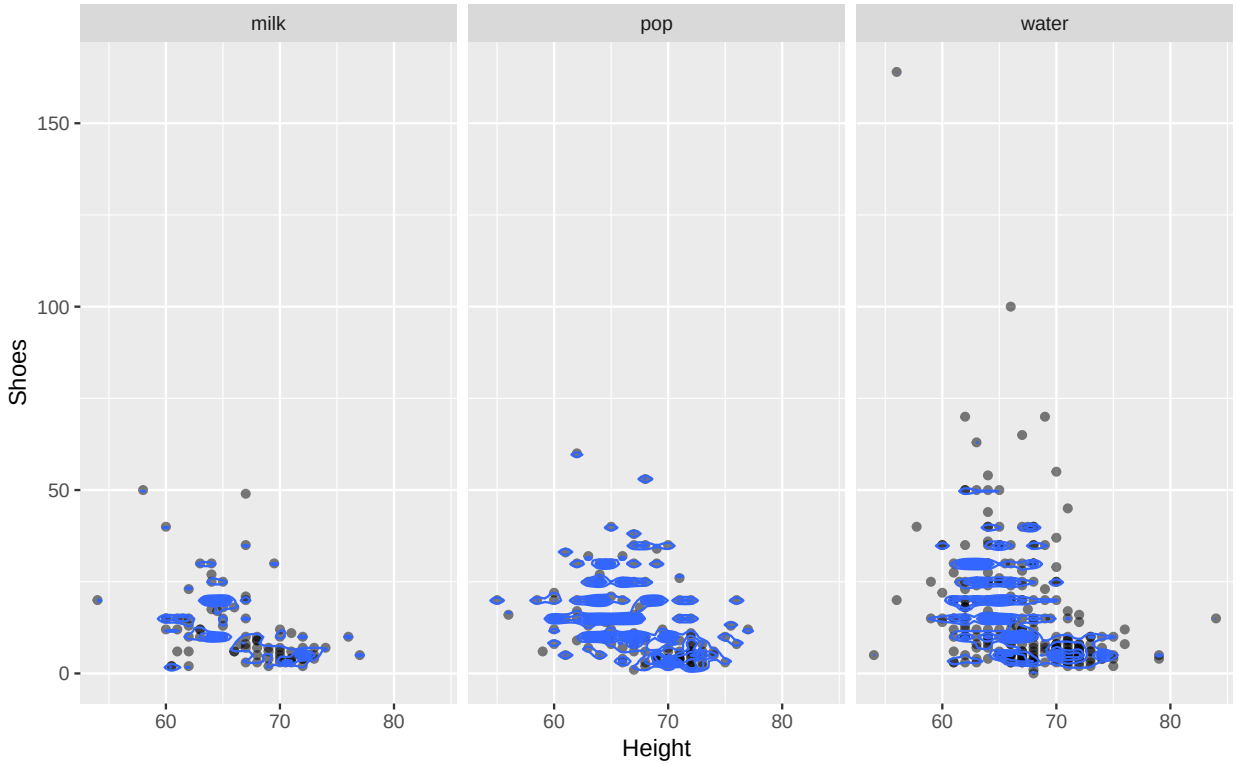
```
#probability
1-0.95^6
```

```
## [1] 0.2649081
```

```
#bonferroni alpha adjusted
0.05/6
```

```
## [1] 0.008333333
```

```
library(mvtnorm)
library(ggExtra)
ggplot(students, aes(x = Height, y = Shoes))+geom_point(alpha = 0.5)+
  geom_density_2d(h=2)+facet_wrap(~Drink)
```




```
covmats<- students%>%group_by(Drink)%>%do(covs = cov(.[c(1,3)]))
for(i in 1:3){print(as.character(covmats$Drink[i]));print(covmats$covs[i])}
```

```
## [1] "milk"
## [[1]]
##            Height     Shoes
## Height  18.35809 -18.70580
## Shoes  -18.70580  93.78028
##
## [1] "pop"
## [[1]]
##            Height     Shoes
```

```
## Height  17.85651 -15.29649
## Shoes  -15.29649 106.12359
##
## [1] "water"
## [[1]]
##           Height     Shoes
## Height  17.35355 -24.48873
## Shoes  -24.48873 262.30374
```

Randomization Test

A randomization test was then performed between the variables of Gender and Job, to see if there was a difference in the number of hours working on a job per week between males and females. Assumptions for the independent t-test were violated. The null hypothesis would be: the mean number of hours working on a job per week is the same for males and females while the alternative hypothesis would be: the mean number of hours working on a job per week is different for males and females.

The actual mean differences between the groups was first found and calculated to be 0.6631868. Then, the job hours were randomly scrambled and the mean differences between the gender groups was found from the randomly scrambled variables; this was done through a vector and for loop for 5,000 times. A histogram shows the distribution of the randomized values and the actual mean difference between the original data (as a red line).

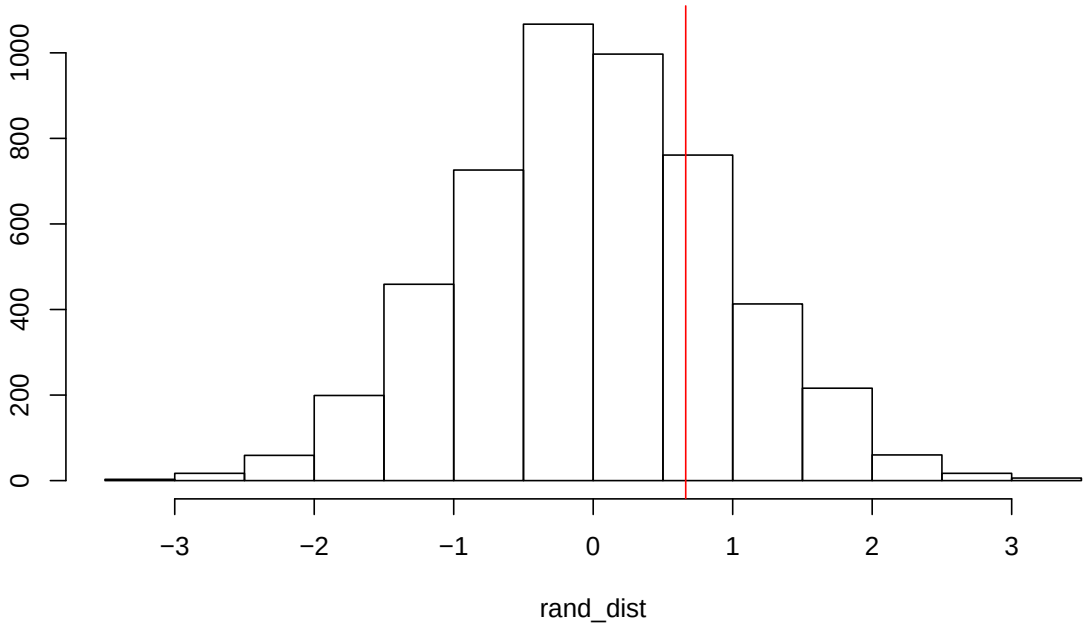
Lastly, the two tailed p-value was calculated to be 0.479, which would correspond to the probability of observing a mean difference as extreme as the one I observed in the original data under this randomized distribution. Therefore, I would fail to reject the null hypothesis and conclude that there is no significant difference between the mean number of hours working on a job per week between males and females.

```
students%>%group_by(Gender)%>%summarize(means=mean(Job))%>%
  summarize(`mean_diff:`=diff(means))
```

```
## # A tibble: 1 x 1
##   `mean_diff:`
##          <dbl>
## 1        0.663
```

```
rand_dist<-vector()
for(i in 1:5000){
new<-data.frame(gender = students$Gender,job = sample(students$Job))
rand_dist[i]<-mean(new[new$gender == "male",]$job)-mean(new[new$gender == "female",]$job)}

{hist(rand_dist, main = "",ylab =""); abline(v = 0.6631868, col="red")}
```




```
mean(rand_dist>0.6631868|rand_dist< -0.6631868)
```

```
## [1] 0.4812
```

Linear Regression Model

Next, a linear regression model was used to predict Height from the variables of Shoes (number of Shoes) and Gender, while also taking into consideration their interaction. The variable of Shoes was mean centered since it was a numeric variable. The null hypotheses would be that controlling for Shoes, Gender does not explain variation in Height and that controlling for Gender, Shoes does not explain variation in Height. In contrast, the alternative hypotheses would be that controlling for Shoes, Gender does explain variation in Height and that controlling for Gender, Shoes does explain variation in Height.

Based on these coefficient estimates, when people have the average number of Shoes, and are female, the average Height will be 64.94709. When controlling for Shoes, males will have a 5.41655 average increase in Height compared to females. When the person is a female, there is a decrease of 0.02413 in Height for every 1 unit increase in Shoes. Lastly, the slope for Shoes on Height is 0.02254 higher for males as compared to females.

This regression was then plotted. Assumptions of linearity, normality and homoskedasticity appear to be met after looking at the graphs and seeing relative linearity, normality and equal variances within the regression.

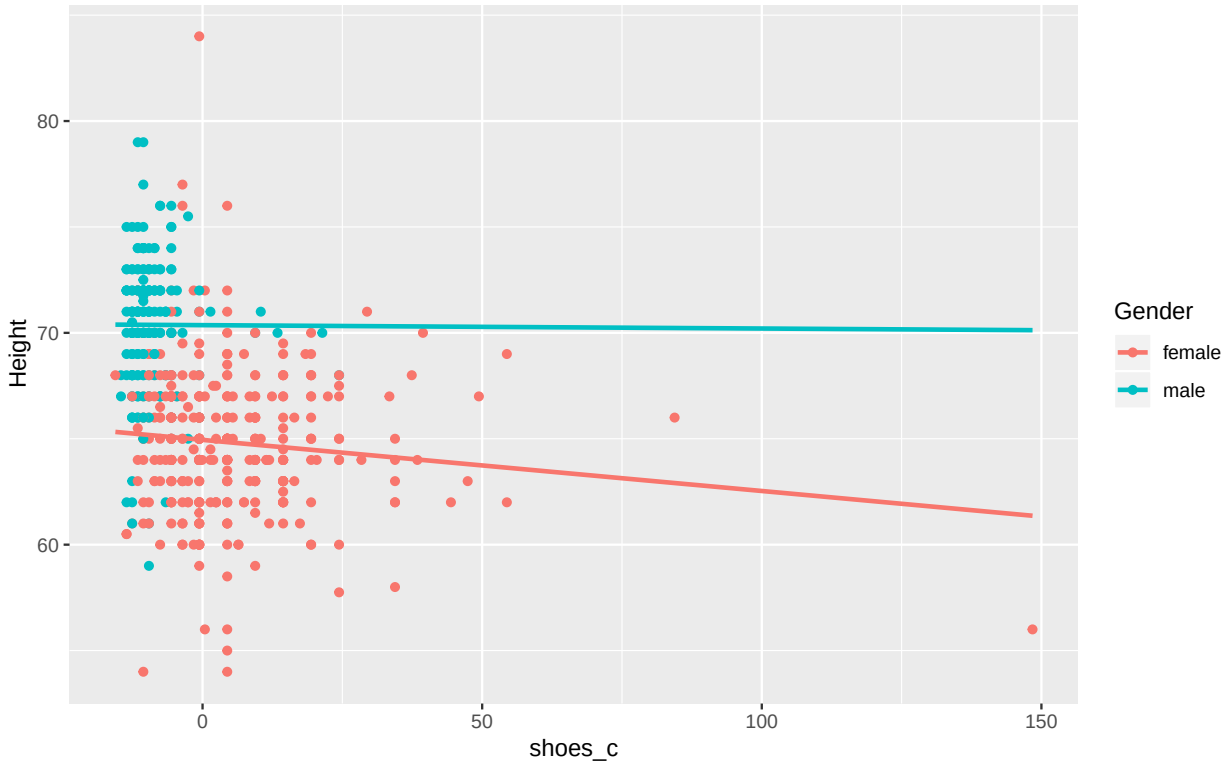
Despite meeting homoskedasticity, the regression was then recomputed with robust standard errors. The result was that there was barely any change to the coefficients but the standard errors did decrease slightly. This would probably be due to the fact that the variables were originally homoskedastic, since conducting coeftest(. . . ,vcov=vcovHC(. . .)) would normally lead to higher SEs.

The adjusted R^2 value from summary(fit) was found to be 0.3976, which is the proportion of variation in the response variable explained by the overall model. In other words, my model explains 39.76% variation in the outcome/response variable.

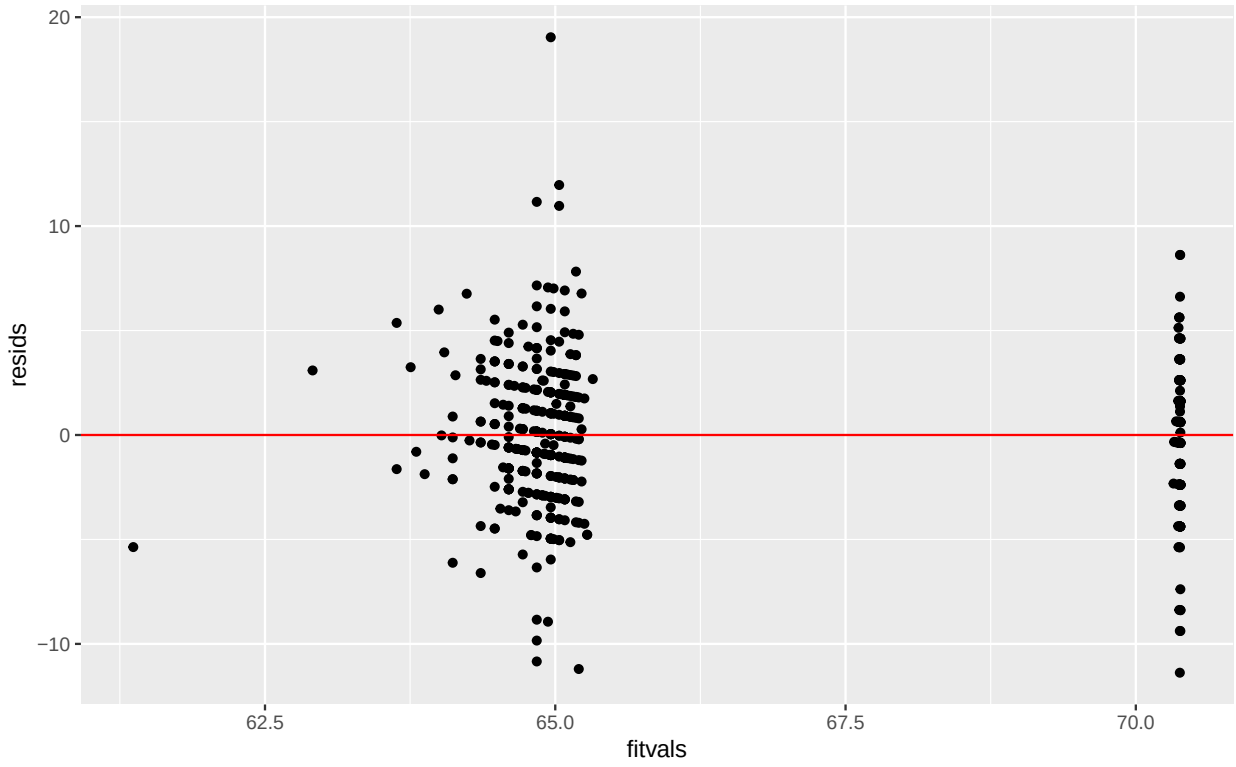
```r
students$shoes_c <- students$Shoes-mean(students$Shoes)
fit<- lm(Height~Gender*shoes_c, data = students)
summary(fit)
```

```
##
## Call:
## lm(formula = Height ~ Gender * shoes_c, data = students)
##
## Residuals:
## Min 1Q Median 3Q Max
## -11.3789 -1.9613 -0.0337 1.8819 19.0387
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 64.94709 0.18041 360.006 <2e-16 ***
## Gendermale 5.41655 0.51342 10.550 <2e-16 ***
## shoes_c -0.02413 0.01170 -2.062 0.0397 *
## Gendermale:shoes_c 0.02254 0.04744 0.475 0.6349
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 3.264 on 555 degrees of freedom
## Multiple R-squared: 0.4008, Adjusted R-squared: 0.3976
## F-statistic: 123.8 on 3 and 555 DF, p-value: < 2.2e-16
```

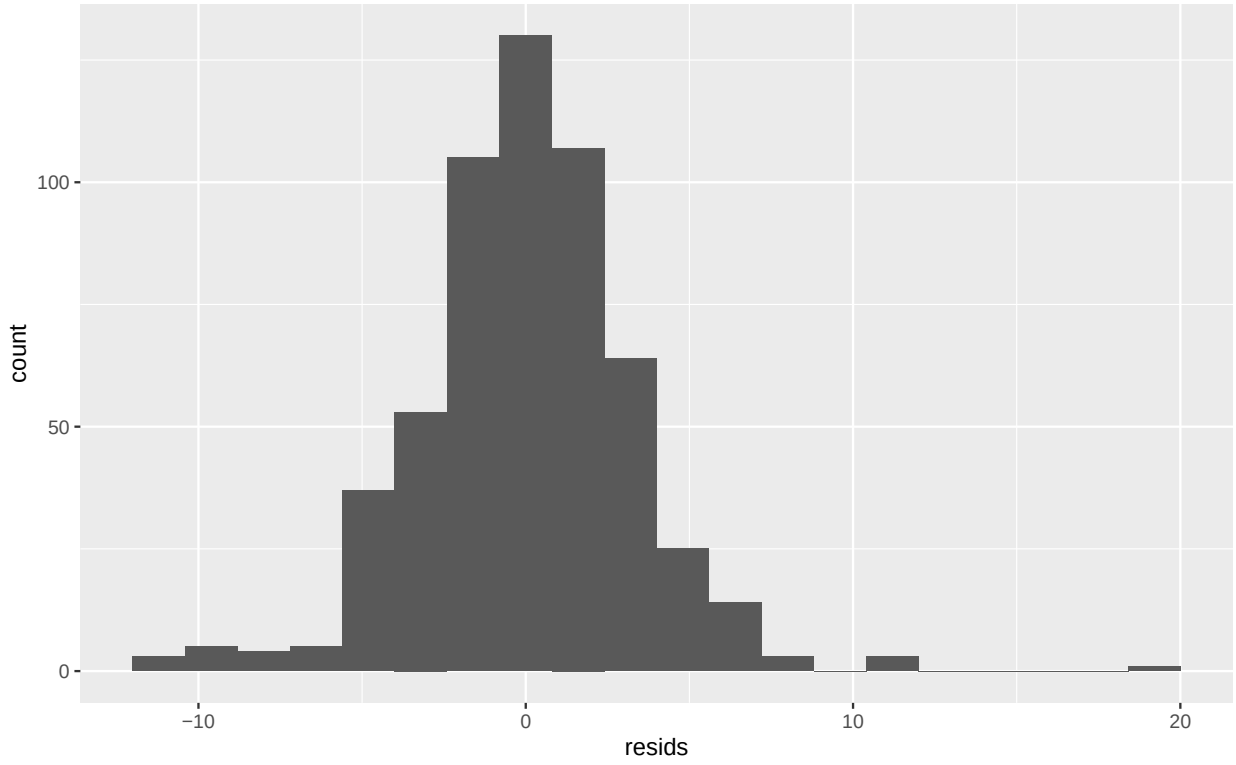
```r
qplot(x=shoes_c, y = Height, color = Gender, data = students)+
  stat_smooth(method = "lm", se = FALSE, fullrange = TRUE)
```



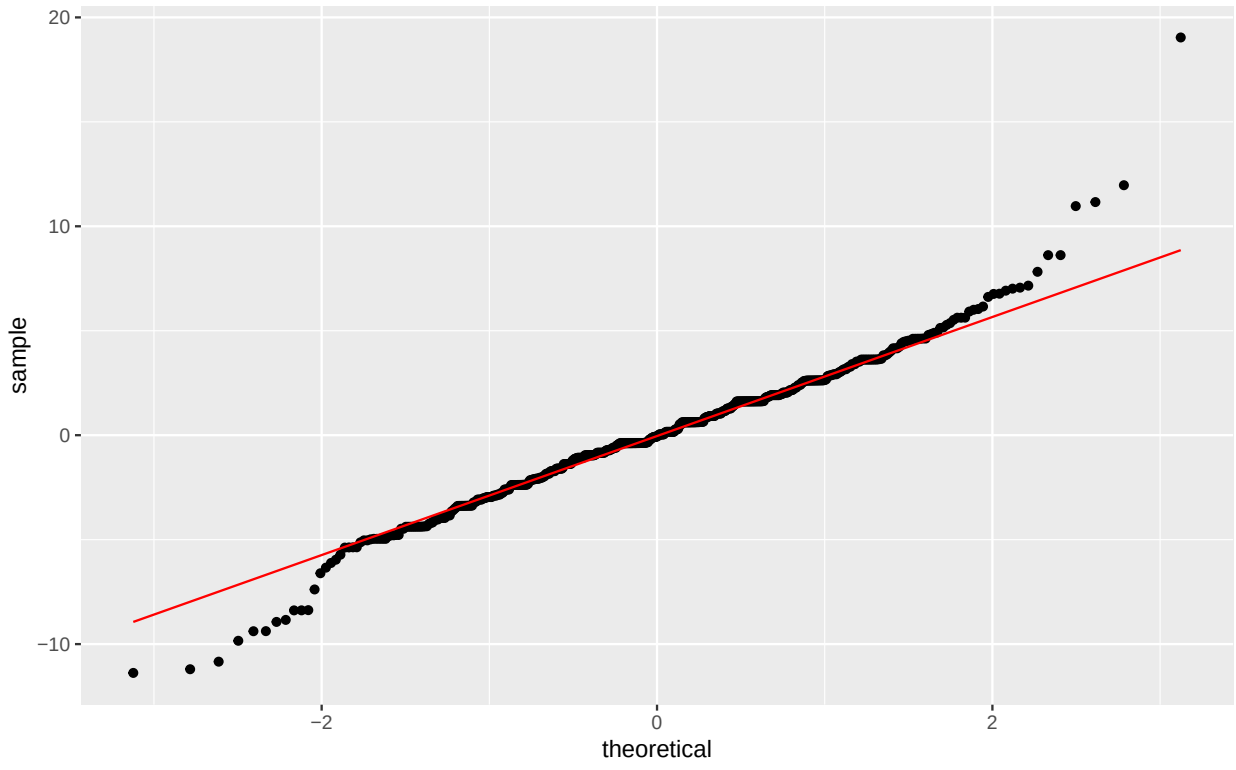
```r
#linearity and homoskedasticity
resids<- fit$residuals
fitvals <- fit$fitted.values
ggplot()+geom_point(aes(fitvals, resids))+geom_hline(yintercept = 0, color = "red")
```




```r
#normality graphs
ggplot()+geom_histogram(aes(resids),bins=20)
```




```
ggplot()+geom_qq(aes(sample=resids))+geom_qq_line(aes(sample=resids), color='red')
```




```
library(sandwich);library(lmtest)
coeftest(fit, vcov=vcovHC(fit))[,1:4]
```

```
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 64.94709075 0.19275962 336.9330695
0.000000e+00
## Gendermale 5.41655493 0.39887856 13.5794586 1.802270e-36
## shoes_c -0.02413067 0.01699139 -1.4201704 1.561196e-01
## Gendermale:shoes_c 0.02253710 0.03730836 0.6040765
5.460397e-01
```

Bootstrapped Standard Errors

In this section, I reran my same regression model from the previous section but am going to compute bootstrapped standard errors. Bootstrapping will allow me to randomly sample rows from my datset with replcement and calculate coefficient estimates on the bootstrapped sample.

The bootstrapped SEs are very similar to the robust SEs and the original SEs for this model. These SEs vary slightly above or slightly below the robust SEs (above for Gendermale and Gendermale:shoes_c). However, the bootstrapped SEs are above the original SEs for the Intercept and the shoes_c variables in contrast.

```
boot_dat<-students[sample(nrow(students),replace=TRUE),]
samp_distn<-replicate(5000, {
  boot_dat<-boot_dat<-students[sample(nrow(students),replace=TRUE),]
  fit<- lm(Height~Gender*shoes_c, data=boot_dat)
  coef(fit)
})
## Estimated SEs
samp_distn%>%t%>%as.data.frame%>%summarize_all(sd)
```

```
##   (Intercept) Gendermale    shoes_c Gendermale:shoes_c
## 1    0.186253  0.4117641 0.01423664         0.03765789
```

Logistic Regression

Next, a logistic regression was performed to try to predict y from the variables of Dvds (number of DVDs owned), WakeUp (time they woke up), and Number (a number that the person randomly chose, from 1-10). I wanted to see if there was any relationship between these variables and if the person was male or female.

Based on the logistic regression, the odds of being a male (y = 1), when Dvds is 0, WakeUp is 0, and Number is 0, is 0.07548153. When controlling for WakeUp and Number, for every 1 additional increase in number of Dvds, odds of being a male increase by a factor of 1.00213996. When controlling for Dvds and Number, for every 1 additional hour increase in WakeUp time, odds of being a male increase by a factor of 1.23823637. When controlling for Dvds and WakeUp, for every 1 additional increase in Number, odds of being a male increase by a factor of 1.01352330.

Next, a confusion matrix was created to compare the model predictions versus the true outcomes. From this table the accuracy, sensitivity, specificity and Precision was calculated. In this context, accuracy would be the proportion of correctly classified cases, sensitivity would be the proportion of males correctly classified, specificity would be the proportion of females correctly classified, and precision (or PPV) would be the proportion of classified males who actually are males. The accuracy was found to be 0.6601073, sensitivity was 0.06666667, specificity was 0.978022, and PPV was 0.6190476. Based on these values, the TNR was very high but the TPR, PPV and accuracy could all be improved.

Using ggplot, I then plotted the density of log-odds by the y variable. Based on this graph, the majority of the gray area is found on the left of 0, which corresponds with the proportion of false negatives, where the

value is actually male but was predicted to be female. For the part of the gray area found on the right of 0, this area is the proportion of females that was predicted as males (false positive).

From the model, I also created an ROC curve/plot to compare TPR to FPR. Based on this ROC plot and the calculation of AUC, which was found to be 0.591878, we can conclude that this model is a bad predictor overall. In other words, it is hard to predict Gender from the variables of Dvd, WakeUp and Number.

Lastly, I performed a 10-fold CV. After first running the class_diags function, I set the fold number to 10, created training and test sets, trained the model on the training set, trained the model on the test set and then conducted average diagnostics across all k folds. From this test, the average out-of-sample accuracy, sensitivity and precision (PPV) was 0.6600325, 0.06505952, and 0.65, respectively. In addition, the AUC was 0.5900491. Compared to the confusion matrix, these values are very similar and only differ very slightly.

```
fit2 <- glm(y~Dvds+WakeUp+Number, data = students, family = "binomial")
coeftest(fit2)
```

```
##
## z test of coefficients:
##
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.5838673 0.6074401 -4.2537 2.103e-05 ***
## Dvds 0.0021377 0.0014064 1.5200 0.1285169
## WakeUp 0.2136881 0.0632481 3.3786 0.0007286 ***
## Number 0.0134327 0.0413475 0.3249 0.7452772
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

```
exp(coef(fit2))
```

```
## (Intercept)        Dvds      WakeUp      Number
##  0.07548153  1.00213996  1.23823637  1.01352330
```

```
students$prob<-predict(fit2,type="response")
students$predicted<-ifelse(students$prob>.5,"male","female")
students$Gender<-factor(students$Gender,levels=c("female","male"))

table(truth=students$Gender, prediction=students$predicted)%>%addmargins
```

```
##         prediction
## truth    female male Sum
##   female    356    8 364
##   male      182   13 195
##   Sum       538   21 559
```

```
#accuracy
(356+13)/559
```

```
## [1] 0.6601073
```

```
#sensitivity (TPR)
13/195
```

```
## [1] 0.06666667
```

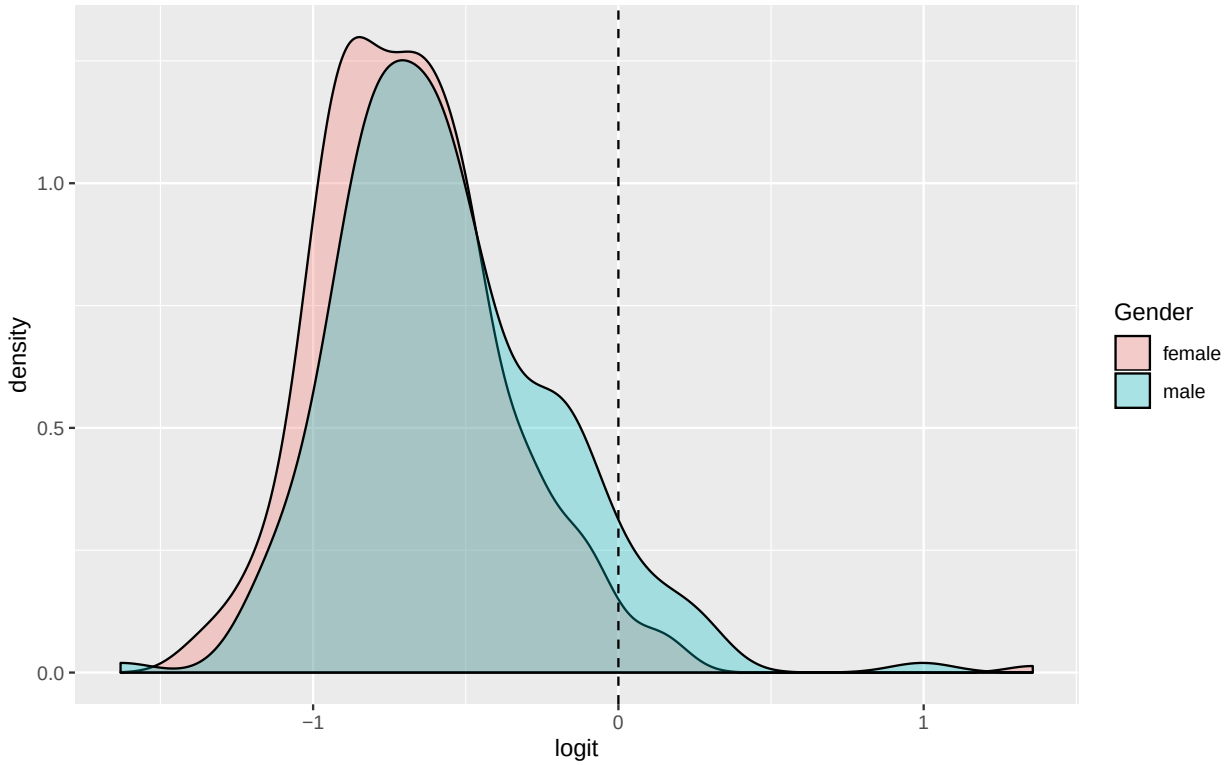
```
#specificity (TNR)
356/364
```

```
## [1] 0.978022
```

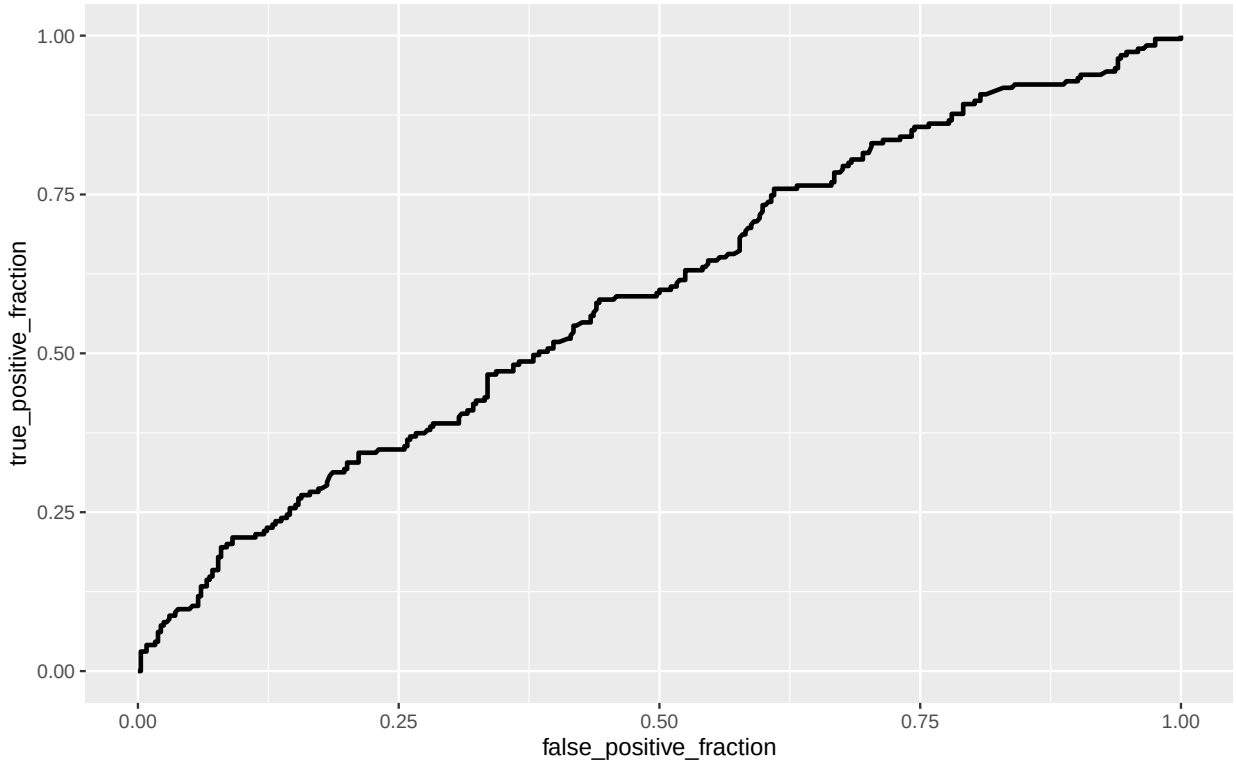
```
#precision (PPV)
13/21
```

```
## [1] 0.6190476
```

```
students$logit<-predict(fit2) #get predicted log-odds
ggplot(students,aes(logit, fill=Gender))+geom_density(alpha=.3)+
  geom_vline(xintercept=0,lty=2)
```




```
library(plotROC)
ROCplot <- ggplot(students)+geom_roc(aes(d = y, m = prob), n.cuts = 0)
ROCplot
```




```
calc_auc(ROCplot)
```

```
##   PANEL group      AUC
## 1     1    -1 0.591878
```

```
#K-fold CV
class_diag<-function(probs,truth){

  tab<-table(factor(probs>.5,levels=c("FALSE","TRUE")),truth)
  acc=sum(diag(tab))/sum(tab)
  sens=tab[2,2]/colSums(tab)[2]
  spec=tab[1,1]/colSums(tab)[1]
  ppv=tab[2,2]/rowSums(tab)[2]

  if(is.numeric(truth)==FALSE & is.logical(truth)==FALSE) truth<-as.numeric(truth)-1

  #CALCULATE EXACT AUC
  ord<-order(probs, decreasing=TRUE)
  probs <- probs[ord]
  truth <- truth[ord]

  TPR=cumsum(truth)/max(1,sum(truth))
  FPR=cumsum(!truth)/max(1,sum(!truth))

  dup<-c(probs[-1]>=probs[-length(probs)], FALSE)
  TPR<-c(0,TPR[!dup],1)
  FPR<-c(0,FPR[!dup],1)

  n <- length(TPR)
```

```
  auc<- sum( ((TPR[-1]+TPR[-n])/2) * (FPR[-1]-FPR[-n]) )

  data.frame(acc,sens,spec,ppv,auc)
}

set.seed(1234)
k=10

data<-students[sample(nrow(students)),]
folds<-cut(seq(1:nrow(students)),breaks=k,labels=F)

diags<-NULL
for(i in 1:k){
  ## Create training and test sets
  train<-data[folds!=i,]
  test<-data[folds==i,]
  truth<-test$y

  ## Train model on training set
  fit2 <- glm(y~Dvds+WakeUp+Number, data = students, family = "binomial")
  probs<-predict(fit2,newdata = test,type="response")

  ## Test model on test set (save all k results)
  diags<-rbind(diags,class_diag(probs,truth))
}
summarize_all(diags,mean)
```

```
##         acc       sens      spec  ppv       auc
## 1 0.6600325 0.06505952 0.9784034 0.65 0.5900491
```

LASSO Regression

For this last regression, I chose the variable of y to predict from the rest of my variables. After running this lasso regression, I found out that the variables of Height, Shoes, ToSleep, WakeUp and Haircut are all important predictors of y.

Next, I performed a 10-fold CV using this new model with only the important predictors. From this cross validation test, I got an accuracy value of 0.9158442, a sensitivity value of 0.8905952, a specificity value of 0.9349463, a PPV value of 0.8770614, and an AUC value of 0.9568835. Compared to the logistic regression completed in the previous section, this 10-fold CV was much more accurate, specific and an overall better predictor of the y variable. This is probably due to using the variables that are most important in predicting y, which was found through the lasso function.

```
students$Gender <- NULL
students$shoes_c <- NULL
students$prob <- NULL
students$predicted<- NULL
students$logit<- NULL

fit3<-glm(y~., data = students, family = "binomial")

set.seed(1234)
x<-model.matrix(fit3)
```

```
y<-as.matrix(students$y)

library(glmnet)
cv<-cv.glmnet(x,y, family = "binomial")
lasso<-glmnet(x,y,lambda=cv$lambda.1se)
coef(lasso)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept) -2.869381202
## (Intercept)  .
## Height       0.049784646
## Shoes       -0.007499315
## Number       .
## Dvds         .
## ToSleep      0.006561616
## WakeUp       0.007879163
## Haircut     -0.002358318
## Job          .
## Drinkpop     .
## Drinkwater   .
```

```
#10-fold CV
set.seed(1234)
data1 <- students %>% sample_frac
folds1 <- ntile(1:nrow(students),n=10)

diags1<-NULL
for(i in 1:k){
  train1<- data1[folds!=i,] #create training set (all but fold i)
  test1<- data1[folds==i,] #create test set (just fold i)
  truth1<- test1$y #save truth labels from fold i

  fit4 <- glm(y~Height+Shoes+ToSleep+WakeUp+Haircut, data=train1, family="binomial")
  probs1 <- predict(fit4, newdata=test1, type="response")

  diags1<-rbind(diags1,class_diag(probs1,truth1))
}

summarize_all(diags1,mean)
```

```
##         acc      sens      spec       ppv       auc
## 1 0.9158442 0.8905952 0.9349463 0.8770614 0.9568835
```

This concludes the end of my project 2.