Project 2: An Analysis of Child Measurements

Jena Georgopulos

4/15/2021

Introduction

This report will analyze a dataset containing the body measurements for a random sample of 198 children between the ages of 8 and 18 in the year of 1977. The data contains a total of 5 variables: height of the child (in inches), weight of the child (in pounds), age of the child (in months), sex of the child (male or female), and race of the child (white or other). The variables of height, weight, and age are all considered to be numeric. The variables of sex and race are considered to be categorical variables. These two categorical variables have been coded as dummy variables. For the variable sex, this translates to male = 0 and female = 1, with males being the reference category. For the variable race, this translates to white = 0, other = 1, with white being the reference category. after uploading the data into R-studio, a single column titled “X”, which gave the observation number, was removed from the dataset. I expect to find a significant association between the variables of height and weight, age and weight, and age and height. I also expect to find a siginificant association between sex and weight.

Extrapalitory Data Analysis

```
children <- read.csv("https://vincentarelbundock.github.io/Rdatasets/csv/Stat2Data/Ki
ds198.csv")

# drop the variable "X" from the dataset children and rename the new dataset children
1
library(dplyr)
```

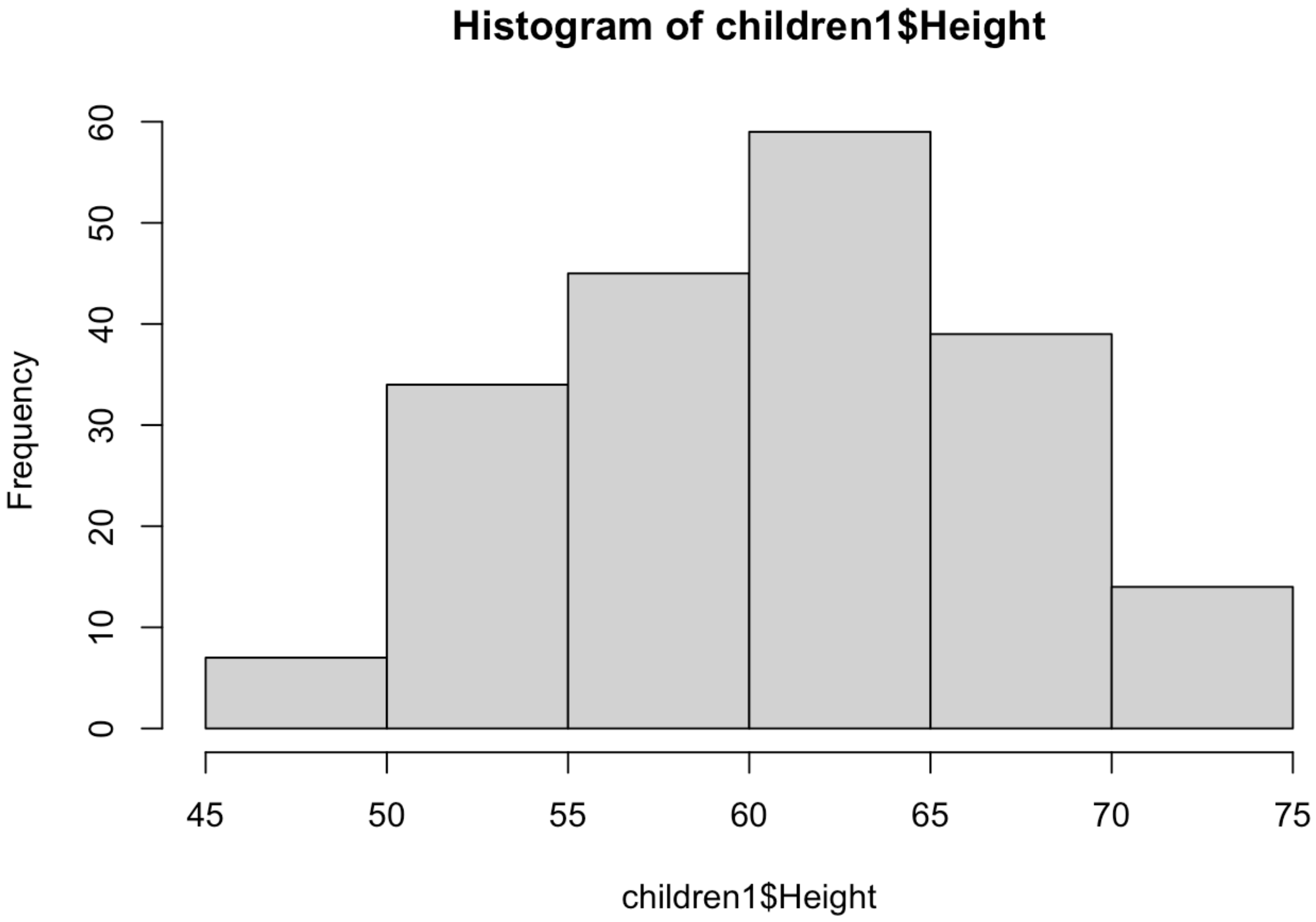
```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

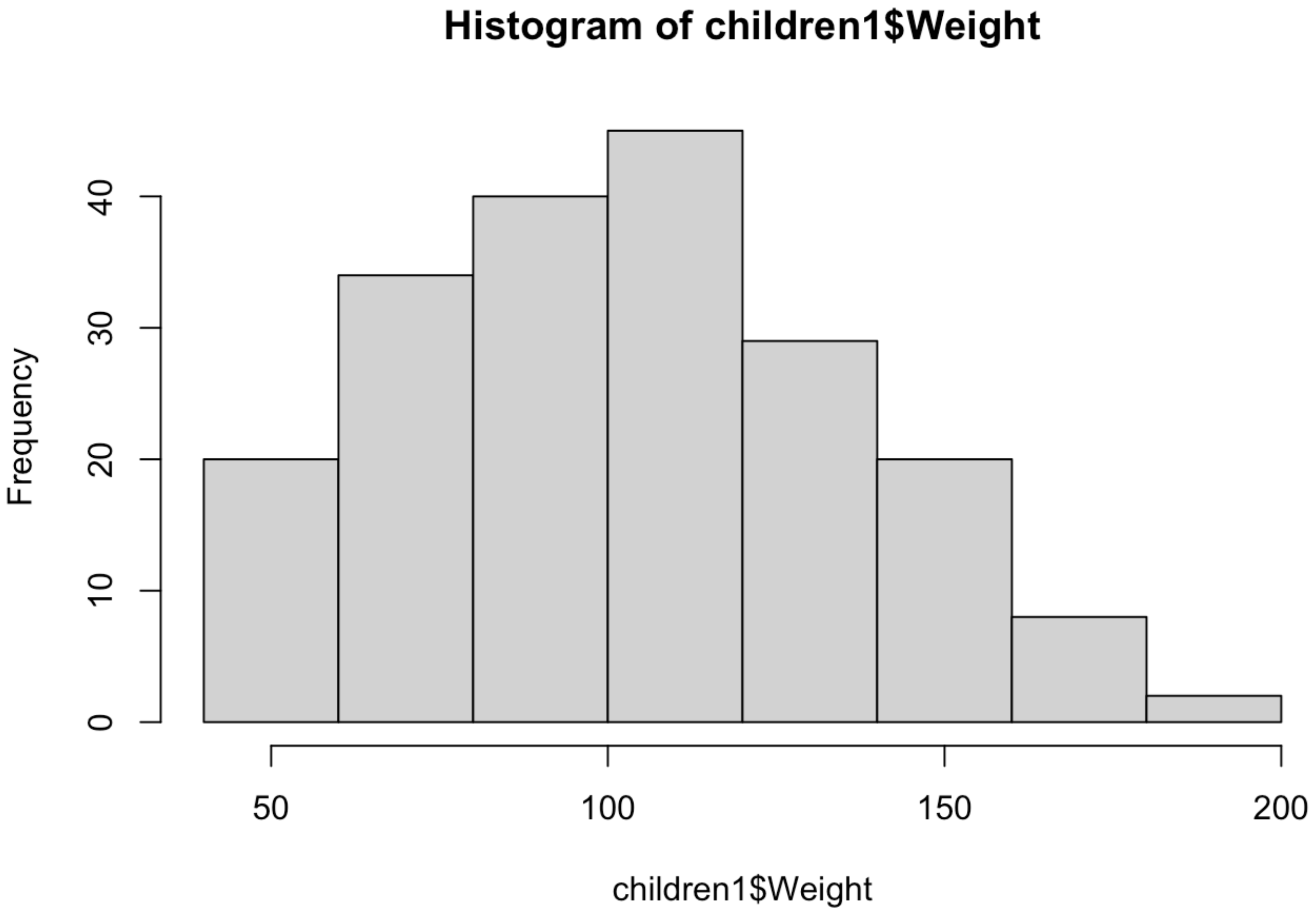
```
children1 <- children %>%
  select(-X)

#summary statistics for dataset children
summary(children1)
```

```
##      Height          Weight          Age             Sex
##  Min.   :47.40   Min.   : 47   Min.   : 99.0   Min.   :0.0000
##  1st Qu.:56.05   1st Qu.: 77   1st Qu.:131.2   1st Qu.:0.0000
##  Median :61.00   Median :103   Median :158.0   Median :1.0000
##  Mean   :60.69   Mean   :104   Mean   :158.4   Mean   :0.5152
##  3rd Qu.:65.55   3rd Qu.:128   3rd Qu.:184.8   3rd Qu.:1.0000
##  Max.   :72.70   Max.   :198   Max.   :221.0   Max.   :1.0000
##       Race
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.1212
##  3rd Qu.:0.0000
##  Max.   :1.0000
```

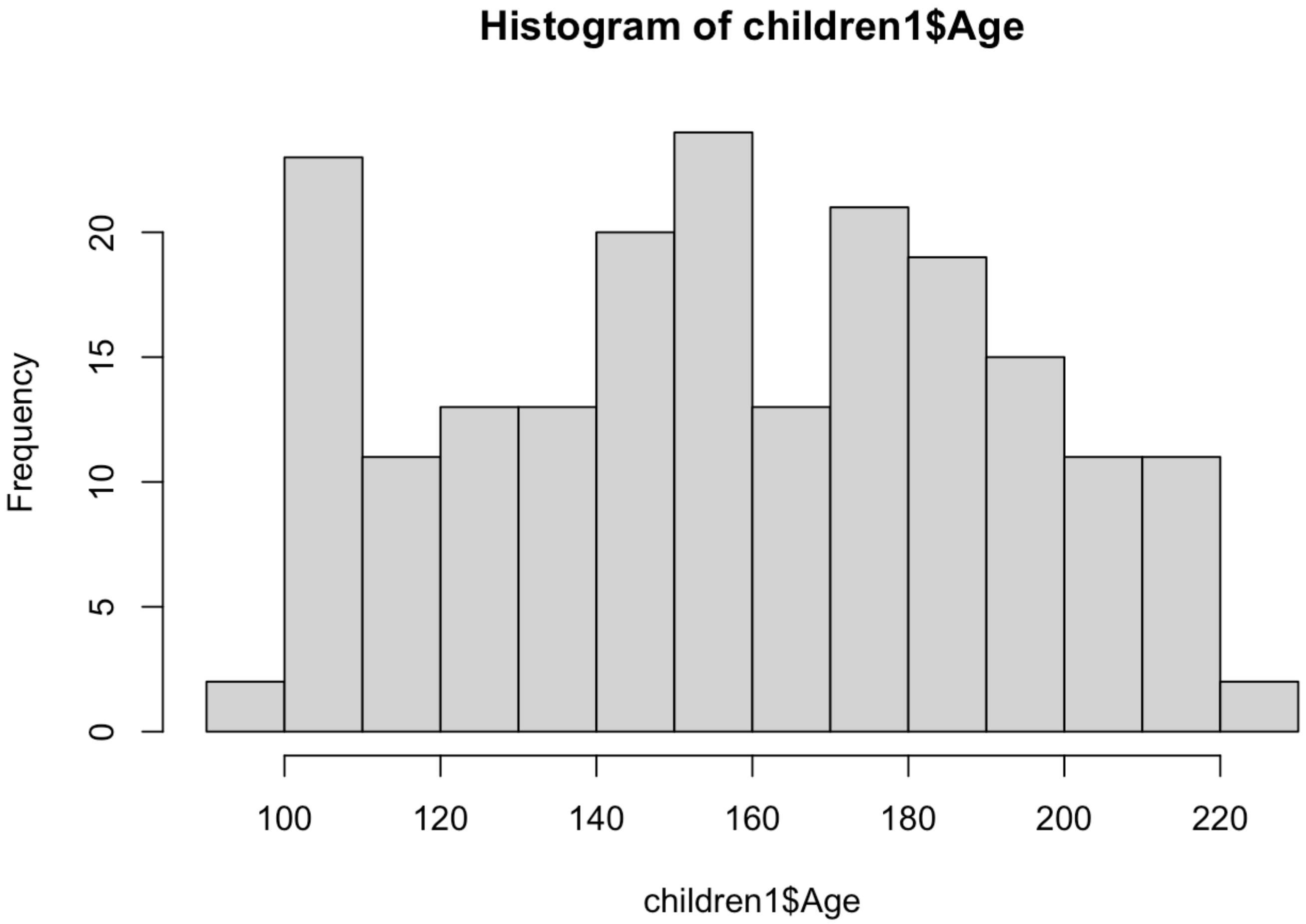
```
#distributions of variables of interest?
hist(children1$Height)
```



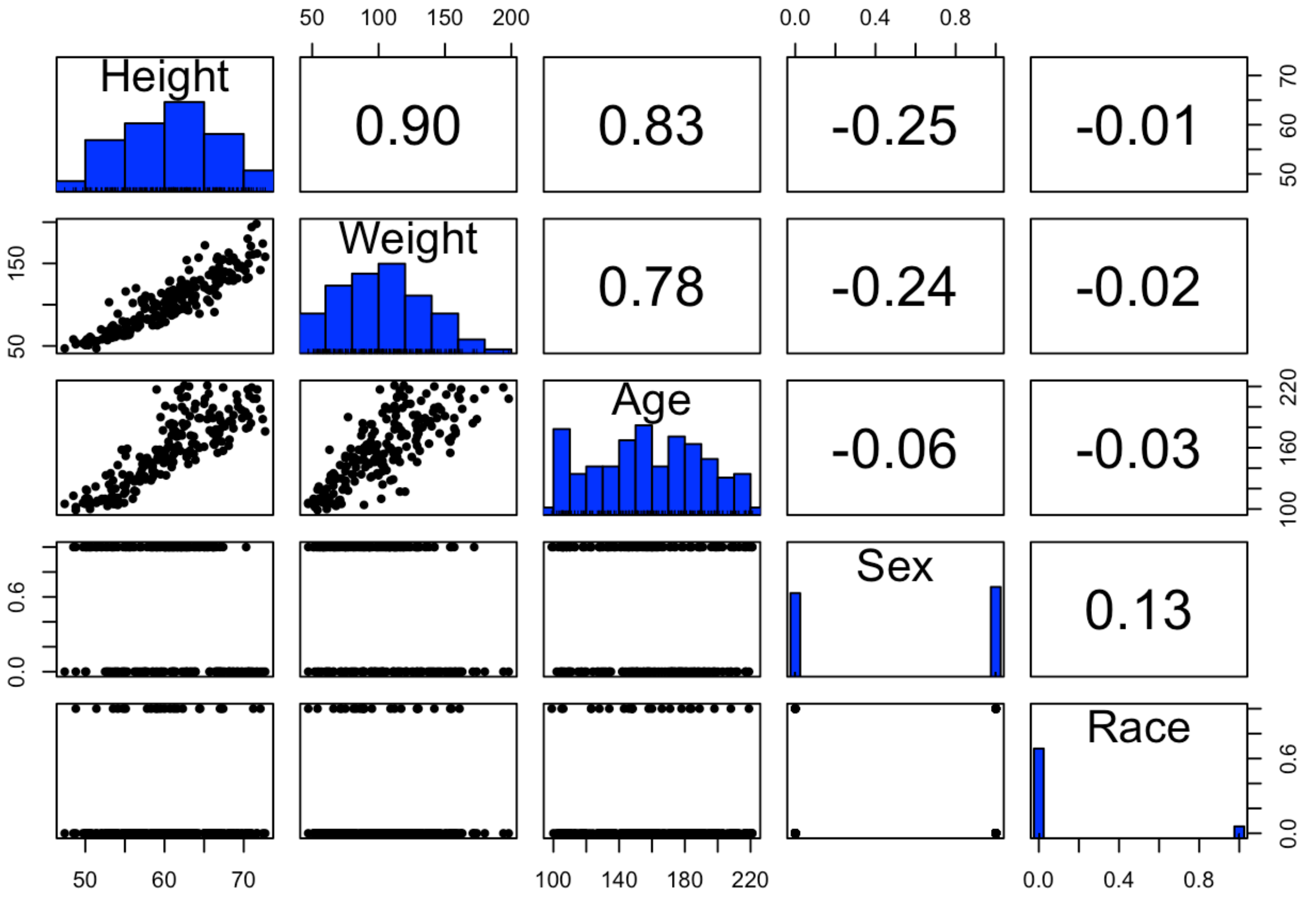
```
hist(children1$Weight)
```



```
hist(children1$Age)

#visualize correlation matrix between all numeric variables? (like in  WS8)?

#explore univariate and bivariate summaries by creating a correlation matrix with uni
variate/bivariate graphs
library(psych)
```



```
pairs.panels(children1[-1,],
method = "pearson", # correlation coefficient method
hist.col = "blue", # color of histogram
smooth = FALSE, density = FALSE, ellipses = FALSE)
```



Upon examining the above correlation matrix, it can be seen that the highest correlation coefficient of R = 0.9 exists between the variables of weight and height. This means that the variable that seems to be the most closely associated with height is weight. The second highest correlation coefficient of R = 0.83 exists between the variables of age and height. This means that the variable that seems to be the second-most closely associated with height is age. Lastly, it can be seen that another high correlation coefficient of R = 0.78 exists between the variables of age and weight, suggesting that age is closely associated with weight.

Histograms were also created to show the distribution of height, weight, and age. The histogram for height appears somewhat normally distributed. The distribution for weight appears less normal, and seems to skew to the right. The distribution for age also does not appear to be completely normal.

MANOVA

```
#perform a MANOVA to test whether any of the numeric variables (height, weight or age
) show a mean difference across the categorical variable of sex.

#HO: The mean height, the mean weight, and the mean age are the same across sexes.

#HA: for at least one of the response variables of height weight and age, at least on
e of the group means differs

#Inspect homogeneity of (co)variances
covchild <- children1 %>% group_by(Sex) %>% do(covs=cov(.[3:2]))
# Covariance matrices per sex
for(i in 1:3){print(as.character(covchild$Sex[i])); print(covchild$covs[i])}
```

```
## [1] "0"
## [[1]]
##              Age    Weight
## Age    1039.2359  944.3638
## Weight  944.3638 1288.1890
##
## [1] "1"
## [[1]]
##              Age   Weight
## Age    1241.3349 778.9225
## Weight  778.9225 791.3415
##
## [1] NA
## [[1]]
## NULL
```

```
#The data passes the assumptions for MANOVA, so now perform the MANOVA
manova_children1 <- manova(cbind(Height, Weight, Age) ~ Sex, data = children1)
summary(manova_children1)
```

```
##            Df  Pillai approx F num Df den Df    Pr(>F)
## Sex         1 0.14251   10.747      3    194 1.447e-06 ***
## Residuals 196
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Results of MANOVA are significant, so perform a one-way ANOVA for each variable
summary.aov(manova_children1)
```

```
##  Response Height :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Sex           1  480.7  480.67   13.99 0.0002412 ***
## Residuals   196 6733.9   34.36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Weight :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Sex           1  12975 12974.5   12.57 0.0004902 ***
## Residuals   196 202303  1032.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Age :
##              Df Sum Sq Mean Sq F value Pr(>F)
## Sex           1   1020  1020.4  0.8924  0.346
## Residuals   196 224102  1143.4
```

```
#Results of the ANOVA were significant, but there are only two categories for the var
iable sex (male or female), so it is not necessary to perform pairwise t-tests

#Calculate probability of at least one type I error. We performed a total of 4 tests
1-0.95^4
```

```
## [1] 0.1854938
```

```
#use the Bonferonni correction to adjust alpha and more confidently reject null hypot
hesis (alpha' = alpha/# of tests)
0.05/4
```

```
## [1] 0.0125
```

The assumptions for the MANOVA are likely met because there is no ratio of four seen between the values of the covariance matrices

We have performed a total of 4 tests above. The probability of having committed at least one type one error was determined to be 0.1854938. After using the Bonferonni correction, the adjusted significance level was determined to be 0.0125. Comparing our above P-values to this corrected significance level will allow us to more confidently reject the null hypotheses of the various tests.

The MANOVA test results above show a P-value less than 0.0125 (P-value = 1.447E-06), so we must reject the null hypothesis that the mean height, the mean weight, and the mean age are the same across sexes. This means that we have significant evidence to say that for at least one of the response variables of height, weight and age, at least one of the group means differs.

When examining the results of the one-way ANOVA test (also shown above), it can be seen that the results were significant for the groups height (P-value = 0.0002412) and weight (P- value = 0.0004902) but were not significant for the group age (P-value = 0.346). This means that for the groups of weight and height, we reject the null hypothesis that the means are equal across sexes.We therefore have sufficient evidence to say the following: Average height differs significantly by sex. Average weight differs significantly by sex. Average age does not differ significantly by sex.

Because there are only two possible categories for sex (Male or Female) it is not necessary to perform a post-hoc analysis to determine which categories of sex have differing means. This is because the ANOVA already confirmed that we have a difference between the mean weight and height across sex. Post-hoc analysis would only be necessary if there was a third category for the variable sex, such as intersex.

Randomization Test for weight with sex as the condition

```
#HO: Mean weight for children is the same for males vs females

#HA: Mean weight for children is different for males vs females


# first calculate the observed test statistic, which is the original mean difference
of weight between the two sexes
true_diff <- children1 %>% group_by(Sex) %>% summarize(means = mean(Weight)) %>% summ
arize(mean_diff = diff(means)) %>% pull
true_diff
```

```
## [1] -16.1973
```

```
#Now randomly mix up the association by purmuting one variable. So keep the same cond
ition, and re-sample the weight across the conditions
perm1 <- data.frame(condition = children1$Sex, weight = sample(children1$Weight))
head(perm1)
```

```
##   condition weight
## 1         0     66
## 2         1    112
## 3         0    161
## 4         1     77
## 5         0    101
## 6         0     88
```

```
#Find the new mean difference
perm1 %>%
  group_by(condition) %>%
  summarize(means = mean(weight)) %>%
  summarize(mean_diff = diff(means))
```

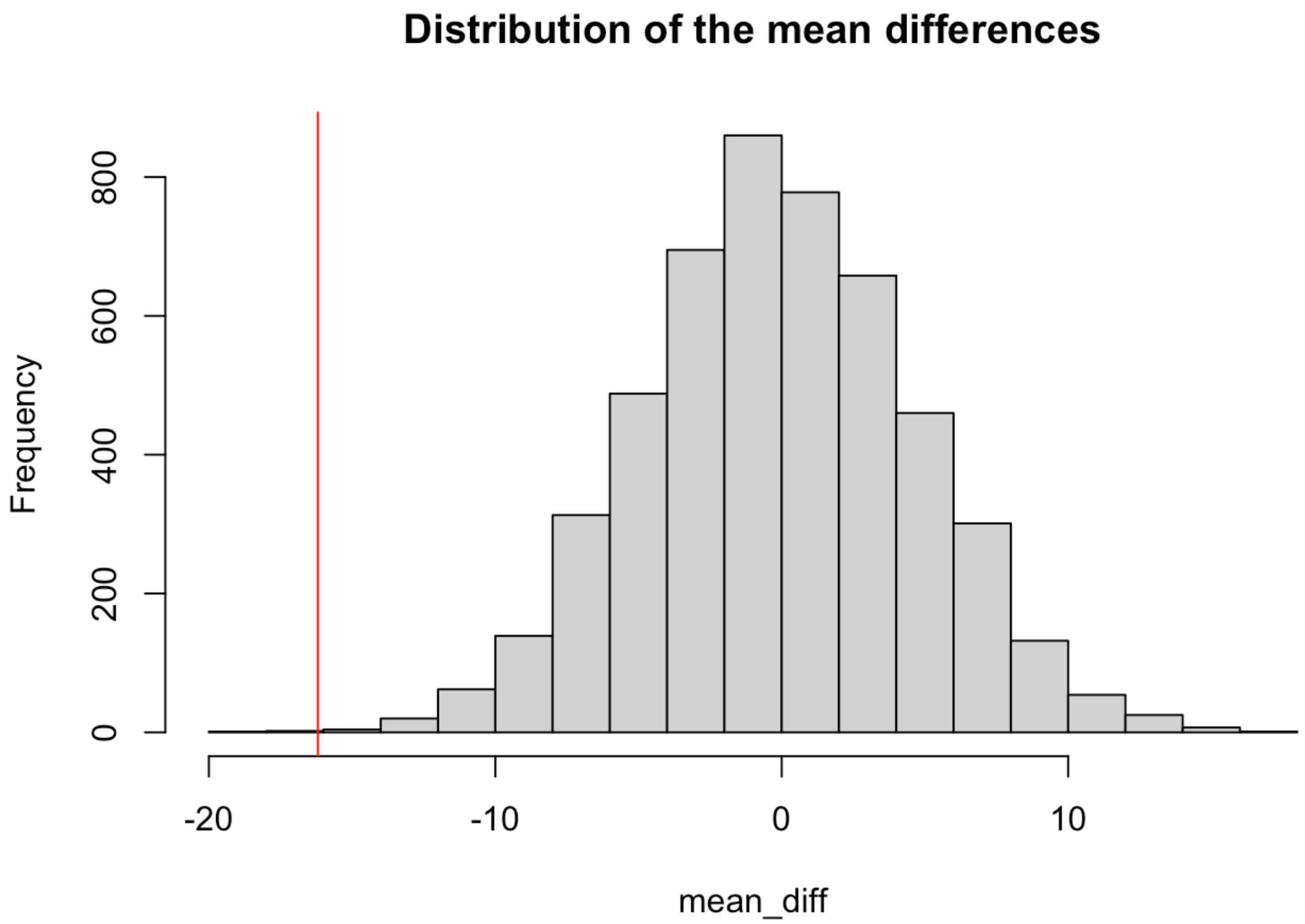
```
## # A tibble: 1 x 1
##   mean_diff
##       <dbl>
## 1     -6.94
```

```
#Repeat randomization
# Keep the same condition, re-sample the weight across conditions
perm2 <- data.frame(condition = children1$Sex, weight = sample(children1$Weight))
head(perm2)
```

```
##   condition weight
## 1         0     55
## 2         1    117
## 3         0     62
## 4         1     79
## 5         0     63
## 6         0     71
```

```
#Find the new mean difference
perm2 %>%
  group_by(condition) %>%
  summarize(means = mean(weight)) %>%
  summarize(mean_diff = diff(means))
```

```
## # A tibble: 1 x 1
##   mean_diff
##       <dbl>
## 1      5.32
## Repeat randomization many times
# Create an empty vector to store the mean differences
mean_diff <- vector()
# Create many randomizations with a for loop
for(i in 1:5000){
temp <- data.frame(condition = children1$Sex, weight = sample(children1$Weight))
mean_diff[i] <- temp %>% group_by(condition) %>% summarize(means = mean(weight)) %>%
summarize(mean_diff = diff(means)) %>% pull}
#Represent the distribution of the mean differences with a vertical line showing the
true difference.
#In my first submission I included two ablines in my plot. As per the comments on gradescope, I have changed this so that there is only one abline representing the value
that was previously calculated
{hist(mean_diff, main="Distribution of the mean differences"); abline(v = -16.197, col = "red")}
```



```
#Calculate the corresponding two-sided p-value
mean(mean_diff > -true_diff | mean_diff < true_diff)
```

```
## [1] 8e-04
```

```
mean(mean_diff > 16.197 | mean_diff < -16.197)
```

```
## [1] 8e-04
```

Null Hypothesis: Mean weight for children is the same for males vs females

Alternative Hypothesis: Mean weight for children is different for males vs females

The resulting P-value of the randomization test was below 0.05 (P-value = 0.0004), so we reject the null hypothesis that the mean weight for children is the same for males vs females.Therefore, we have sufficient evidence that the mean weight for children varies significantly between males and females.

Multiple Linear Regression Model Predicting Weight From Height and Age

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

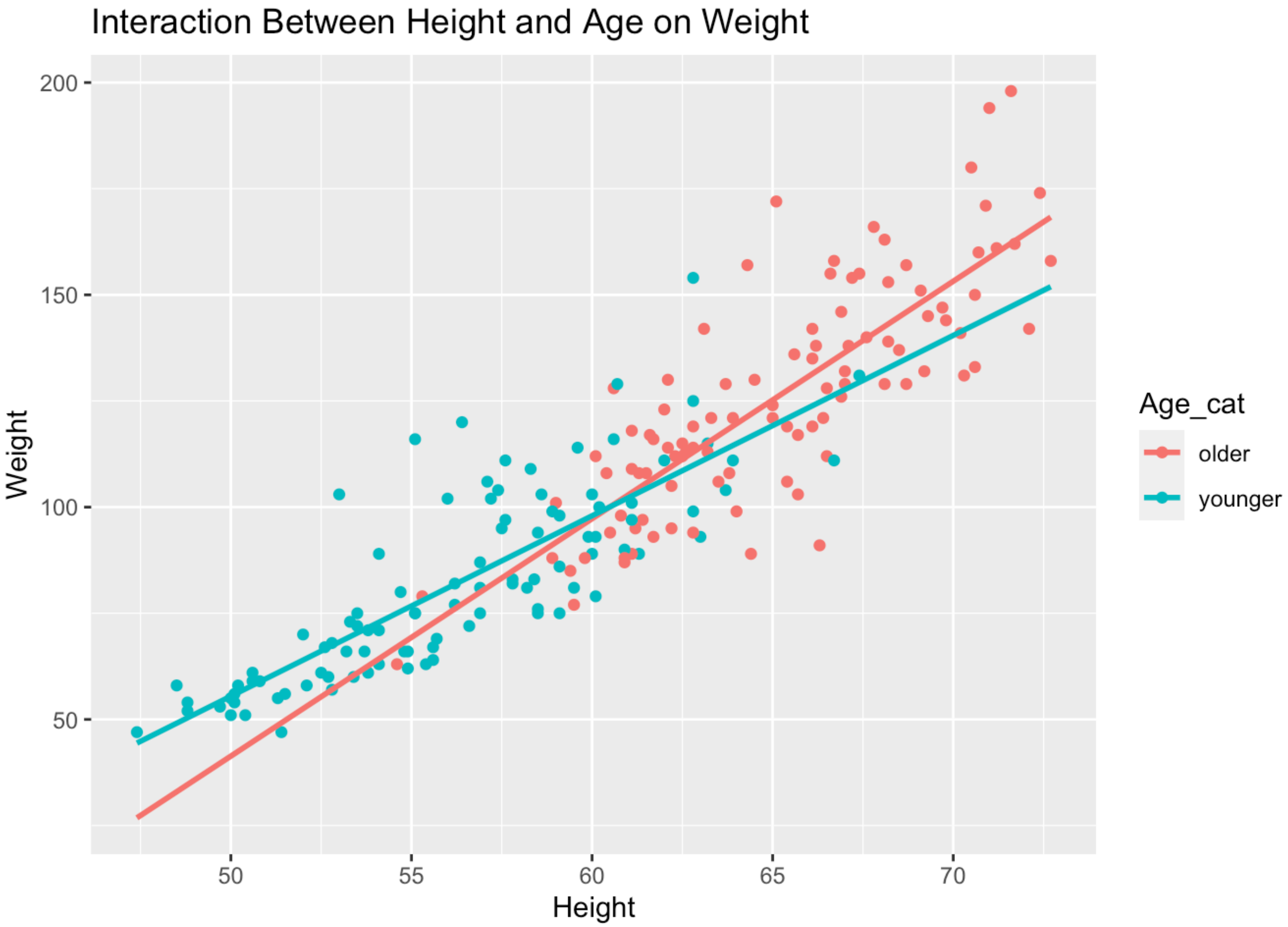
```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
#center the means for the numeric variables involved in the interaction, height and a
ge. (center the data around the means, so the intercept becomes more informative).
children1$Height_c <- children1$Height - mean(children1$Height)
children1$Age_c <- children1$Age - mean(children1$Age)

#Run a MLR model using the centered mean values and include an interaction term
fit <- lm(Weight ~ Height_c * Age_c, data = children1)
summary(fit)
```

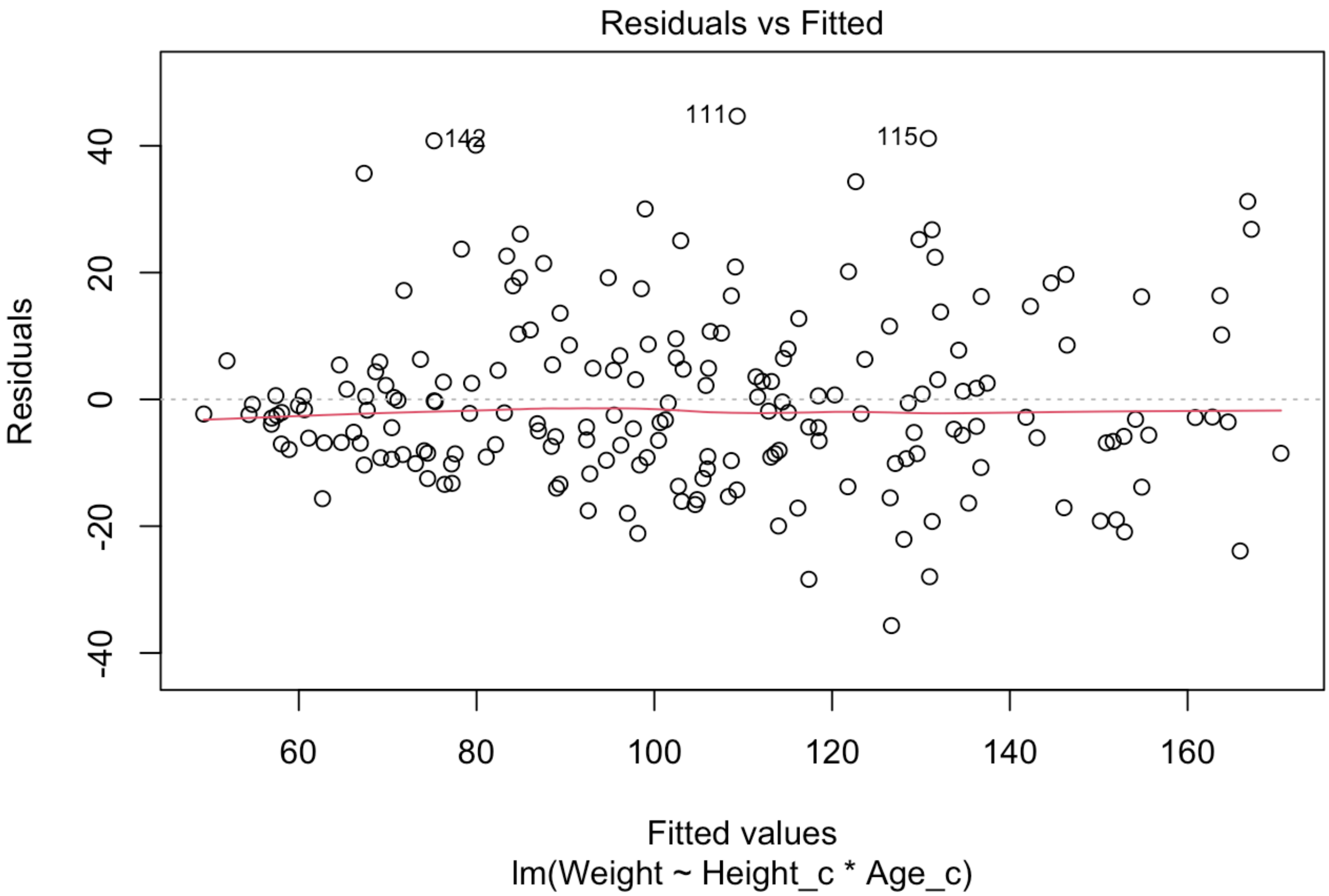
```
##
## Call:
## lm(formula = Weight ~ Height_c * Age_c, data = children1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.683  -8.675  -2.350   6.315  44.676
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.004e+02  1.363e+00  73.695  < 2e-16 ***
## Height_c       4.483e+00  2.984e-01  15.024  < 2e-16 ***
## Age_c          1.216e-01  5.344e-02   2.275 0.024022 *
## Height_c:Age_c 2.106e-02  5.496e-03   3.833 0.000171 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14 on 194 degrees of freedom
## Multiple R-squared:  0.8235, Adjusted R-squared:  0.8207
## F-statistic: 301.7 on 3 and 194 DF,  p-value: < 2.2e-16
```

```
#create a graph to visualize the interaction between Height and Age on Weight
#consider one of the numeric variables in the interaction (age) as a categorical vari
able.
children1 %>%
  mutate(Age_cat = case_when(
    Age_c < median(Age_c) ~ "younger",
    Age_c >= median(Age_c) ~ "older")) %>%
  ggplot(aes(x = Height, y = Weight, color = Age_cat)) +
  geom_point() + geom_smooth(method = lm, se = FALSE, fullrange = "True") + ggtitle("
Interaction Between Height and Age on Weight")
```

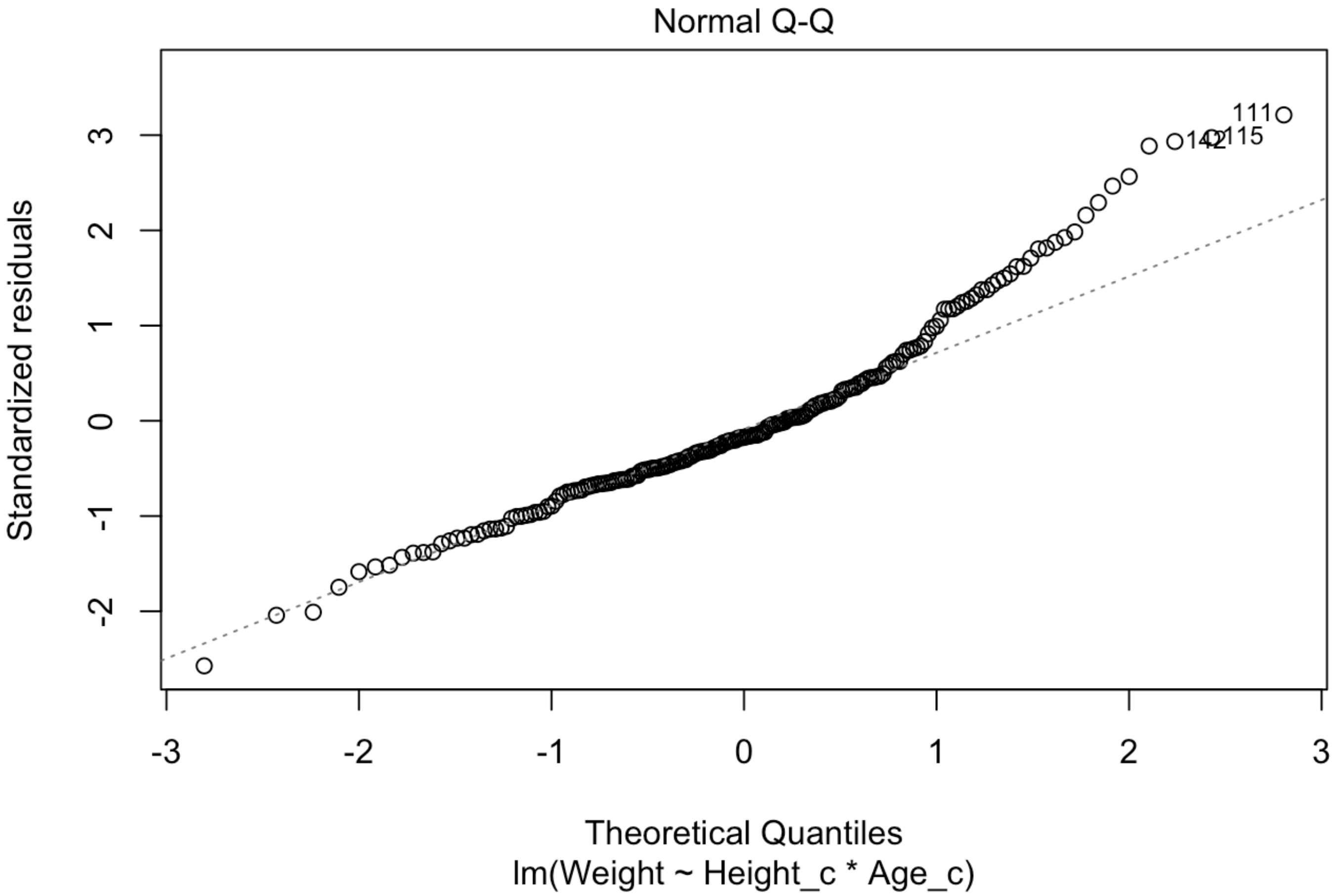
```
## `geom_smooth()` using formula 'y ~ x'
```



```
##check the assumptions of linearity, normality, and homoscedasticity graphically

#Check for equal variance with residuals vs fitted plot
plot(fit, which =1)
```



```
#check for normality with a Q-Q plot of the residuals
plot(fit, which =2)
```



```
#confirm normality, results of Q-Q plot appear to differ from a normal distribution.
use a Shapiro-Wilk test in which the null hypothesis is that the data is normal
shapiro.test(fit$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit$residuals
## W = 0.95744, p-value = 1.165e-05
```

```
##confirm equal variance (homoscedasticity) with a Breusch-Pagan test in which the nu
ll hypothesis is that the data shows homoscedasticity.
#first install new packages
#install.packages("sandwich")
library(sandwich)
#install.packages("lmtest")
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

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

```
#Now perform the Breusch-Pagan test with HO:homoscedasticity
bptest(fit)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 6.6116, df = 3, p-value = 0.08536
```

HO: slopes do not differ significantly from zero HA: slopes do differ significantly from zero

Using centered numeric variables of height and age to interpret this result.

When controlling for Age, there is a significant relationship between weight and height (P-value = 2E-16, t= 15.024, df= 194).When age remains constant, weight increases on average by 4.483 pounds with every 1 inch increase in height.

When controlling for height, there is a significant relationship between weight and age (P-value = 0.024022, t= 2.275, df= 194). When height remains constant, weight increases on average by 1.216E-01 pounds with every 1 month increase in age.

There is a very significant interaction between the variables height and age. Because the interaction between height and age was determined to be significant, the model is actually being driven by this interaction, rather than by the individual relationships between weight and height. So on average taller, “older” people (who are above the median age) tend to weigh more, and Shorter, “Younger” people (who are below the median age) tend to weigh less. The slope for height on weight is 2.106E-2 higher for older people compared to younger people.

The multiple R-squared value from the above summary shows that 82.35% of the variation in weight is explained by height, age, and the interaction between height and age. The adjusted R-squared value from the summary shows that 82.07% of the variation in weight is explained by height, age, and the interaction between height and age.

The Shapiro-Wilk test was used to confirm the normality of the data. The resulting p-value was less than 0.05, so we reject the null hypothesis that the data is normal. This means that the data violated the normality assumption (W = 0.95744, p-value = 1.165e-05).

The Breusch-Pagan test was used to confirm equal variance (homoscedasticity). The resulting P-value was larger than 0.05, so we fail to reject the null hypothesis that the data meets the equal variance assumption. This means that the data has approximately equal variance (BP = 6.6116, df = 3, p-value = 0.08536)

```
#Recompute regression results with the robust standard errors (regardless of meeting assumption of equal variance)
coeftest(fit, vcov = vcovHC(fit))
```

```
##
## t test of coefficients:
##
##                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    1.0044e+02 1.3880e+00 72.3635 < 2.2e-16 ***
## Height_c       4.4829e+00 2.8330e-01 15.8237 < 2.2e-16 ***
## Age_c          1.2157e-01 5.3310e-02  2.2804 0.0236698 *
## Height_c:Age_c 2.1064e-02 5.4731e-03  3.8487 0.0001611 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##Now compute bootstrapped standard errors
# Use the function replicate to repeat the process (similar to a for loop)
samp_SEs <- replicate(5000, {
  # Bootstrap your data (resample observations)
boot_data <- sample_frac(children1, replace = TRUE)
  # Fit regression model
fitboot <- lm(Weight ~ Height_c * Age_c, data = boot_data)
  # Save the coefficients
  coef(fitboot)
})
# Estimated SEs
samp_SEs %>%
  # Transpose the obtained matrices
t %>%
  # Consider the matrix as a data frame
as.data.frame %>%
  # Compute the standard error (standard deviation of the sampling distribution)
summarize_all(sd)
```

```
##   (Intercept)  Height_c      Age_c Height_c:Age_c
## 1    1.358927 0.2760752 0.05327714    0.005351136
```

```
#compare the bootstrapped results with the uncorrected standard errors
coeftest(fit)[,1:2]
```

```
##                    Estimate  Std. Error
## (Intercept)    100.43902941 1.362897675
## Height_c         4.48288993 0.298380950
## Age_c            0.12156820 0.053445059
## Height_c:Age_c   0.02106404 0.005496044
```

```
#compare the bootstrapped results with the robust standard errors
coeftest(fit, vcov = vcovHC(fit))[,1:2]
```

```
##                    Estimate  Std. Error
## (Intercept)    100.43902941 1.387978681
## Height_c         4.48288993 0.283301592
## Age_c            0.12156820 0.053309521
## Height_c:Age_c   0.02106404 0.005473065
```

The regression results were recomputed with robust standard errors and there were no changes in the significance of the results.When controlling for Age, there is still a significant relationship between weight and height.When controlling for height, there is still a significant relationship between weight and age. Lastly, the interaction between height and age remains highly significant. This means that the model is still being driven by the interaction between height and age.

The bootstrapped standard errors were also computed. The bootstrapped standard error was found to be 0.2858167 fro height, 0.05399789 for age, and 0.005392686 for the interaction between height and age.

The uncorrected standard error for height is larger than the bootstrapped standard error for height (0.298380950 > 0.2858167). The uncorrected standard error for age is smaller than the bootstrapped standard error for age (0.053445059 < 0.05399789). Lastly, the uncorrected standard error for the interaction between height and age is larger than the bootstrapped standard error (0.005496044 > 0.005392686).

The bootstrapped standard errors for height and age are larger than the robust standard errors. The bootstrapped standard error for the interaction between age and height is smaller than the robust standard error.

Logistic Regression Model Preditcting Sex from Weight and Height

```
# Create a binary variable coded as 1 and 0 for sex being male
# Remember that for the original dataset children1, the varibale Sex is coded as male
= 0, female = 1
children1_male <- children1 %>%
mutate(Male = ifelse(Sex == "0", 1, 0))

#Define the definition of odds = p/(1-p)
odds <- function(p)p/(1-p)
# Simulate probability values (varying between 0 and 1 by 0.1)
p <-seq(0, 1, by = .1)

# Define the logit link function (logarithm of odds)
logit <- function(p) log(odds(p))

# Fit a regression model predicting Sex from weight and height
fitlg <- glm(Sex ~ Weight + Height, data = children1, family = binomial(link="logit")
)
summary(fitlg)
```

```
##
## Call:
## glm(formula = Sex ~ Weight + Height, family = binomial(link = "logit"),
##     data = children1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7228  -1.1186   0.7788   1.1276   1.4900
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.663540   2.490679   1.872   0.0612 .
## Weight      -0.004519   0.010128  -0.446   0.6555
## Height      -0.068022   0.055552  -1.224   0.2208
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 274.30  on 197  degrees of freedom
## Residual deviance: 260.57  on 195  degrees of freedom
## AIC: 266.57
##
## Number of Fisher Scoring iterations: 4
```

```
# Interpret the coefficients by considering the odds (inverse of log(odds))
exp(coef(fitlg))
```

```
## (Intercept)      Weight      Height 
## 106.0107062   0.9954916   0.9342401
```

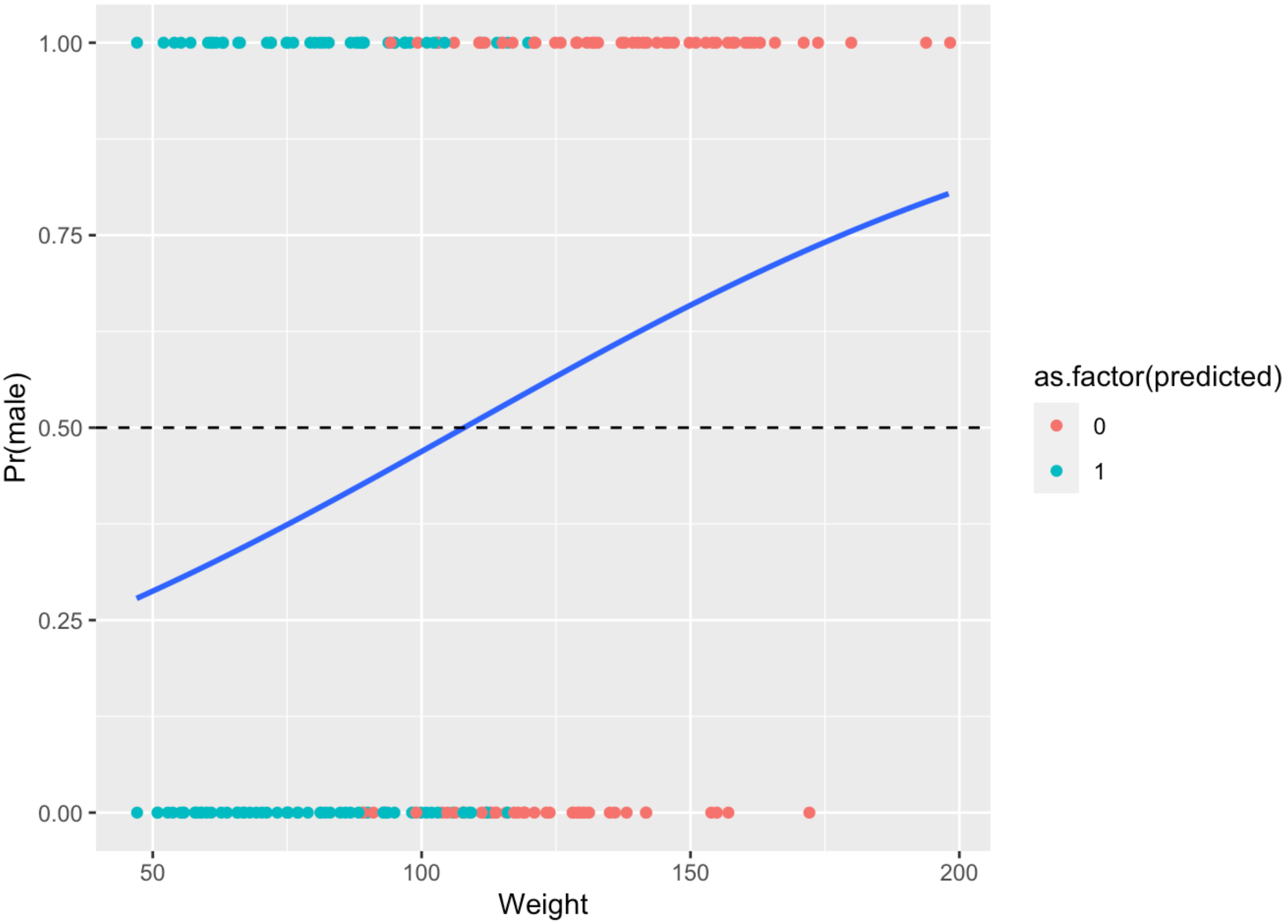
The results of the logistic regression model were not found to be significant (all P-values > 0.05), but we will interpret the coefficient estimates regardless:

When controlling for height, every one unit increase in weight decreases the log odds of a child’s sex being male by 0.004519. When controlling for weight, every one unit increase in height decreases the log odds of a child’s sex being male by 0.068022.

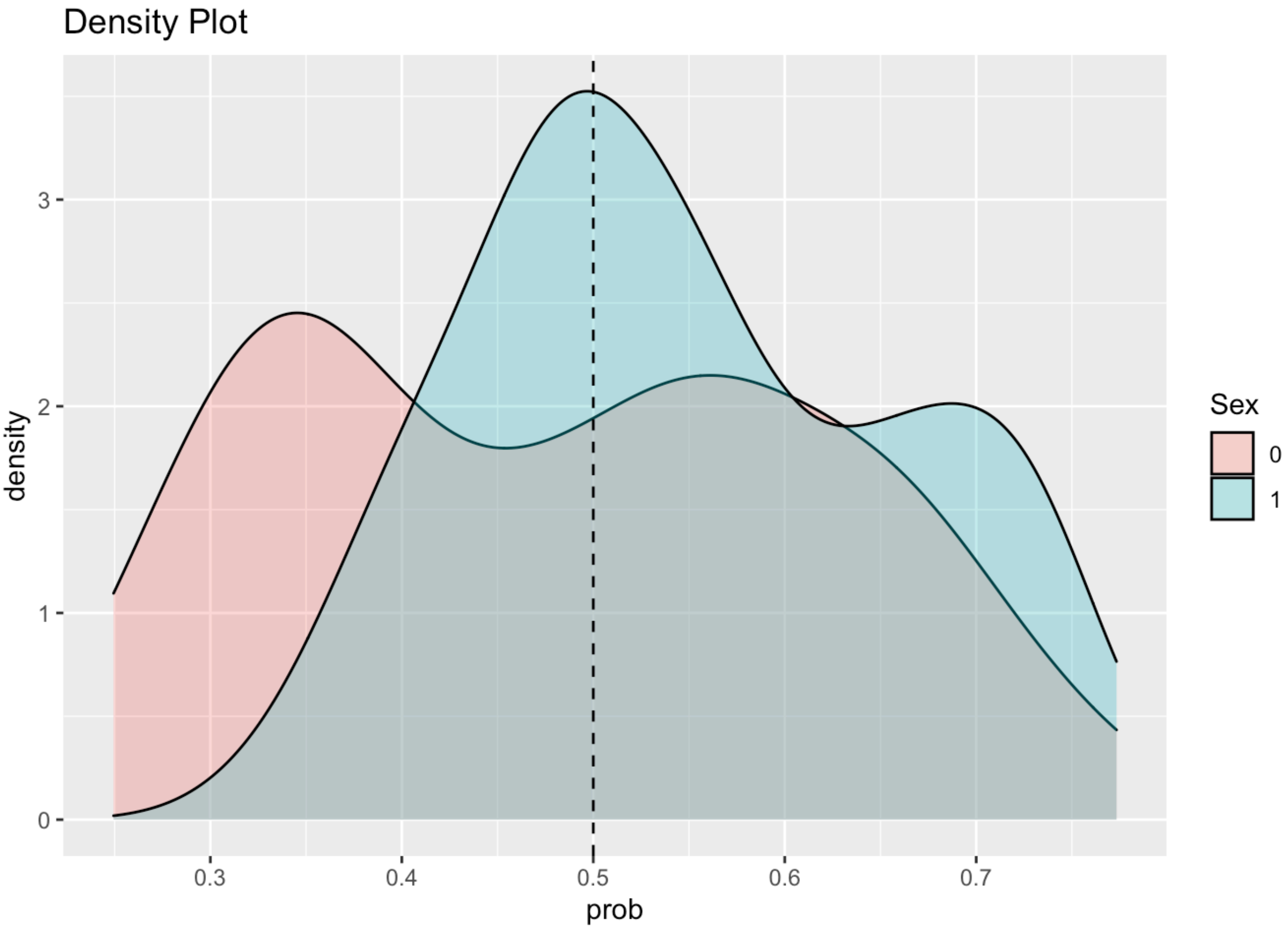
Every one pound increase in weight multiplies the odds of a child’s sex being male by 0.9954916. Every one inch increase in height multiplies the odds of a child’s sex being male by 0.9342401.

```
# Add predicted probabilities to the dataset
children1_male$prob <- predict(fitlg, type = "response")
# Predicted outcome is based on the probability of being male
# if the probability is greater than 0.5, the person is considered to be male
children1_male$predicted <- ifelse(children1_male$prob > .5, "1", "0")
 # Plot the model
ggplot(children1_male, aes(Weight,(Male))) +
geom_jitter(aes(color = as.factor(predicted)), width = .3, height = 0) + stat_smooth(method="glm", method.args = list(family="binomial"), se = FALSE) + geom_hline(yintercept = 0.5, lty = 2) +
ylab("Pr(male)")
```

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



```
#Density plot
ggplot(children1_male, aes(prob, fill = as.factor(Sex))) + geom_density(alpha = .3) +
geom_vline(xintercept = 0.5, lty = 2) +
labs(fill = "Sex") + ggtitle("Density Plot")
```



The red portion of the density plot above represents the true negative observations. In the context of this model, it represents the number of females that were correctly classified as being female.

The blue portion of the density plot above represents the true positive observations. In the context of this model, it represents the number of males that were correctly classified as being male.

The gray portion of the density plot above represents the misclassified observations. In the context of this model, it represents the number of males and females that were classified under the incorrect gender.

```
#confusion matrix
table(truth = children1_male$Male, prediction = children1_male$predicted)
```

```
##      prediction
## truth  0  1
##     0 41 61
##     1 50 46
```

```
#Compute accuracy (proportion of correctly classified cases)
(41 + 46)/198
```

```
## [1] 0.4393939
```

```
#compute the sensitivity (true positive rate)
46/96
```

```
## [1] 0.4791667
```

```
#compute the specificity (true negative rate)
41/102
```

```
## [1] 0.4019608
```

```
#compute the precision (positive predictive value)
46/107
```

```
## [1] 0.4299065
```

The accuracy represents the portion of correctly classified cases, meaning the proportion of individuals who’s gender was correctly classified by the model. the accuracy of the model is low (0.4393939), meaning that there are a large portion of people who’s genders were misclassified in the model. This can be see in the density plot above, in which the misclassified values are represented by the color gray.

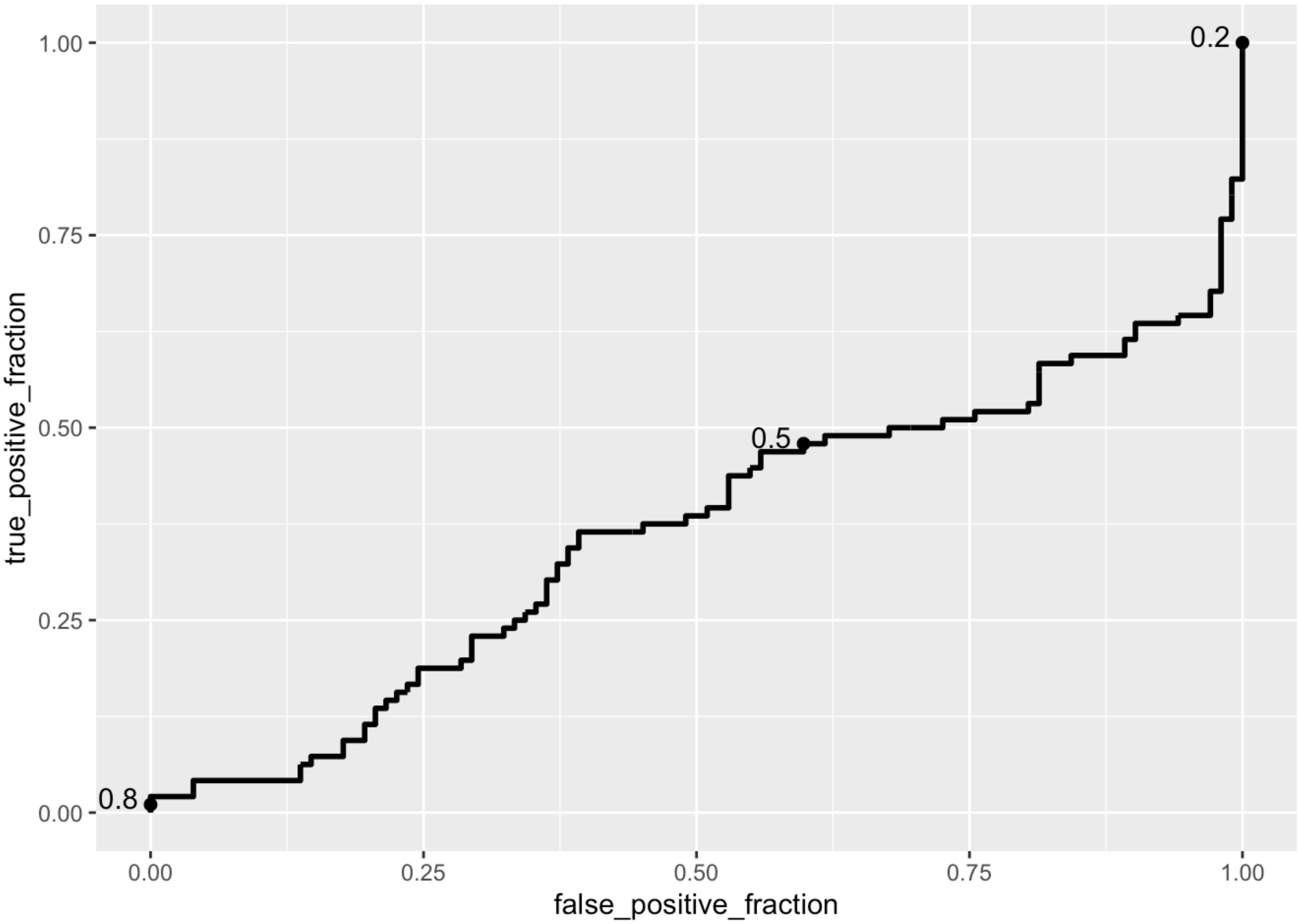
The sensitivity represents the true positive rate. In the context of this model, it represents the amount of males correctly detected compared to the amount of males that there actually were. The sensitivity of this model is low (0.4791667), so it does not detect positive cases accurately.

The specificity represents the true negative rate. In the context of this model, it represents the amount of females correctly detected compared to the amount of females that there actually were. The specificity of this model is low (0.4019608), so it does not detect negative cases accurately.

The precision represents the proportion of the true positive compared to the total amount predicted. In the context of this model, it represents amount of correctly classified males compared to the total number of predicted males. The precision of the model is low (0.4299065).

```
library(plotROC)
```

```
#Plot ROC curve
ROCplotproject2 <- ggplot(children1_male) +  geom_roc(aes(d = Male, m = prob), cutoffs.at = list(0.1, 0.5, 0.9))
ROCplotproject2
```



```
#calculate AUC
calc_auc(ROCplotproject2)
```

```
##   PANEL group      AUC
## 1     1    -1 0.361826
```

The AUC can be interpreted as the probability that a randomly selected male child has a higher predicted probability than a randomly selected female child.

On average, 36% of the time males will have higher probabilities than females.

The AUC value is very low (0.361826), so this model is poor at predicting sex from weight and height.