BIOST 515/518 Homework 1

Latera Tesfaye Olana

11 January, 2023

Answers:

1. **Yes.** I believe there is a linear relationship between weight (pounds) and height (inches). From the given sample data, with an increase in height (inches) there seems to be an increase in weight (pounds). Accordingly, the first order trend suggestive of a tendency for higher average weight in taller groups. As shown on Figure 1, there seems to be some suggestion of greater variability in weight in taller groups than there is in the shorter groups.
   In order to accurately determine or quantify the linear relationship between these two variables, statistical tests such as Winsorized Pearson correlation coefficient - robust, can be applied (eyes might be deceiving). The results of this test suggest a moderate (1), statistically significant correlation ($\hat{r} = 0.50$, $p < 0.05$) between weight and height. In another word, we reject the null hypothesis of no association and conclude that there is evidence for a linear association between weight and height. Figures 1 and 2 provide further visual representation of this relationship through scatter and histogram plots, as well as more detailed statistical information.

2. In a study of 3,154 male volunteers aged 39 to 59 years old, we found evidence of an association between weight and height ($p < 2.2e\text{-}16$). An increase of one inch in height was associated with a 4.45 pounds increase in weight (95% confidence interval [4.18; 4.72]).

3. Depending on the values of their standard error, RSE is an appropriate choice. This is mainly due to the RSE has higher standard error as compared to naive.
   The choice between OLS (Naive) and Robust Standard Error (RSE) depends on the assumptions of our model and the properties of the data. OLS assumes that the errors are normally distributed with constant variance (homoscedasticity) and that there are no outliers in the data. If these assumptions are not met, OLS can produce biased and inefficient estimates of the model parameters.
   RSE, alternatively, can be utilized to acquire more robust assessments of the model coefficients when the OLS assumptions are contravened (not that we must verify it beforehand). The fundamental concept behind RSE is to implement a weighting function that diminishes the impact of outliers in the data. This can result in more stable and precise evaluations of the model coefficients, even when the OLS assumptions are not fulfilled.
   As demonstrated in tables 1 and 2, the coefficient for height remains consistent, however, variations in the estimated standard errors, as well as the corresponding "t" and "p-values," are observed (with slight deviations). The OLS standard errors were under-estimating the standard error for the coefficient of height. The coefficient, on the other hand, retained its significance as the p-values were below 0.05. It can be inferred that heteroscedasticity is not a major concern in this scenario as there was no alteration in the significance of coefficients. They were statistically significant with OLS standard errors and continue to be statistically significant with Robust Standard Errors.
   However, it is really **naive** for me to say the selection of the model only depends on the standard error. Standard error is a statistical measure used to evaluate the precision of a model's estimates. However, it is important to note that it does not account for other crucial considerations such as the overall accuracy of the model, its generalizability to new data, and its interpretability. To comprehensively

assess and compare models, other metrics such as the mean squared error, coefficient of determination (R-squared), and cross-validation scores should also be taken into consideration. All in all, selecting the **RSE** relatively poses few risks.

4. It seems reasonable. For men aged 39 to 59 years old and free of heart disease, we can estimate their expected weight given their height, as long as height is scientifically meaningful.

5. The extrapolation of estimating the mean weight of adolescent boy is not scientifically meaningful as the population of interest was men aged 39 to 59 years old and free of heart disease.

6. In a study of 3,154 male volunteers aged 39 to 59 years old, the study found evidence of an association between weight and height ($p < 2.2e\text{-}16$). An increase of one centimeter in height was associated with a 1.75 pounds increase in weight (95% confidence interval [1.64; 1.86]).
   An increase of one centimeter was approximeted to be increase of 2.54 inches. The rescaling of the dependent variable (height) only changes the difference in the estimated expected mean of weight between groups varying by one unit of height (slope or $\hat{\beta}$). Where as, it does not change the estimated expected value of weight for a given male individual with a zero unit of height (intercept $\hat{\beta}_0$). The degree and statistical significance of the association (correlation) with weight will not change with changing unit as well. In addition, measure of fits such as R^2 will not change. The scaling of the independent variable with a constant c, results in $\frac{1}{c} * se(\hat{\beta})$. Results are shown on table 3 and 4.

7. The fitted value will be the best prediction. This is because the fitted value takes into account the relationship between height and weight for the entire study population (3,154 male volunteers aged 39 to 59 years old) including the weight of individuals who are 73 inches tall, rather than just the weight of one individual (i.e., participant 2001). The fitted value is calculated using the equation of the line of best fit, which takes into account the average weight with proper uncertainty for a given height. This allows for a more accurate prediction of weight based on the height of an individual within the given study population. In addition, the choice of 73 inches, seems rather arbitrary there are multiple observations where the weight of individuals are 150 pounds, but with different height group.

Supplementary results - Tables

Table 1: Linear fit of weight and height (inches) - OLS

	Estimate	Naive SE	Robust SE	95%L	95%H	t value	Pr(>|t|)
(Intercept)	-140.2811	8.7793	9.6141	-159.1317	-121.4306	-14.5912	0
height0	4.4460	0.1257	0.1384	4.1746	4.7175	32.1167	0

Table 2: Linear fit of weight and height (inches) - RSE

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-140.281132	9.6110754	-14.59578	0
height0	4.446043	0.1383899	32.12693	0

Table 3: Linear fit of weight and height (centimeters) - OLS

	Estimate	Naive SE	Robust SE	95%L	95%H	t value	Pr(>|t|)
(Intercept)	-140.2811	8.7793	9.6141	-159.1317	-121.4306	-14.5912	0
height0_cm	1.7504	0.0495	0.0545	1.6435	1.8573	32.1167	0

Table 4: Linear fit of weight and height (centimeters) - RSE

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-140.281132	9.6110754	-14.59578	0
height0_cm	1.750411	0.0544842	32.12693	0

Supplementary results - Plots

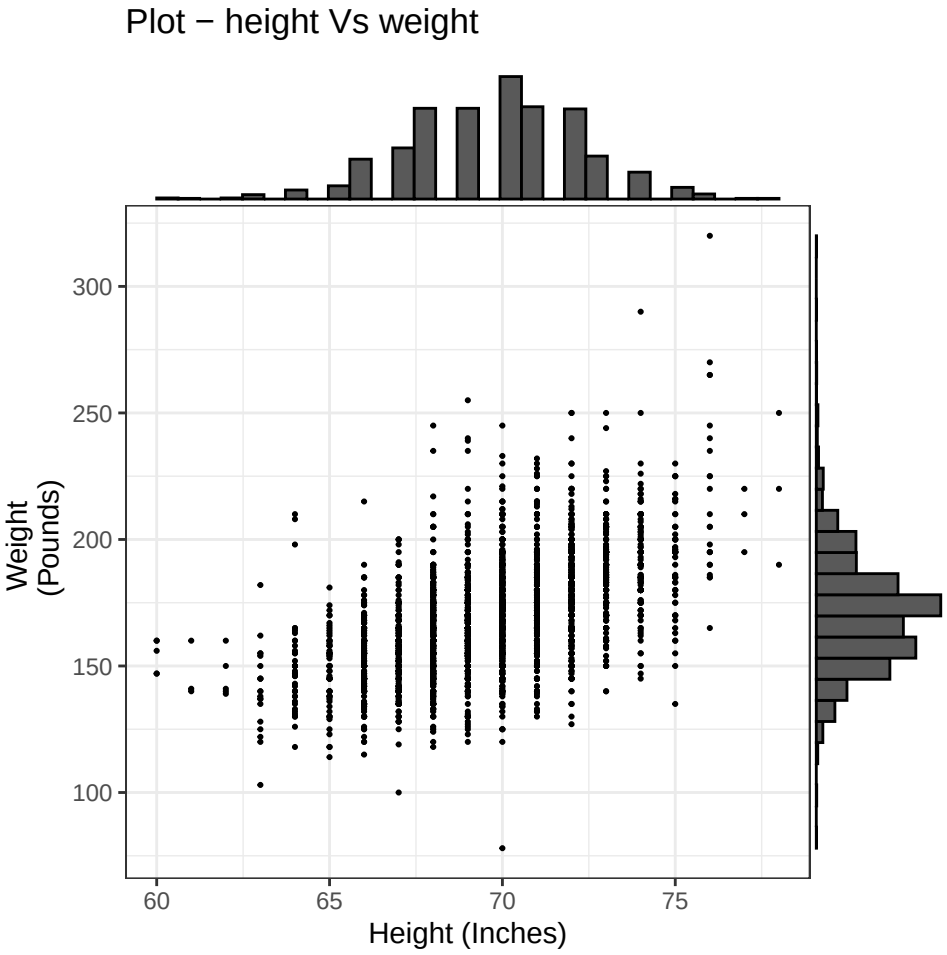


Figure 1: Height and weight scatter plot

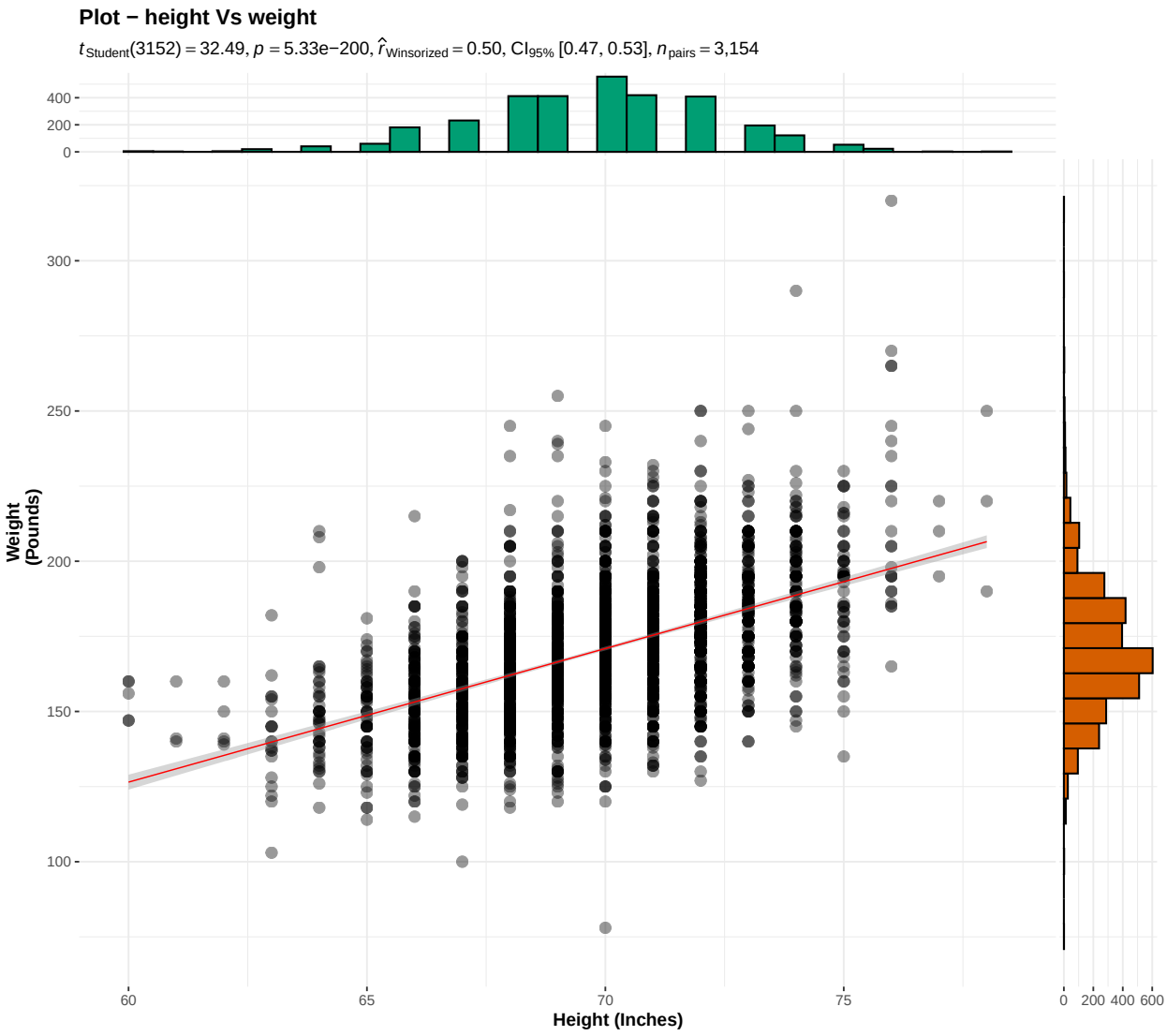
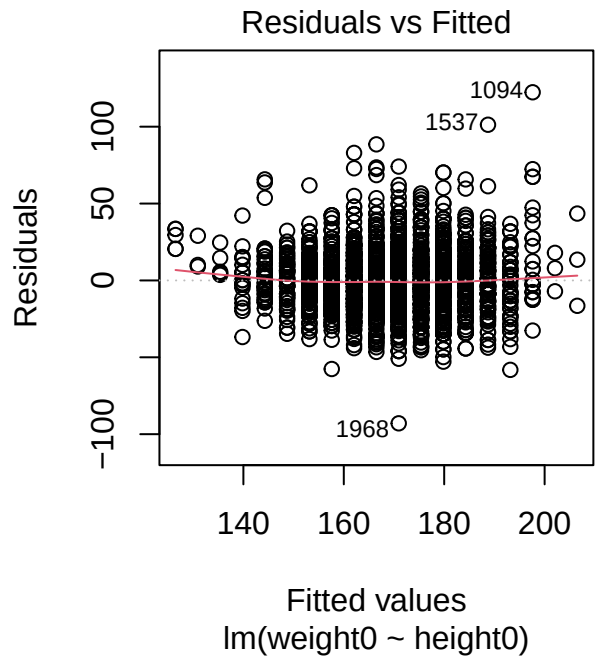
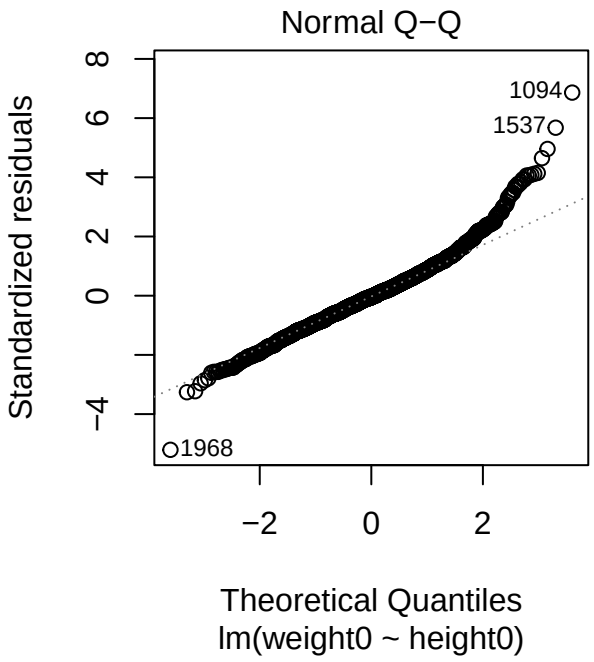
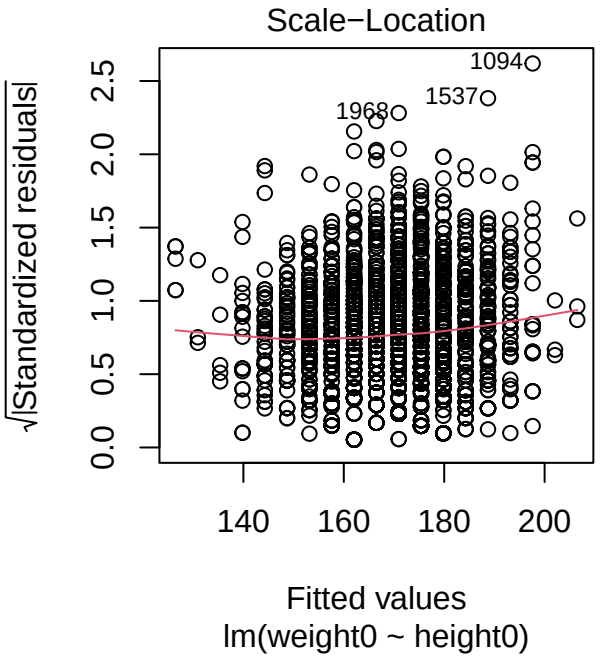
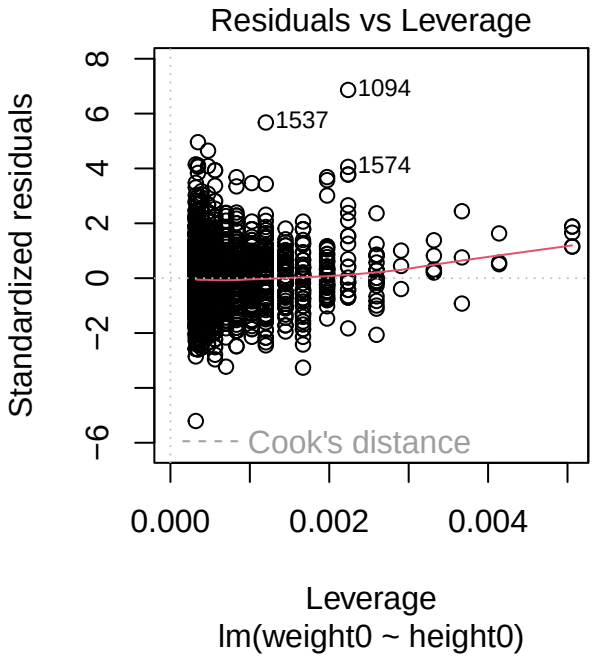


Figure 2: More detailed statistical scatter plot


Residuals vs Fitted
Residuals
Fitted values
lm(weight0 ~ height0)
1094
1537
1968


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(weight0 ~ height0)
1094
1537
1968


Scale−Location
√|Standardized residuals|
Fitted values
lm(weight0 ~ height0)
1094
1537
1968


Residuals vs Leverage
Standardized residuals
Leverage
lm(weight0 ~ height0)
1094
1537
1574
Cook's distance

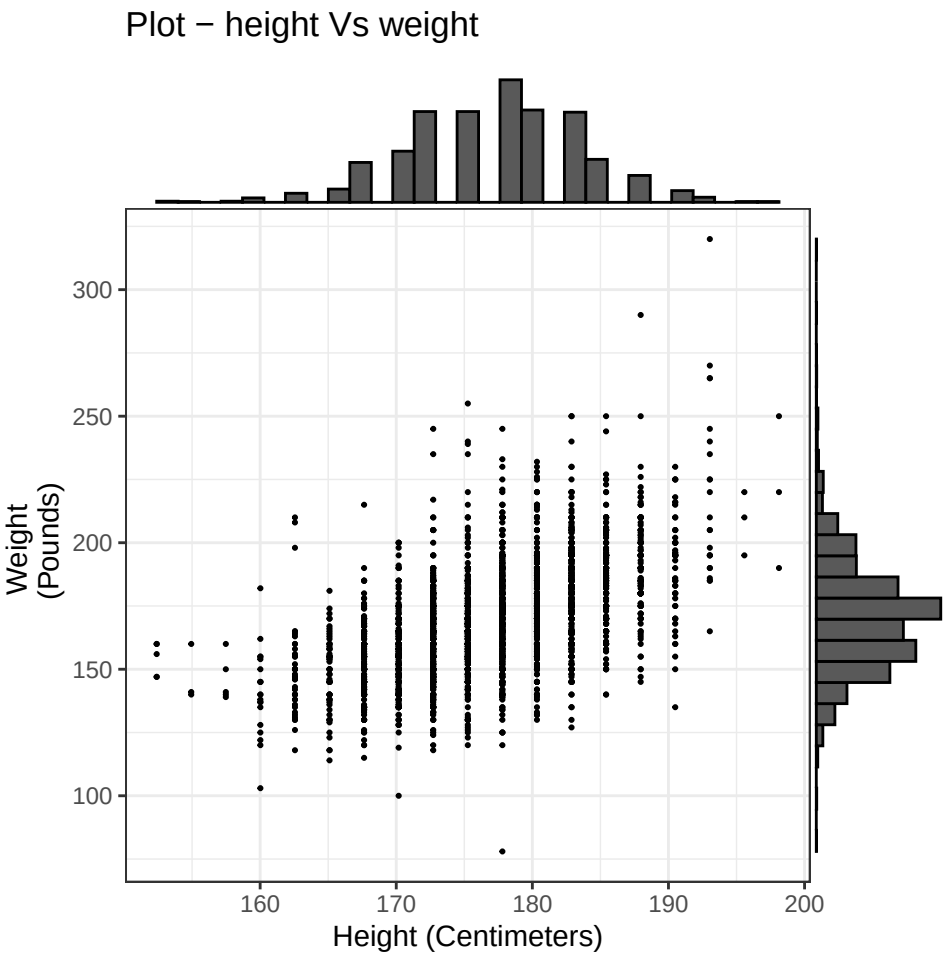


Figure 3: Height (cm) and weight scatter plot

Reference

1. Schober, Patrick MD, PhD, MMedStat; Boer, Christa PhD, MSc; Schwarte, Lothar A. MD, PhD, MBA. Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia & Analgesia 126(5):p 1763-1768, May 2018. | DOI: 10.1213/ANE.0000000000002864

Code Appendix

```
### Setting up the packages
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
# check if packages are installed; if not, install them
packages <- c("dplyr", "readr", "ggExtra", "plotly",
              "ggplot2","ggstatsplot","ggside","rigr","nlme","lmtest",
              "sandwich")
not_installed <- setdiff(packages, rownames(installed.packages()))
if (length(not_installed)) install.packages(not_installed)

# load packages
library(dplyr)
library(sandwich)
library(readr)
library(lmtest)
library(nlme)
library(ggstatsplot)
library(ggside)
library(rigr)
library(ggExtra)
library(plotly)
library(ggplot2)
# library(tidyverse) # don't load tidyverse package due to conflict with dplyr
### ------------------------------------------------------------
#Loading working directory of the raw data

#Please load your data/directory by changing it with your work directory
#Throughout this code module you will see a tone of places, where
#data is read and written, so please make sure to change them to your
#working directory folder format

working_directory_data <- setwd("C:/Users/latera/Desktop/Bio_ass")

wcgs <- read_csv("Data/wcgs.csv")
### ------------------------------------------------------------
### Q2
# fitting: linear regression with weight as the response variable
# and height as the predictor variable

weight_lm <- regress("mean", weight0 ~ height0, data = wcgs)
coef(weight_lm)[,c('Estimate','Naive SE', 'Robust SE','95%L',
                   '95%H','Pr(>|t|)')]
as.data.frame(coef(weight_lm)[,c('Estimate','Naive SE',
                   'Robust SE','95%L','95%H','Pr(>|t|)')])

coef(weight_lm)

#Generate summary statistics for Robut SE
tests<-coeftest(lm(weight0 ~ height0, data = wcgs),
                vcov=vcovHC(lm(weight0 ~ height0, data = wcgs),
                            type = "HC0"))
```

```
### ------------------------------------------------------------
### Q6
# fitting: linear regression with weight as the response
# variable and height as the predictor variable (height in centimeters)

#Assumption: 1 inch is 2.54 cm

wcgs$height0_cm <- (wcgs$height0) * 2.54
weight_lm_cm <- regress("mean", weight0 ~ height0_cm, data = wcgs)
coef(weight_lm_cm)[,c('Estimate','Naive SE',
                      'Robust SE','95%L','95%H','Pr(>|t|)')]
as.data.frame(coef(weight_lm_cm)[,c('Estimate','Naive SE',
                      'Robust SE','95%L','95%H','Pr(>|t|)')])

coef(weight_lm_cm)

tests_cm<-coeftest(lm(weight0 ~ height0_cm, data = wcgs),
                   vcov=vcovHC(lm(weight0 ~ height0_cm, data = wcgs),
                               type = "HC0"))



#Generating tables
kable(coef(weight_lm)%>%round(4), caption = "Linear fit of weight
      and height (inches) - OLS")

kable((tests)[,], caption = "Linear fit of weight and height
      (inches) - RSE")

kable(coef(weight_lm_cm)%>%round(4), caption = "Linear fit of weight
      and height (centimeters) - OLS ")

kable((tests_cm)[,], caption = "Linear fit of weight and height
      (centimeters) - RSE")

#Plotting scatter plot between weight and height (inches)

p <- ggplot(wcgs, aes(x=height0, y=weight0)) +
      geom_point(cex = 0.4)+
      theme_bw()+
      xlab("Height (Inches)") + ylab("Weight \n (Pounds)")+
      ggtitle("Plot - height Vs weight")+
      theme(legend.position="none")

# with marginal histogram
p1 <- ggMarginal(p, type="histogram")
p1

#Plotting detailed scatter plot
ggscatterstats(
  data = wcgs,
  x = height0,
```

```
  y = weight0,
  xlab="Height (Inches)",
  ylab="Weight \n (Pounds)",
  title = "Plot - height Vs weight",
  bf.message = FALSE,
  conf.level = 0.95,
  type = "robust",
  smooth.line.args = list(linewidth = 0.35,
                          color = "red",
                          method = "lm",
                          formula = y ~ x))

#checking OLS fit (residuals Vs fitted plots)
plot(lm(weight0 ~ height0, data = wcgs))


### Plotting scatter plot between weight and height (cm)

p_cm <- ggplot(wcgs, aes(x=height0_cm, y=weight0)) +
      geom_point(cex = 0.4)+
      theme_bw()+
      xlab("Height (Centimeters)") + ylab("Weight \n (Pounds)")+
      ggtitle("Plot - height Vs weight")+
      theme(legend.position="none")

# with marginal histogram
p1_cm <- ggMarginal(p_cm, type="histogram")
p1_cm
```