BIOST 515/518 Homework 2

Latera Tesfaye Olana

19 January, 2023

Responses:

1. I believe there is a linear relationship between forced expiratory volume (FEV) - liter per second and height inches. From the given sample data, with an increase in height (inches) there seems to be an increase in FEV. Accordingly, the first order trend suggestive of a tendency for higher average FEV in taller groups. As shown on Figure 1, there seems to be some suggestion of greater variability in FEV in taller groups than there is in the shorter groups. There are no striking outliers.
   In order to accurately determine or quantify the presumed linear relationship between these two variables, statistical tests such as Pearson correlation coefficient, can be applied. The results of this test suggests the data is consistent with a significant ($p<0.05$) a strong (1) linear correlation ($\hat{r}$ = 0.87, 95% CI [0.85 - 0.89]) between height and FEV.

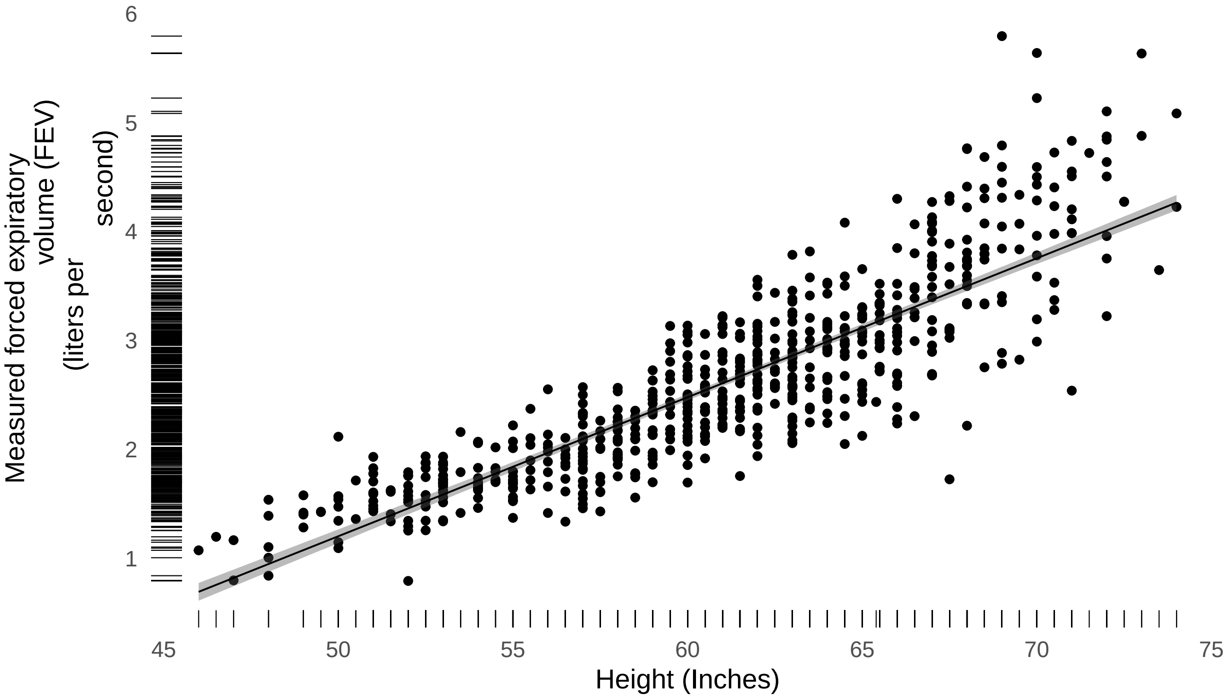


Figure 1: Scatter plot of FEV (liters per second) vs height (inches)

2. Based on a simple linear model, we estimate that the difference in mean forced expiratory volume (FEV) in a study group of 654 children differing by one inches in body height is 0.13 liters per second,

with the taller group having higher mean FEV (95% confidence interval [0.12; 0.14] liters per second). In a study of 654 children, we found evidence of an association between FEV and height ($p < 0.05$). Accordingly, we reject the null hypothesis of non linear trend in the expected value of FEV as a function of height in this study group.

3. Even though we have strong linear correlation (0.87), the association between FEV and height is not *perfectly linear*. Furthermore, it is evident that, while the average FEV generally exhibits an increasing trend, there may be instances of deviation whereby, between two groups, differing by one inches of height, the FEV may either increase or decrease. Figure 2, gives more descriptions.
   Regardless of our assumption about linearity the interpretation of the result will not change (as we are approximating best linear fit). Our interpretation and conclusion about the association between FEV - liters per second and height - inches, in the above question, did not imply anything about linear relationship.

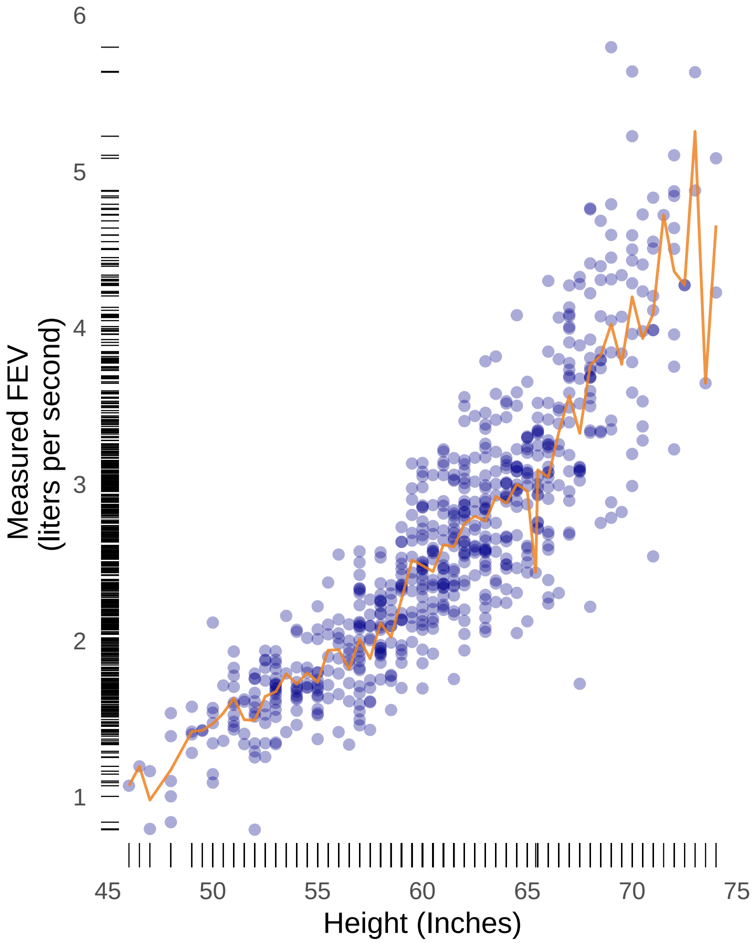


Figure 2: Plotting averaged trendline between groups differing by one inche

4. Let's see what log transformation on the dependent variable looks like:

$$\log(Fev_i) = \beta_0 + \beta_1 * height_i$$

$$FEV_i = e^{(\beta_0 + \beta_1 * height_i)}$$

The estimated difference in the geometric mean of forced expiratory volume (FEV) -liters per second for two groups differing in one inches is 1.0512711. In another word, for every group of children differing by one inche of height, we estimate that the geometric-mean of FEV in the taller children is 5.35 (95% confidence interval 5.12 to 5.58; $p<0.05$) % greater than the geometric-mean of FEV in the shorter children. The estimated expected value of forced expiratory volume (FEV) -liters per second for a children with a measured height of zero inches is 0.1033122 (95% confidence interval [0.09; 0.12]). *The intercept is just a mathematical construct allowing us to fit a line over the range of our data* and it might not be scientifically meaningful. In addition, intercept might not estimate a quantity of interest. After transforming the dependent variable (forced expiratory volume (FEV)) by natural log the linear correlation index increased from 0.87 to 0.89. The robust standard error decreased from $3.42 * 10^{-3}$ - standard error for the un-transformed to $1.12 * 10^{-3}$ - standard erro for the log transformed fit. Log transforming the dependent variable can help to linearize the relationship between the independent and dependent variables, which can improve the accuracy of the linear regression model. The QQ-plot on the supplementary plots, also shows how the plot between standard residuals and theoretical quantile changed. In addition, log transforming the dependent variable can also make outliers less influential on the linear regression model, since the log function tends to compress large values (uniform variance). Log transforming the dependent variable can make it more difficult to interpret the coefficients of the linear regression model, since the coefficients represent the change in the log of the dependent variable, rather than the change in the dependent variable itself. The figures in the supplementary provides general comparison of these two models.

5. The new formula would be:

$$\text{F}ev_i = \beta_0 + \beta_1 * (height_i - 45)$$

The estimated difference in expected value of FEV for two groups varying by one inches of height is 0.13 (95% confidence interval [0.13; 0.14]) liters per second. The interpretation drawn from the summary of this model remains invariant across different height groups, as the slope of the relationship remains constant, regardless of any alteration (shifting) in the independent variable. The estimated expected value of forced expiratory volume (FEV) for a children with a measured height of 45 inches is 0.51 (95% confidence interval [0.09; 0.12]). Changing of the independent variable (height) by constant factor only changes the estimated expected value of FEV for individuals with a zero unit of height (intercept $\hat{\beta_0}$). Where as, it does not change the difference in the estimated or expected mean of FEV between groups differing by one unit of height (slope or $\hat{\beta}$). In another word, the determination of the intercept is contingent upon the constants that have been added to the independent variable (height - inches). In addition, shifting our fitted model does not change the standard error and other accuracy metrics of the model. Figure 3, presents the scatter plot and robust fit (method: MM) of children above the height of 45 inches and their FEV liters per second.

6. The log transformed fit will give us the best prediction for a child with a height of 48 inches. This is mainly due, as shown on Figure 4 and 5, the log transformed fit has smaller residuals at height of 48 inches. In addition, the residual plot for the log transformed seems more *pattern-less* (random scatter of points forming an approximately constant width band around the identity line). Comparing the performance of both models globally, the log transformed fit has 0.8 coefficient of determination (R^2) value, where as the non-transformed has `round(r fev_lm$r.squared,2)`. Another way to compare the accuracy of these two linear models is to use the root mean square error (RMSE) value. The RMSE is the square root of the mean squared error, which is the average of the squares of the differences between the predicted values and the actual values. A lower RMSE value indicates a better fit of the model to the data and a higher accuracy. Accordingly, the log transformed fit has 0.15 RMSE, where as the non-transformed fit has 0.43.

7. In a study of 654 children, the estimated difference in the mean or expected value of FEV between female and male is 0.36 with female having smaller FEV (95% confidence interval [0.49; 0.23]) liters per second. We can reject the null hypothesis that mean FEV is the same in male and female children

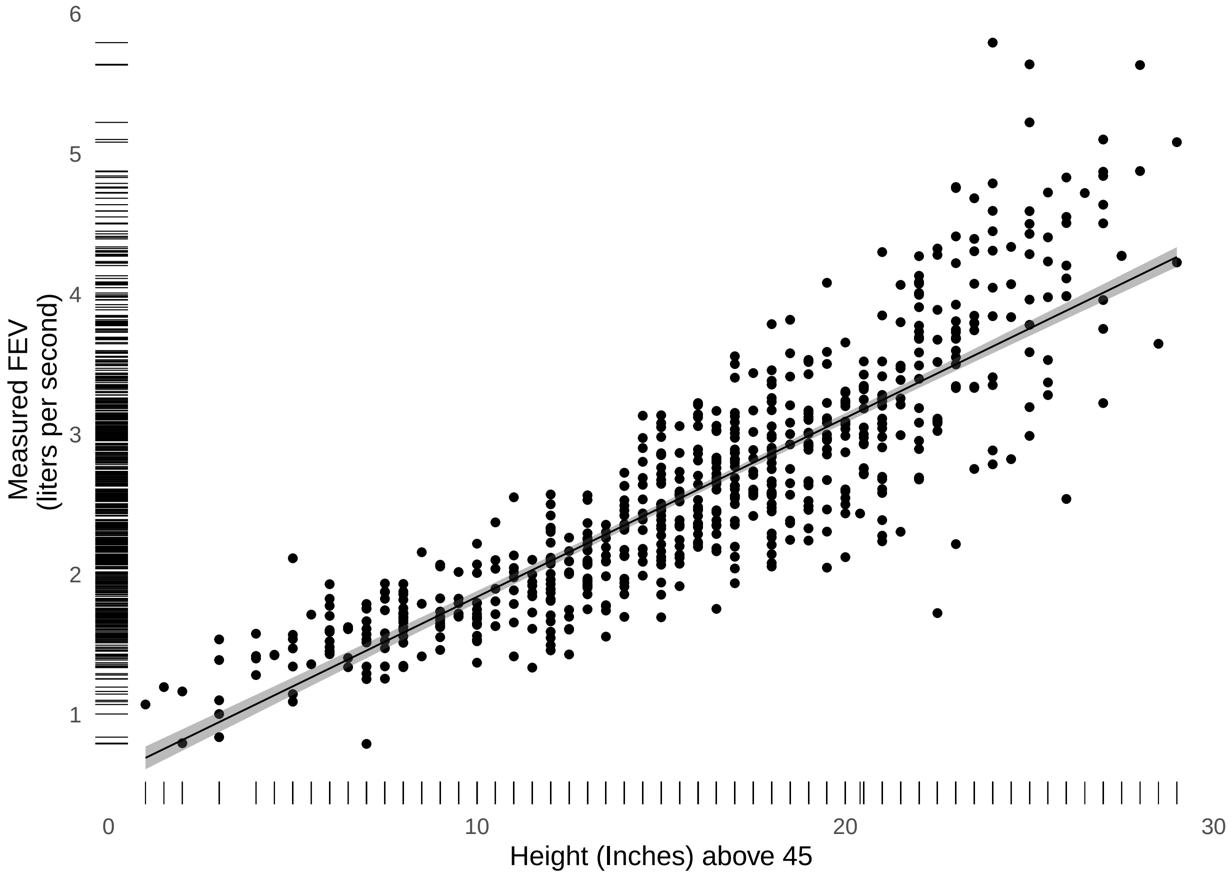


Figure 3: Scatter plot of FEV (liters per second) vs height (inches) - 45 above

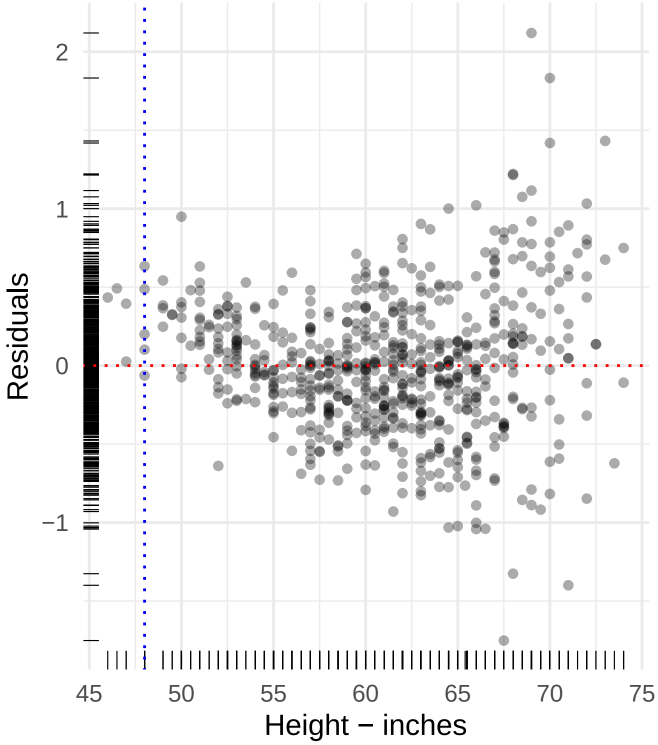


Figure 4: Residuals plot for FEV - liters per second and height - inches fit (no log transformation)

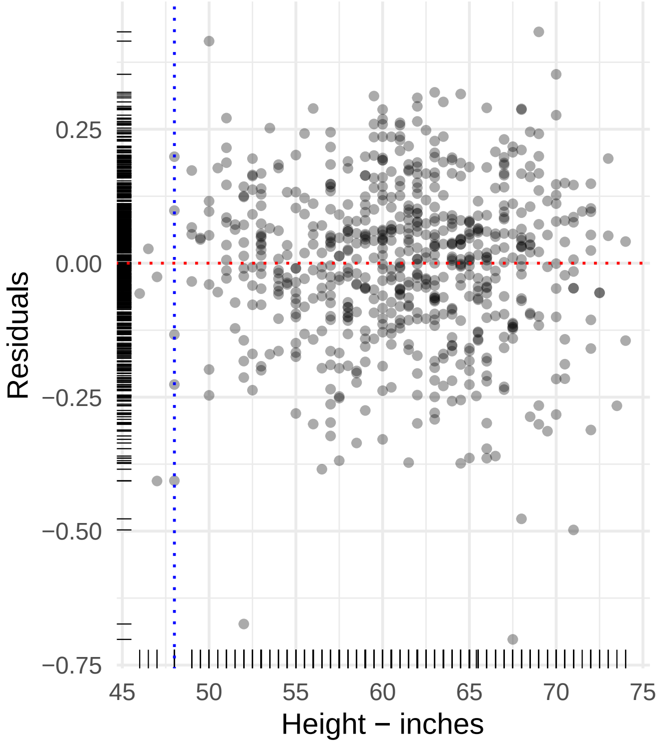


Figure 5: Residuals plot for FEV - liters per second and height - inches fit with log transformation

(p<0.05). The estimated mean FEV for female is 2.81 (95% confidence interval from 2.70 to 2.92) liters per second. Figure 8, provides a description on how the data is distributed and fitted across the given genders.

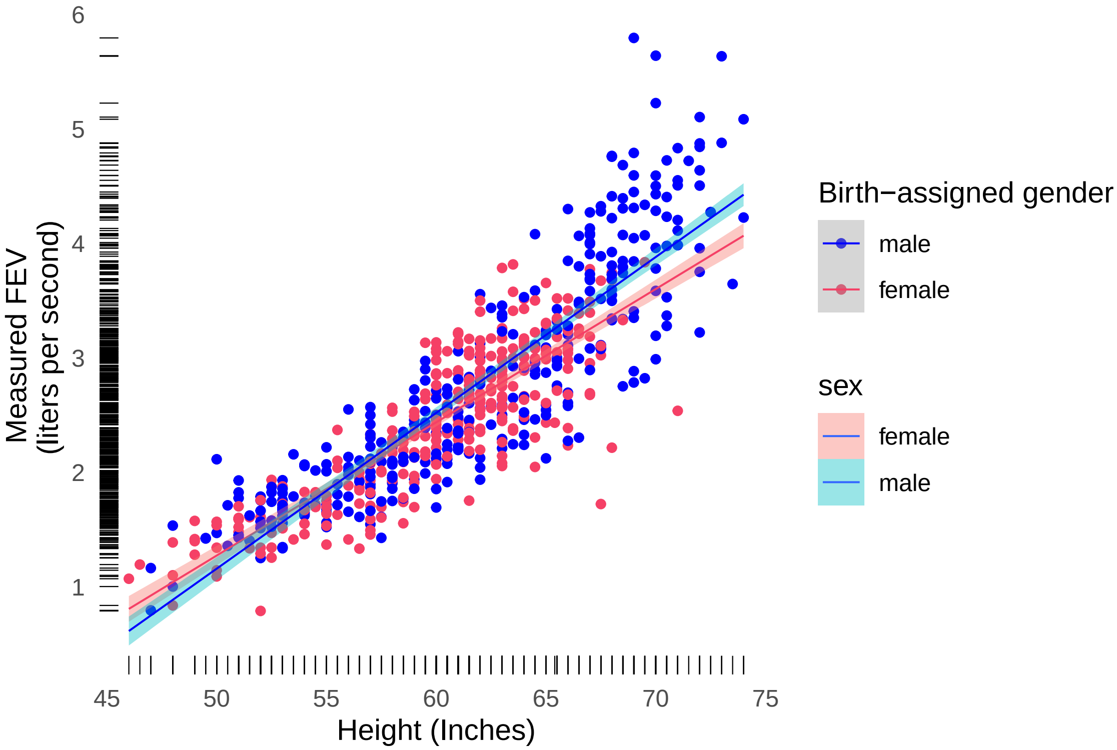


Figure 6: Scatter plot of FEV - liters per second vs height - inches disaggregated over gender

8. The above interpretation, might also apply. We can also rephrase it as follow: In a study of 654 children, the estimated mean difference in the mean or expected value of FEV between female and male is 0.36 liters per second with male having higher FEV [95% confidence interval [0.4902; 0.2324]. We can reject the null hypothesis that mean FEV is the same in male and female children (p<0.05). The estimated mean FEV for male is 2.45 (95% confidence interval from 2.38 to 2.52) liters per second.

Supplementary results - Tables

Table 1: Linear fit of FEV - liters per second and height inches

			Robust				
	Estimate	Naive SE	SE	95%L	95%H	t value	Pr(>|t|)
(Intercept)	-5.4327	0.1815	0.2008	-5.8270	-5.0383	-27.0523	0
height	0.1320	0.0030	0.0034	0.1253	0.1387	38.6446	0

Table 2: Linear fit of log transformed FEV - liters per second and height inches

			Robust				
	Estimate	Naive SE	SE	95%L	95%H	t value	Pr(>|t|)
(Intercept)	-2.2713	0.0635	0.0686	-2.4059	-2.1367	-33.1328	0
height	0.0521	0.0010	0.0011	0.0499	0.0543	46.4228	0

Table 3: Linear fit of FEV - liters per second and height inches (above 45)

		Naive	Robust				
	Estimate	SE	SE	95%L	95%H	t value	Pr(>|t|)
(Intercept)	0.5062	0.0506	0.0488	0.4103	0.6021	10.3678	0
height_above_45_inches	0.1320	0.0030	0.0034	0.1253	0.1387	38.6446	0

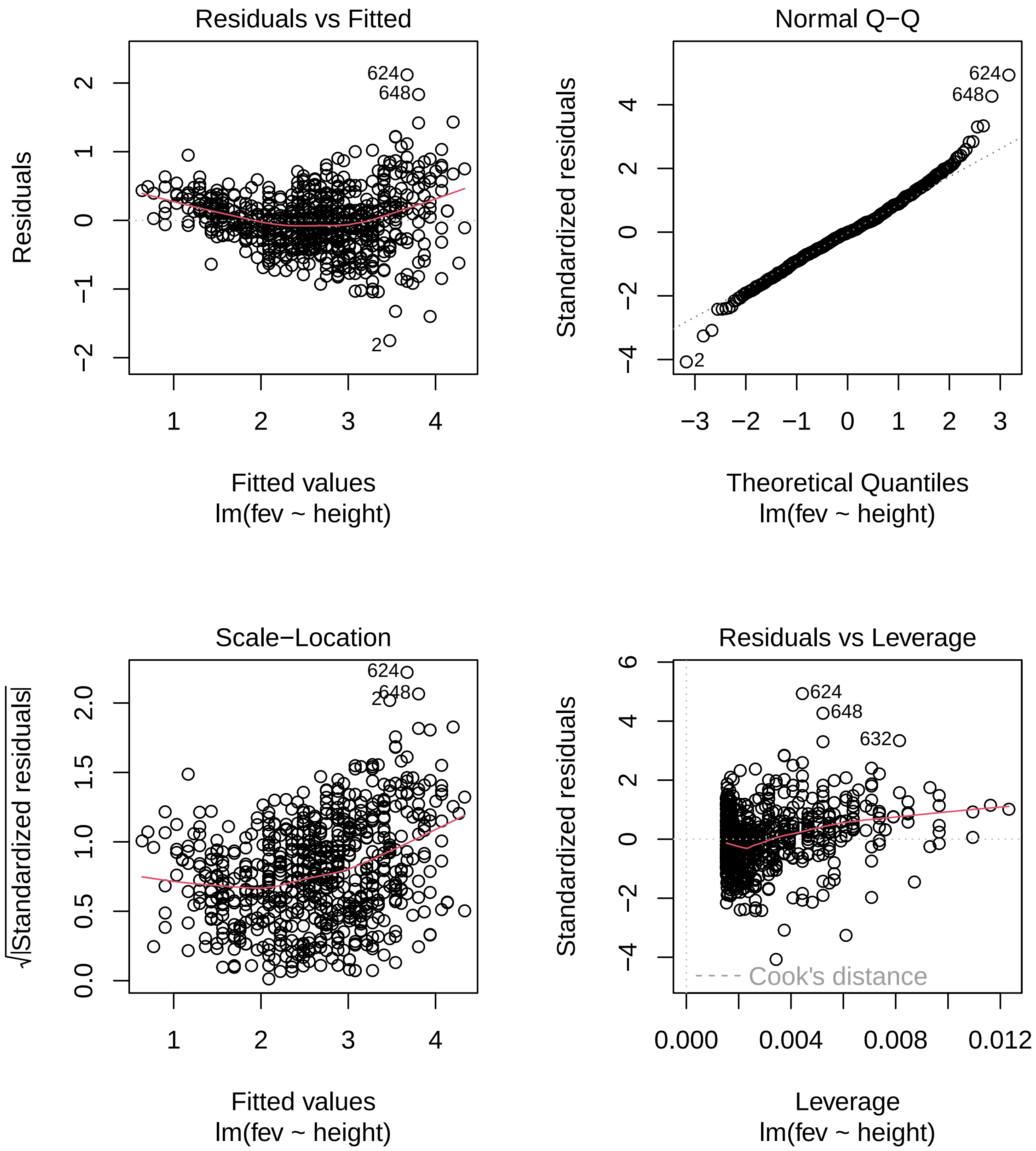
Table 4: Linear fit of FEV - liters per second and gender inches (female)

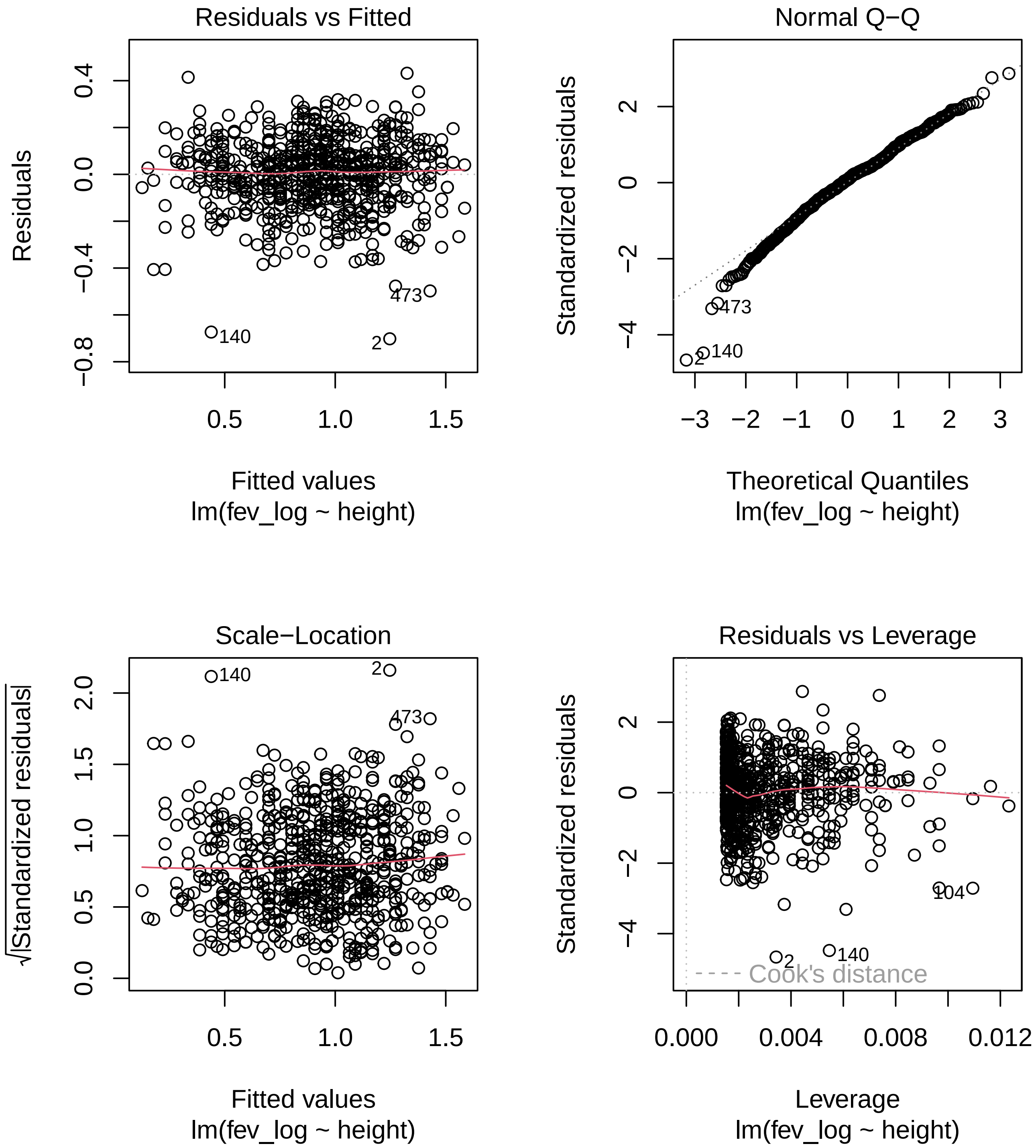
			Robust				
	Estimate	Naive SE	SE	95%L	95%H	t value	Pr(>|t|)
(Intercept)	2.8124	0.0463	0.0548	2.7049	2.9200	51.3661	0
female	-0.3613	0.0664	0.0656	-0.4902	-0.2324	-5.5036	0

Table 5: Linear fit of FEV - liters per second and gender inches (male)

	Estimate	Naive SE	Robust SE	95%L	95%H	t value	Pr(>|t|)
(Intercept)	2.4512	0.0476	0.0362	2.3801	2.5223	67.6941	0
male	0.3613	0.0664	0.0656	0.2324	0.4902	5.5036	0

Supplementary results - Plots




Residuals vs Fitted
Residuals
0.4
0.0
−0.4
−0.8
0.5
1.0
1.5
473
140
2
Fitted values
lm(fev_log ~ height)
Normal Q−Q
Standardized residuals
2
0
−2
−4
−3
−2
−1
0
1
2
3
473
2
140
Theoretical Quantiles
lm(fev_log ~ height)
Scale−Location
√|Standardized residuals|
2.0
1.5
1.0
0.5
0.0
0.5
1.0
1.5
140
2
473
Fitted values
lm(fev_log ~ height)
Residuals vs Leverage
Standardized residuals
2
0
−2
−4
104
2
140
Cook's distance
0.000
0.004
0.008
0.012
Leverage
lm(fev_log ~ height)

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("MASS")
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")

fev <- read_csv("Data/fev.csv")

#Plotting scatter plot
p <- ggplot(fev, aes(x=height, y=fev)) +
      geom_point(size=1.2)+
      xlab("Height (Inches)") + ylab("Measured forced expiratory
                                     volume (FEV) \n (liters per
                                     second)")+
  #scale_color_manual(name="Smoking Status",breaks=c('nonsmoker', 'smoker'),
    #   values=c('nonsmoker'='#409df4', 'smoker'='#f54066'))+
  geom_rug(col="black",linewidth=0.20)+
  theme_bw() +
  theme(axis.line = element_line(colour = "white"),
        axis.ticks = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())+
  theme(legend.background = element_rect
        (fill = "transparent"))

p+stat_smooth(method=function(formula,data,weights=weight)
  rlm(formula, data,weights=weight, method="MM"),
  fullrange=TRUE,color="black", fill="#555555",
  level=0.95,linewidth = 0.35)


### -----------------------------------------------------------
### Q2
# fitting: linear regression with weight as the response variable
# and height as the predictor variable

# Testing correlation between height and fev
cor.test(fev$height, fev$fev, method = "pearson")

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

coef(fev_lm)


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



fev$fev_log <- log(fev$fev)
fev_log_lm <- regress("mean", fev_log ~ height, data = fev)
coef(fev_log_lm)[,c('Estimate','Naive SE',
                      'Robust SE','95%L','95%H','Pr(>|t|)')]
as.data.frame(coef(fev_log_lm)[,c('Estimate','Naive SE',
                      'Robust SE','95%L','95%H','Pr(>|t|)')])

coef(fev_log_lm)


#Check correlation with new transformation
cor.test(fev$height, fev$fev_log, method = "pearson")

fev$height_above_45_inches = fev$height - 45

fev_45_above <- regress("mean", fev ~ height_above_45_inches, data = fev)
coef(fev_45_above)[,c('Estimate','Naive SE',
                      'Robust SE','95%L','95%H','Pr(>|t|)')]
as.data.frame(coef(fev_45_above)[,c('Estimate','Naive SE',
                      'Robust SE','95%L','95%H','Pr(>|t|)')])
```

```
coef(fev_45_above)
mean_data <- group_by(fev, height) %>%
             summarise(fev_new = mean(fev))

ggplot(data=fev, aes(x=height, y=fev)) +
  geom_point(data=fev, aes(x=height, y=fev), col='darkblue',
             alpha = 1/3) +
  geom_line(data=mean_data, aes(x = height, y = fev_new), col='#ee8324',
            alpha=0.85)+
  xlab("Height (Inches)") + ylab("Measured FEV \n (liters per second)")+
  geom_rug(col="black",linewidth=0.20)+
  theme_bw() +
  theme(axis.line = element_line(colour = "white"),
        axis.ticks = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())

#Plotting detailed scatter plot
p <- ggplot(fev, aes(x=height_above_45_inches, y=fev)) +
      geom_point(size=1.2)+
      xlab("Height (Inches) above 45") + ylab("Measured FEV \n (liters per second)")+
  #scale_color_manual(name="Smoking Status",breaks=c('nonsmoker', 'smoker'),
   #    values=c('nonsmoker'='#409df4', 'smoker'='#f54066'))+
  geom_rug(col="black",linewidth=0.20)+
  theme_bw() +
  theme(axis.line = element_line(colour = "white"),
        axis.ticks = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())+
  theme(legend.background = element_rect
        (fill = "transparent"))

p+stat_smooth(method=function(formula,data,weights=weight) rlm(formula,
                                                               data,
                                                               weights=weight,
                                                                method="MM"),
              fullrange=TRUE,color="black", fill="#555555", level=0.95,linewidth = 0.35)


ggplot(fev, aes(x=height, y=resid(fev_lm))) +
      geom_point(size=1.2,alpha = 1/3)+
      xlab("Height - inches") + ylab("Residuals")+
  geom_rug(col="black",linewidth=0.20)+
  theme_bw() +
  theme(axis.line = element_line(colour = "white"),
        axis.ticks = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())+
  theme(legend.background = element_rect
        (fill = "transparent"))+
  geom_hline(yintercept = 0, color="red",
             linetype="dotted",)+
  geom_vline(xintercept = 48, color="blue",
             linetype="dotted")
ggplot(fev, aes(x=height, y=resid(fev_log_lm))) +
      geom_point(size=1.2,alpha = 1/3)+
      xlab("Height - inches") + ylab("Residuals")+
  geom_rug(col="black",linewidth=0.20)+
  theme_bw() +
  theme(axis.line = element_line(colour = "white"),
        axis.ticks = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())+
  theme(legend.background = element_rect
        (fill = "transparent"))+
  geom_hline(yintercept = 0, color="red",linetype="dotted")+
  geom_vline(xintercept = 48, color="blue",linetype="dotted")
fev<-fev %>%
  mutate(female = case_when(sex == "female" ~ 1,
                                 sex == "male" ~ 0))
fev<-fev %>%
  mutate(male = case_when(sex == "male" ~ 1,
                                 sex == "female" ~ 0))
fev_lm_gender <- regress("mean", fev ~ female, data = fev)
coef(fev_lm_gender)[,c('Estimate','Naive SE',
                      'Robust SE','95%L','95%H','Pr(>|t|)')]
as.data.frame(coef(fev_lm_gender)[,c('Estimate','Naive SE',
                      'Robust SE','95%L','95%H','Pr(>|t|)')])

coef(fev_lm_gender)


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

coef(fev_lm_gender_male)

#Plotting detailed scatter plot
p <- ggplot(fev, aes(x=height, y=fev, color=sex)) +
      geom_point(size=1.2)+
      xlab("Height (Inches)") + ylab("Measured FEV \n (liters per second)")+
  scale_color_manual(name="Birth-assigned gender",breaks=c('male', 'female'),
       values=c('male'='blue', 'female'='#f54066'))+
  geom_rug(col="black",linewidth=0.20)+
  theme_bw() +
  theme(axis.line = element_line(colour = "white"),
        axis.ticks = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())+
  theme(legend.background = element_rect
        (fill = "transparent"))

p+stat_smooth(aes(fill=sex),method=function(formula=y~x,data,weights=weight) rlm(formula,
                                                               data,
                                                               weights=weight,
                                                                method="MM"),
              fullrange=TRUE, level=0.95,linewidth = 0.35)

#Generating tables

kable(coef(fev_lm)%>%round(4), caption = "Linear fit of FEV - liters
per second and height inches")

kable(coef(fev_log_lm)%>%round(4), caption = "Linear fit of log
      transformed FEV - liters per second and height inches")

kable(coef(fev_45_above)%>%round(4), caption = "Linear fit of
      FEV - liters per second and height inches (above 45)")

kable(coef(fev_lm_gender)%>%round(4), caption = "Linear fit of
      FEV - liters per second and gender inches (female)")

kable(coef(fev_lm_gender_male)%>%round(4), caption = "Linear fit of
      FEV - liters per second and gender inches (male)")

#checking OLS fit (residuals Vs fitted plots)
plot(lm(fev ~ height, data = fev))
plot(lm(fev_log ~ height, data = fev))

#abline(0, 0)
```