Apply simple linear regression model on FEV prediction in children with FEV dataset

Eliza Chai

01/19/2022

Responses

1. Based on figure 1 below, the data appears to show a linear relationship between average FEV and height for most of the range of observed heights. The FEV values appear to be positively correlated with higher heights. We could fit a straight line fit with a positive slope that captures most of the variability in FEV across different values of height. However, we observe some evidence of non-linearity for the shortest and tallest children in the FEV cohort.

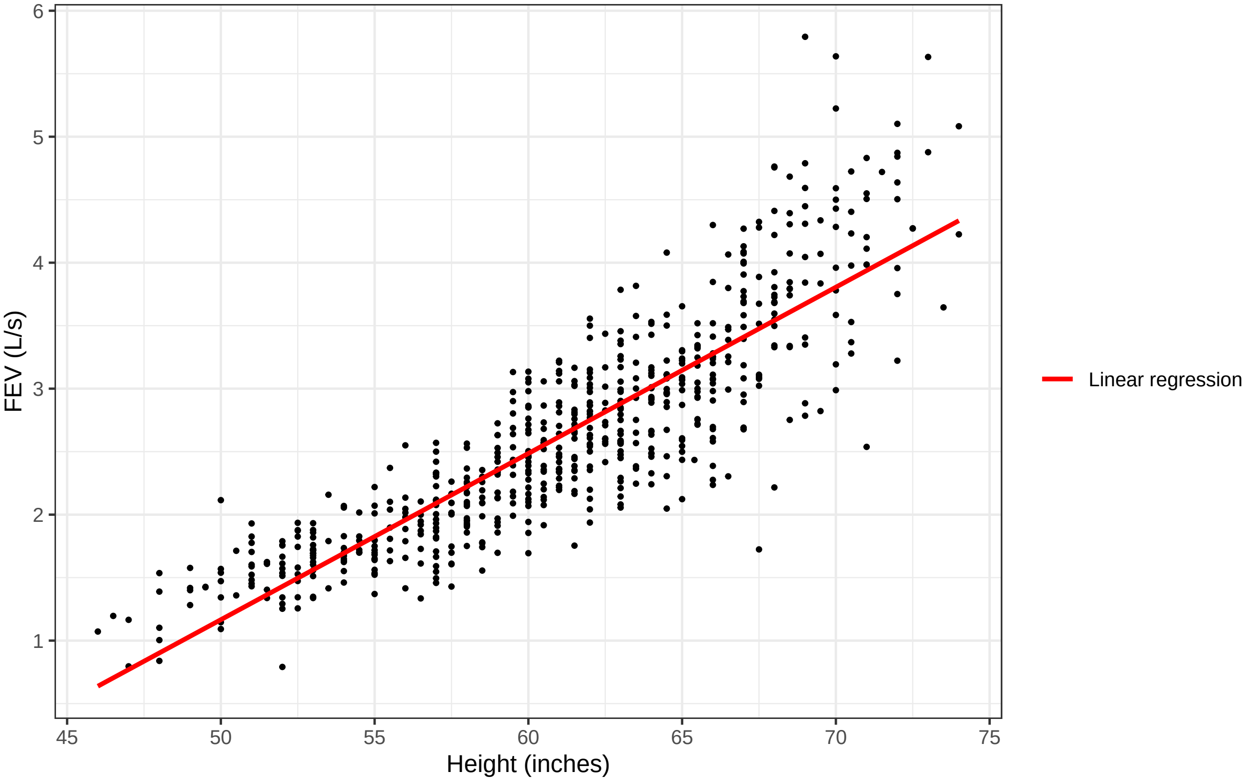


Figure 1: Scatterplot of FEV versus height for children in the FEV cohort

2. Based on a simple linear regression, we estimate that the difference in the mean FEV between two populations differing by one inch in height is 0.13 L/s, with the taller group having higher FEV values

(95% CI based on heteroskedasticity-robust standard errors: 0.13, 0.14). The intercept is -5.43 L/s at zero inches in height, which is an extrapolation from the range of the observed data, and it is not of scientific interest. We reject the null hypothesis of no association (p = < 0.001) and conclude that we have very strong evidence of a first-order linear association between FEV and height.

3. I do not believe the relationship between height and average FEV is exactly linear. From figure 1, we see a positive linear relationship between average FEV and height for most of the range of observed heights. However, we see some evidence of non-linearity for the shortest and tallest children. From figure 2, comparing between the Lowess smoother and linear regression line, we observe a curved smoother due to non-linearity at the smallest and largest height. Since simple linear regression summarizes the average linear trend between predictor and response regardless of whether the relationship is exactly linear, we can use the simple linear regression model to obtain estimators and their associated standard errors. We can conclude the parameters of the population of interest with SLR but note that predictions would be inaccurate if the relationship is not truly linear.

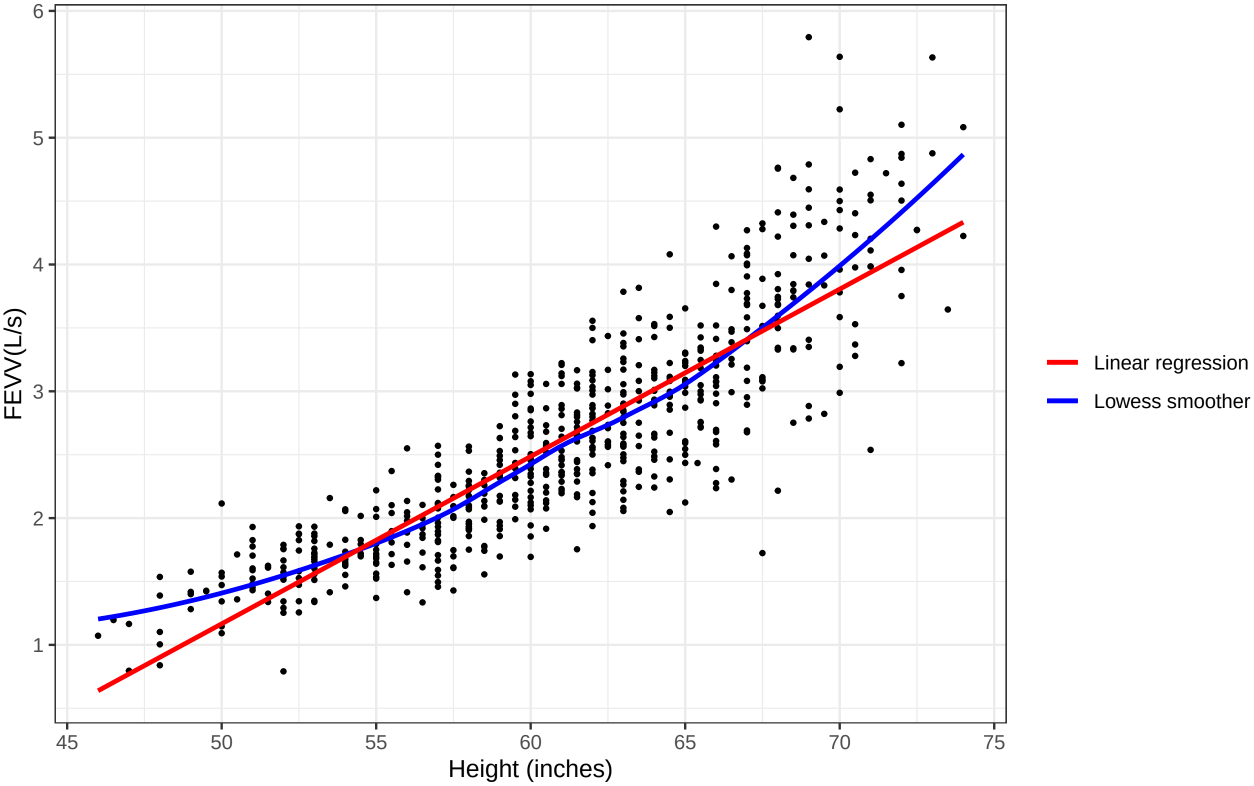


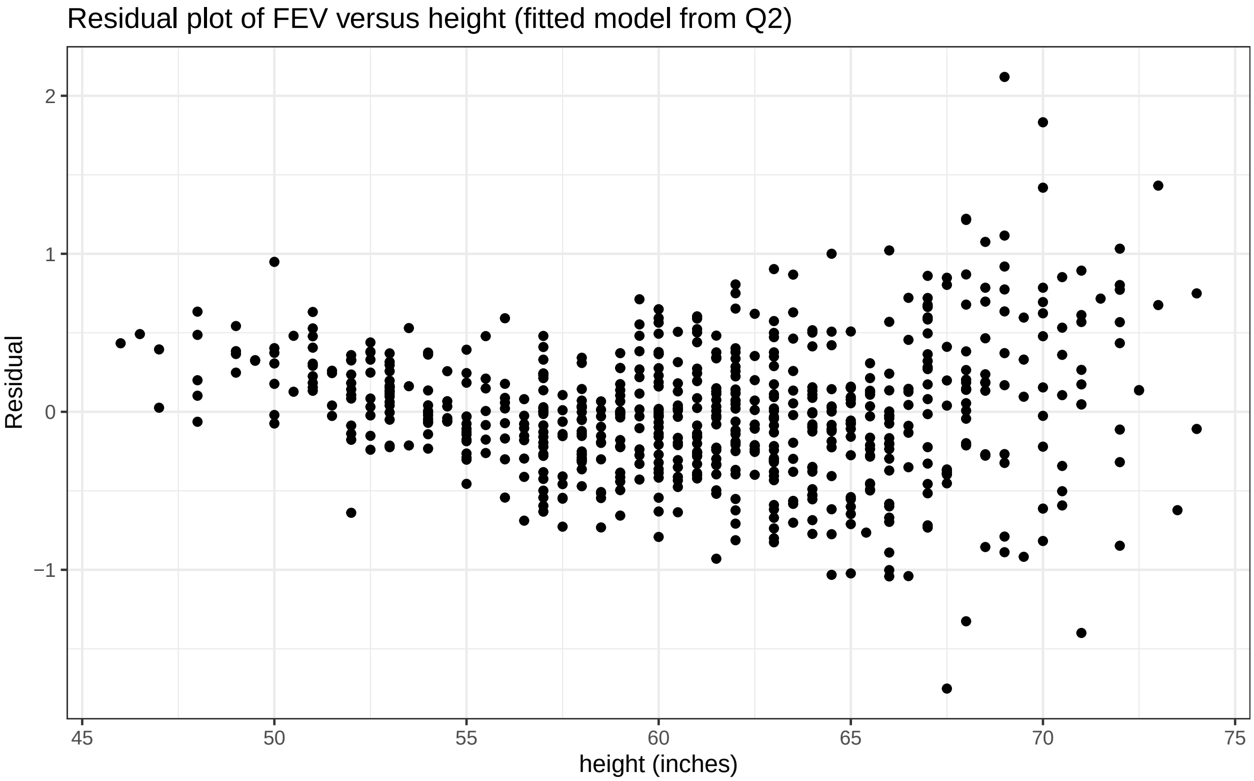
Figure 2: Scatterplot of FEV versus height for children in the FEV cohort

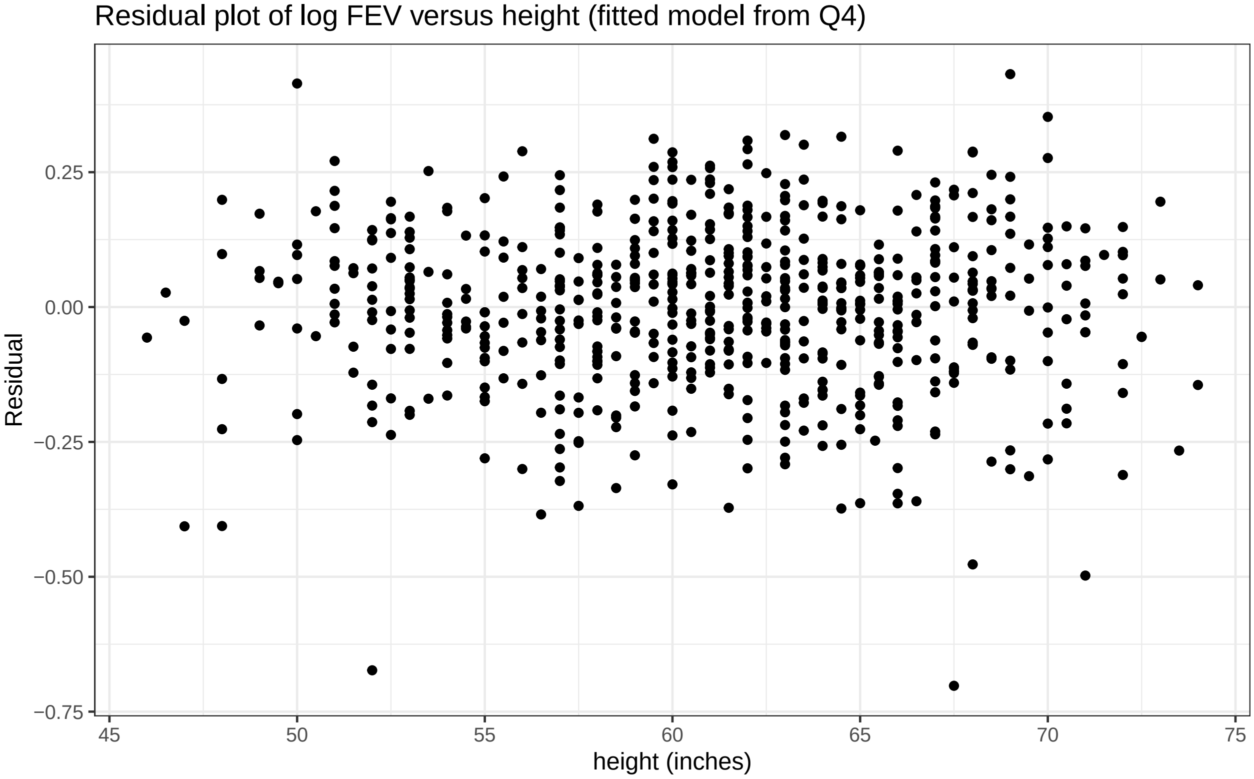
4. Based on a simple linear regression, we estimate that the difference in the mean log FEV between two populations differing by one inch in height is 0.052 log L/s, with the taller group having higher mean log FEV (95% CI based on heteroskedasticity-robust standard errors: 0.050, 0.054). (geometric-mean of FEV in the higher population is 5% greater than the geometric-mean of FEV in the shorter population). The intercept is -2.271 log L/s at zero inches in height, which is an extrapolation from the range of the observed data, and it is not of scientific interest.

5. Based on a simple linear regression, we estimate that the difference in the mean FEV between two populations differing by one inch taller than 45 inches in height is 0.13 L/s, with the taller group having higher mean FEV (95% CI based on heteroskedasticity-robust standard errors: 0.13, 0.14). The intercept is 0.51 L/s at 0 inches taller than 45 inches in height. Comparing this model with the

model fitted in question 2, we observe a leftward shift in the x-axis and a change in the intercept as we deducted 45 inches from all heights. The new fitted model gives us the same slope of 0.13 L/s and 95% CI based on the robust standard error, however, the intercept in the new model is of scientific interest. We conclude that a deduction in the predictor variable in a simple linear regression does not change the conclusions of our analysis, but could give more information on the intercept parameter.

6. From the residual plot of the fitted model in question 2, we see the model underestimates the FEV of children with shorter height in the range of 45 to 50-inch tall child as the average residuals in the range are above zero. This may point to a poorer model fit. From the residual plot of the fitted model in question 4, we see the model has an average residual closer to zero in the range of 45 to 50-inch tall children. Therefore, the model in question 4 would give a better prediction of the FEV of a 48-inch tall child.





7. Based on a simple linear regression, we estimate that the mean FEV for male children is 2.81 L/s. Furthermore, we estimate that the difference in the mean FEV of male group compared with female group is 0.36 L/s, with the male group having higher mean FEV.

8. Based on a simple linear regression, we estimate that the mean FEV for female children is 2.45 L/second. Furthermore, we estimate that the difference in the mean FEV of female children compared with male children is 0.36 L/second, with the male group having higher mean FEV. Comparing the fitted value under model from question 8 to that of question 7, we see the parameter beta 1 remains the same, but beta 0 is smaller (or larger in magnitude) in the fitted model of question 8. Therefore, we can conclude that a switch in the indication of sex between females and males does not change the conclusions of our analysis of the slope parameter, but would affect the conclusions of the intercept parameter.

Code Appendix

```
### Setting up the packages, options we'll need:
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse)
library(rigr)
### -----------------------------------------------------------
### Reading in the data.
fev <- read_csv("fev.csv")
### -----------------------------------------------------------
### Q1
### Plot a scatterplot of FEV on the vertical axis and height on the horizontal axis
fev %>%
  ggplot(aes(x = height, y = fev)) +
  geom_point(cex = 0.7) +
  xlab("Height (inches)") +
  ylab("FEV (L/s)") +
  theme_bw() +
  geom_smooth(aes(col = "Linear regression"), se=F, method = "lm", show.legend = T) +
  scale_color_manual(name = "",
                     values = c("Linear regression"="red")) + NULL
### -----------------------------------------------------------
### Q2
### Here's a linear model for fev dataset using rigr:regress
fev_mod1 <- regress("mean", fev ~ height, data = fev)
fev.ci <- fev_mod1 %>%
  coef() %>%
  as.data.frame() %>%
  select(c("Estimate", "Naive SE", "Robust SE", "95%L", "95%H",
           "Pr(>|t|)")) %>% round(4)
fev.ci
### -----------------------------------------------------------
### Q3
### Plot the scatter plot with lowess smoother and slr line
fev %>%
  ggplot(aes(x = height, y = fev)) +
  geom_point(cex = 0.7) +
  xlab("Height (inches)") +
  ylab("FEVV(L/s)") +
  theme_bw() +
  geom_smooth(aes(col = "Lowess smoother"), se=F, method = "loess", show.legend=T)+
  geom_smooth(aes(col = "Linear regression"), se=F, method = "lm", show.legend = T) +
  scale_color_manual(name = "",
                     values = c("Lowess smoother"="blue",
                                "Linear regression"="red")) +
NULL
### -----------------------------------------------------------
### Q4
### Here's a linear model with log fev using rigr:regress
fev_log <- fev %>%
  mutate(log_fev = log(fev))
```

```r
fev_mod2 <- regress("mean", log_fev ~ height, data = fev_log)
fev_mod2
fev.log.ci <- (fev_mod2 %>% coef)["height",
                                  c("Estimate", "Naive SE", "Robust SE", "95%L",
                                    "95%H","Pr(>|t|)")] %>% round(4)
fev.log.ci

### Another way to make SLM for fev dataset using rigr:regress:geometric mean
fev_lm <- regress("geometric mean", fev ~ height, data = fev)
### -----------------------------------------------------------
### Q5
### Here's a linear model with fev ~ height above 45 inches using rigr:regress
fev_above45 <- fev %>%
  mutate(height_ab45 = height-45)

fev_mod3 <- regress("mean", fev ~ height_ab45, data = fev_above45)

fev_mod3

fev.ab45.ci <- (fev_mod3 %>% coef)["height_ab45",
                                   c("Estimate", "Naive SE", "Robust SE", "95%L",
                                     "95%H","Pr(>|t|)")] %>% round(4)
fev.ab45.ci
### -----------------------------------------------------------
### Q6
### Residual plot of Q2 fev vs. height
mod1_resids <- fev_mod1 %>% residuals
fev_mod1_resids <- fev %>% mutate(mod1_resids)
fev_mod1_resids %>% ggplot(aes(height, mod1_resids)) +
  geom_point() +
  theme_bw() +
  xlab("height (inches)") +
  ylab("Residual") +
  ggtitle(label = "Residual plot of FEV versus height (fitted model from Q2)")

### Residual plot of Q4 log fev vs. height
mod2_resids <- fev_mod2 %>% residuals
fev_mod2_resids <- fev %>% mutate(mod2_resids)
fev_mod2_resids %>% ggplot(aes(height, mod2_resids)) +
  geom_point() +
  theme_bw() +
  xlab("height (inches)") +
  ylab("Residual") +
  ggtitle(label = "Residual plot of log FEV versus height (fitted model from Q4)")
### -----------------------------------------------------------
### Q7
### Here's a linear model with fev ~ height + female using rigr:regress
fev_female <- fev %>%
  mutate(female = ifelse(sex == "female", 1, 0))
fev_mod4 <- regress("mean", fev ~ female, data = fev_female)
fev.female.ci <- fev_mod4 %>% coef %>% round(4)
fev.female.ci
### -----------------------------------------------------------
### Q8
### Here's a linear model with fev ~ height + male using rigr:regress
fev_male <- fev %>%
  mutate(male = ifelse(sex == "male", 1, 0))
fev_mod5 <- regress("mean", fev ~ male, data = fev_male)
fev.male.ci <- fev_mod5 %>% coef %>% round(4)
fev.male.ci
### -----------------------------------------------------------
```