BIOST 515 Homework #2

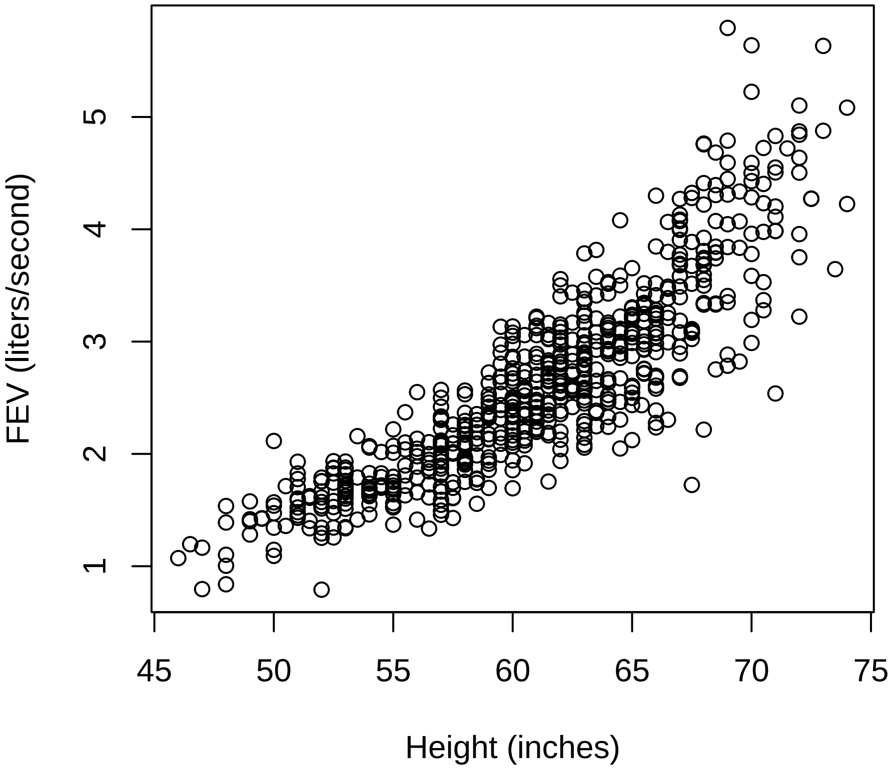
Katie McFarlane

1/21/22

Responses

1. Make a scatterplot with FEV on the vertical axis and height on the horizontal axis. Describe the relationship that you see.

There is a positive linear relationship between height and FEV in this plot: we see an increase in FEV as height increases. We also note that among those with larger height values, the spread of FEV values is larger in comparison to the spread of FEV values among those who smaller height values, indicating that kids in this study who are taller have a larger variance in their FEV values than participants in the study who are shorter.



2. Perform a simple linear regression of FEV on height, and test the null hypothesis that there is no linear trend in the expected value of FEV as a function of height. In 2-3 sentences, summarize the results in language suitable for a scientific publication.

The estimated difference in mean FEV for groups of children differing by one inch in height is 0.132 liters/second, with the taller group having higher levels of FEV (95% CI 0.125, 0.139). The simple linear regression is fitted through the point representing both mean height and mean FEV value in the study, which is observed at 61.14 inches, 2.64 liters/second. We observe a p-value < 0.0001, so we reject the null hypothesis that there is no linear trend in expected value of FEV as a function of height in children.

3. Do you believe that the relationship between height and average FEV is exactly linear? Does this compromise your interpretation and conclusions in Question 2?

I do not believe that the relationship is exactly linear: for a perfectly linear relationship, we would expect a correlation value of -1 or 1, but we observe a correlation of 0.87 between height and FEV. However, we still observe a generally linear trend in the data. The response in question two explains the general linear trend across the data, and this interpretation is still correct, even if it cannot account for every single data point within the dataset.

4. Perform a simple linear regression with the logarithm (base e) of FEV as the response variable and height as the predictor variable. Interpret the fitted parameters of your model.

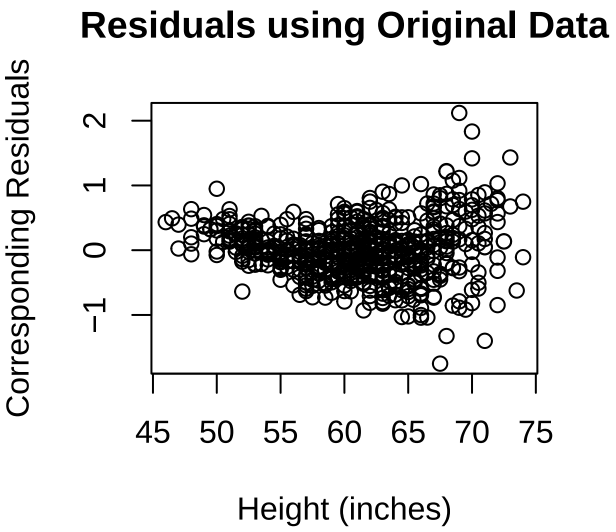
Using a simple linear regression model, we estimate that for two populations of children who differ by one inch of height, the taller population has 1.054-fold (95% CI 1.051, 1.056) greater geometric mean FEV value than the geometric mean of those who are in the shorter population. The y-intercept of 0.103 liters/second represents e to the power of the mean of log FEV for a child who is zero inches tall; yet no such child exists, thus this is not scientifically relevant to our interpretation.

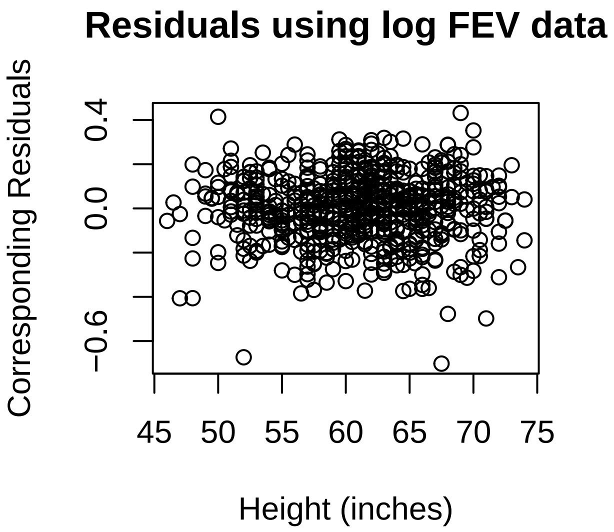
5. Fit a simple linear regression model for FEV on height-above-45-inches. That is, calculate how many inches taller than 45" each participant is and use this as the predictor variable in your model. Interpret the fitted parameters of your model and comment on the differences between this model and the model that you fit in Question 2.

The estimated difference in mean FEV for groups of children differing by one inch in height is 0.132 liters/second, with the taller group having higher levels of FEV (95% CI 0.125, 0.139). We note a y-intercept of 0.506, which is our estimate of the expected FEV value in liters/second of a child who is 45 inches. Our slope paramater estimate is identical to what we observed in question 2, but our y-intercept is different, as our data was shifted so that a height of 45 inches on the original plot is now equivalent to what is portrayed as an x value (height over 45 inches) of 0 in this question.

6. Obtain (or calculate) the residuals from the models that you fit in Questions 2 and 4, and make a plot for each model showing the residuals against the heights. Which model do you think would give a better prediction of the FEV of a 48 inch tall child? Justify your answer.

The model in question 4 would give a better prediction of the FEV of a 48 inch tall child – visually, we can see that the second plot has residuals that are on average smaller than those in the first plot. We can also calculate the mean residuals for a 48 inch tall child in the second and fourth questions, which are respectively 0.27 and -0.09. The absolute value of the residual for the model using log transformed FEV values is roughly 1/3 the size of that from the original model, meaning that data is more closely fitted.





7. Fit the given model for FEV, where female_i=1 if participant i is female and female_i=0 if participant i is male. Interpret the estimated parameters of your model.

The estimated difference in mean FEV for groups of children stratified by sex is -0.36 liters/second, with male children having higher levels of FEV on average (95% CI -0.49, -0.23). Male children have an average FEV value of 2.81 liters/second.

8. Now fit the given model for FEV, where male_i=1 if participant i is male and male_i=0 if participant i is female. Interpret the estimated parameters of your model. Compare the fitted values under this model to the fitted values from the model that you fit in Question 7, and comment.

The estimated difference in mean FEV for groups of children stratified by sex is 0.36 liters/second, with male children having higher levels of FEV on average (95% CI 0.23, 0.49). Female children have an average FEV value of 2.45 liters/second. The estimated slope parameter is the same as in the previous question with the sign reversed. Because sex is a binary variable in this dataset, it makes sense that the slope parameter would be identical, given the male and female variables are the same, just with swapped values. We also note a different y-intercept, with a difference in y-intercepts of 0.36. This also makes sense, given the slope is connecting the mean FEV value for female children to the mean FEV value for male children.

Code Appendix

```r
### Setting up the packages, options we'll need:
library(knitr)
library(rigr)
knitr::opts_chunk$set(echo = FALSE)
### ------------------------------------------------------------
library(tidyverse)
fev <- read.csv("~/Graduate School Work/Winter 2022 - BIOST 515/fev.csv")
### ------------------------------------------------------------
### Q1
plot(fev~height, xlab="Height (inches)", ylab= "FEV (liters/second)", data=fev)
### ------------------------------------------------------------
### Q2
regress("mean", fev~height, data=fev)
mean(fev$height)
mean(fev$fev)
### ------------------------------------------------------------
### Q3
cor(fev$fev, fev$height)
### ------------------------------------------------------------
### Q4
regress("geometric mean", fev~height, data=fev)
### ------------------------------------------------------------
### Q5
fev$height45 = fev$height - 45
regress("mean", fev~height45, data=fev)
### ------------------------------------------------------------
### Q6
lin <- lm(fev~height, data=fev)
x<- resid(lin)
fev$res <- x
plot(x~fev$height)

fev$lnfev = log(fev$fev)
lin2 <-lm(lnfev~height, data=fev)
y <- resid(lin2)
fev$lnres <- y
plot(y~fev$height)

mean(fev[fev$height==48, "res"])
mean(fev[fev$height==48, "lnres"])
plot(x~fev$height, main="Residuals using Original Data", xlab="Height (inches)",
     ylab="Corresponding Residuals")
plot(y~fev$height, main="Residuals using log FEV data", xlab="Height (inches)",
     ylab="Corresponding Residuals")
### ------------------------------------------------------------
### Q7
fev$female = as.integer(fev$sex=="female")
regress("mean", fev~female, data=fev)
mean(fev[fev$female==1, "fev"])

### ------------------------------------------------------------
### Q8
fev$male = as.integer(fev$sex=="male")
regress("mean", fev~male, data=fev)
mean(fev[fev$male==1, "fev"])
```