Apply simple linear regression model to investigate the relationship between weight/height and age in WCGS dataset

Eliza Chai

01/10/2022

Responses

1. Based on the scatterplot below, the data appears to show a linear relationship between weight and height in this population. The weight appears to be positively correlated with higher heights. We could fit a straight line fit with a positive slope that capture most of the variability in weight across different values of height.

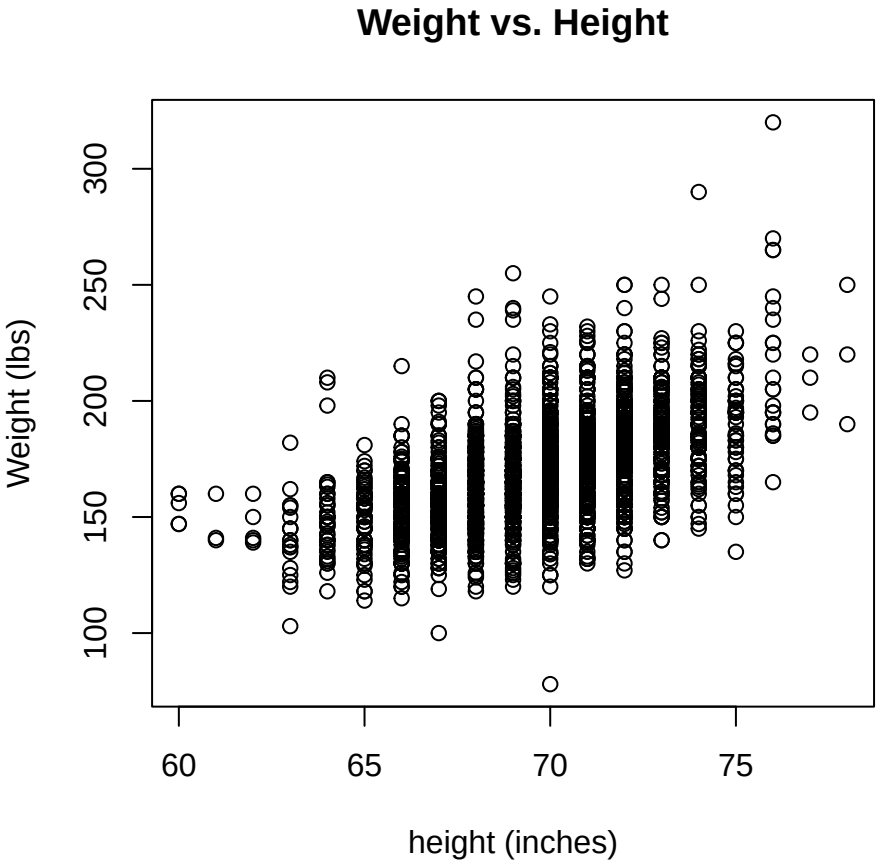

Figure 1: Weight vs. Height

2. We estimate that when comparing two groups that differ in height by one inch, the higher group has on average 4.45 pounds higher in weight with a 95% CI of (4.17, 4.72) using robust SE. We reject the null hypothesis of no association ($p = 4.8 \times 10^{-196}$) and conclude that we have a very strong evidence of a first-order linear association between height and weight. The intercept is -140.28 pounds at zero inches in height, which is an extrapolation from the range of the observed data. Since the weight cannot be negative, the intercept is not of scientific interest.

3. The heteroscedasticity-robust standard error in the estimate of the slope coefficient is 0.138. The homoscedasticity-non robust/ naïve standard error in the estimate of the slope coefficient is 0.126. The robust standard error is slightly larger than non-robust standard error by 0.012 (which is non-significant). I prefer to use robust standard error for the estimate of the slope coefficient since it is generally larger than the non-robust standard error, meaning robust standard error is more conservative. There are also relatively fewer risks of Type 1 error rates if uses robust standard error instead of naïve standard error.

4. It is reasonable to use the fitted linear regression to estimate the mean weight of men from this population who are 70 inches tall. This would be interpolation where we use the linear regression model to predict the value of weight that is within the range of our data. Interpolation is reasonable here because we would have a greater possibility of obtaining a valid estimate.

5. Generally, it would not be reasonable to use the fitted linear regression to estimate the mean weight of men from this population who are 42 inches tall. This would be extrapolation where we use the linear regression model to predict the value of weight that is outside the range of our observed data. We should be caution when using extrapolation and not make invalid estimation. However, under the linear model assumptions, it is possible to legitimize extrapolation and make assumptions that our observed trend continues for values of height outside the range of observed data used to form the linear model.

6. After converting height from inches to centimeters and fit a new simple linear regression model: We estimate that when comparing two groups that differ in height by one centimeter, the higher group has on average 1.75 pounds higher in weight with a 95% CI of (1.64, 1.86) using robust SE. We reject the null hypothesis of no association ($p = 4.8 \times 10^{-196}$) and conclude that we have a very strong evidence of a first-order linear association between height and weight. The intercept is -140.28 pounds at zero height in centimeters, which is an extrapolation from the range of the observed data. Since the weight cannot be negative, the intercept is not of scientific interest. The slope parameter in centimeter of the new linear model (1.75 pounds/cm) is smaller in magnitude comparing to the slope of previous linear model (4.45 pounds/inch). The 95% CI of the new linear model (1.64, 1.86) is smaller in magnitude comparing to the previous linear model (4.17, 4.72). The robust and naïve standard error of the new model (robust se = 0.050, naïve se = 0.055) is also smaller in magnitude comparing to the previous model (robust se = 0.138, naïve se = 0.126). The difference between robust se and naïve se is smaller in the new linear model, although the differences is non-significant. The intercepts (-140.28 pounds) and the p-values ($p = 4.8 \times 10^{-196}$) of the two models are the same.

7. I think the fitted value from the simple linear regress is a better prediction of the weight of another 73 inches tall man from this population. Since 73 inches in height is within the observed data range, we can make interpolation from the linear regression model, and inferred that the weight of another 73 inches tall man would be 184.57 pounds (as the fitter line go through an average weight of 184.57 pounds at 73 inches in height). It would not be valid to use a single data point within the dataset, i.e., the weight of another participant 2001that was 73 inches tall, to make inference of the population.

Code Appendix

```
### Setting up the packages, options we'll need:
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
### ------------------------------------------------------------
### Reading in the data.
library(tidyverse)
wcgs <- read_csv("wcgs.csv")
### ------------------------------------------------------------
### Q1
### Plot a scatterplot of weight on the vertical axis and height on the horizontal axis
plot(weight0 ~ height0, data=wcgs, main ="Weight vs. Height",
     xlab="height (inches)", ylab="Weight (lbs)")

### ------------------------------------------------------------
### Q2
### Here's a linear model for wcgs dataset using rigr:regress
wcgs %>%
  select(height0, weight0)

library(rigr)
mod1 <- regress("mean", weight0 ~ height0, data = wcgs)
mod1 %>%
  coef() %>%
  as.data.frame() %>%
  select(c("Estimate", "Naive SE", "Robust SE", "95%L", "95%H",
           "Pr(>|t|)"))
### ------------------------------------------------------------
### Q3
### Here's the Robust and naive SE
mod1 %>%
  coef() %>% round(3) %>%
  as.data.frame() %>%
  select(c("Naive SE", "Robust SE"))
### ------------------------------------------------------------
### Q6
### Convert height from inch to cm, then run a linear model
### using rigr:regress
wcgs_cm <- wcgs %>%
  select(height0, weight0) %>%
  mutate(height0_cm = height0*2.54)

mod2 <- regress("mean", weight0 ~ height0_cm, data = wcgs_cm)
mod2 %>%
  coef() %>%
  as.data.frame() %>%
  select(c("Estimate", "Naive SE", "Robust SE", "95%L", "95%H",
           "Pr(>|t|)"))

### ------------------------------------------------------------
### Q7
### Average weight at height = 73 inches
```

```r
wcgs %>%
  select(height0, weight0) %>%
  filter(height0 == 73) %>%
  summary()
### ------------------------------------------------------------
```