Nepal Study: Linear Regression Models

Anika Sompuram

2025-02-15

Introduction

This analysis fits separate simple linear regression models with weight as the response variable and height as the predictor for male and female children from the Nepal dataset.

```
# Load libraries
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.4     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# Load the data
library(readr)
nepal_data <- read.csv("nepal.csv")

# Subset the data by sex
male_data <- subset(nepal_data, sex == 1)
female_data <- subset(nepal_data, sex == 2)

# Fit the models for male and female data
male_model <- lm(weight ~ height, data = male_data)
female_model <- lm(weight ~ height, data = female_data)

# Summaries of the models
summary(male_model)
```

```
##
## Call:
## lm(formula = weight ~ height, data = male_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.7192 -0.5064 -0.0510  0.4496  3.2427
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.086925   0.288998  -31.44   <2e-16 ***
## height       0.239343   0.003341   71.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8373 on 453 degrees of freedom
## Multiple R-squared:  0.9189, Adjusted R-squared:  0.9187
## F-statistic:  5131 on 1 and 453 DF,  p-value: < 2.2e-16
```

```
summary(female_model)
```

```
##
## Call:
## lm(formula = weight ~ height, data = female_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.82127 -0.57982 -0.02652  0.50813  3.15115
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.371211   0.303580  -27.57   <2e-16 ***
## height       0.228194   0.003551   64.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8916 on 420 degrees of freedom
## Multiple R-squared:  0.9077, Adjusted R-squared:  0.9075
## F-statistic:  4129 on 1 and 420 DF,  p-value: < 2.2e-16
```

```
# Coefficients for both models
male_coeffs <- coef(male_model)
female_coeffs <- coef(female_model)

cat("Male Model Coefficients:\n")
```

```
## Male Model Coefficients:
```

```
print(male_coeffs)
```

```
## (Intercept)      height
##  -9.0869252   0.2393433
```

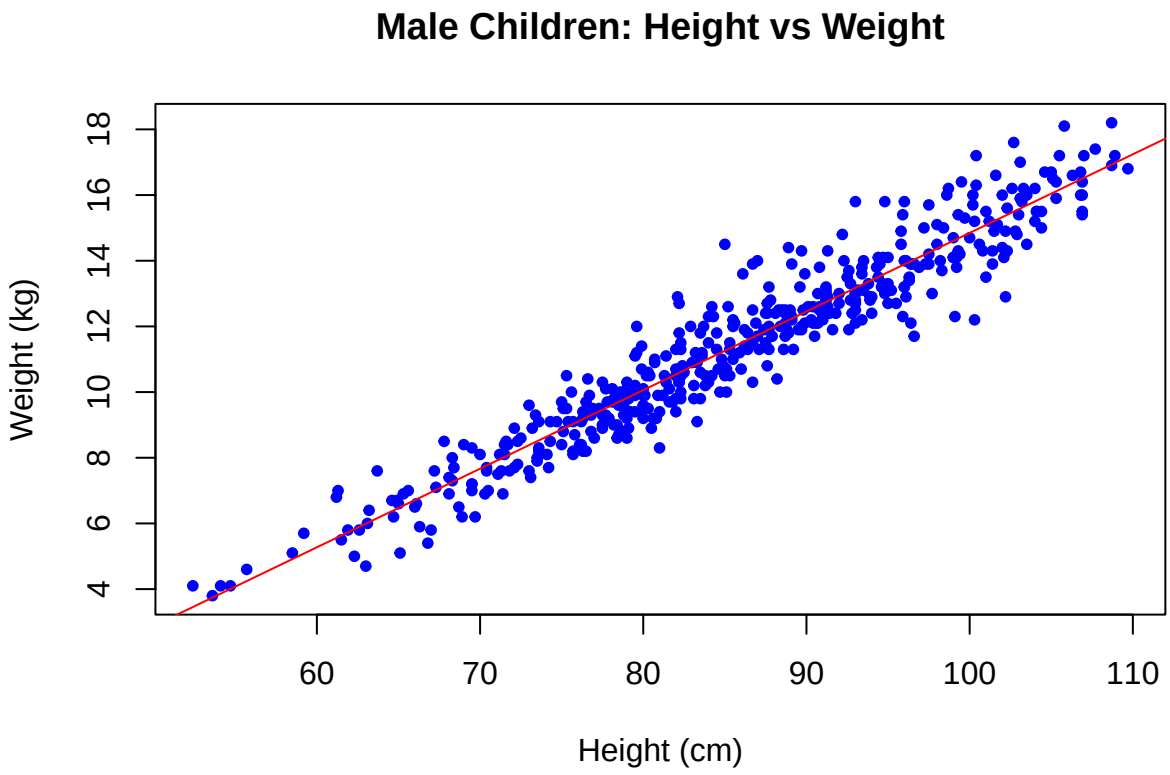
```
cat("\nFemale Model Coefficients:\n")
```

```
##
## Female Model Coefficients:
```

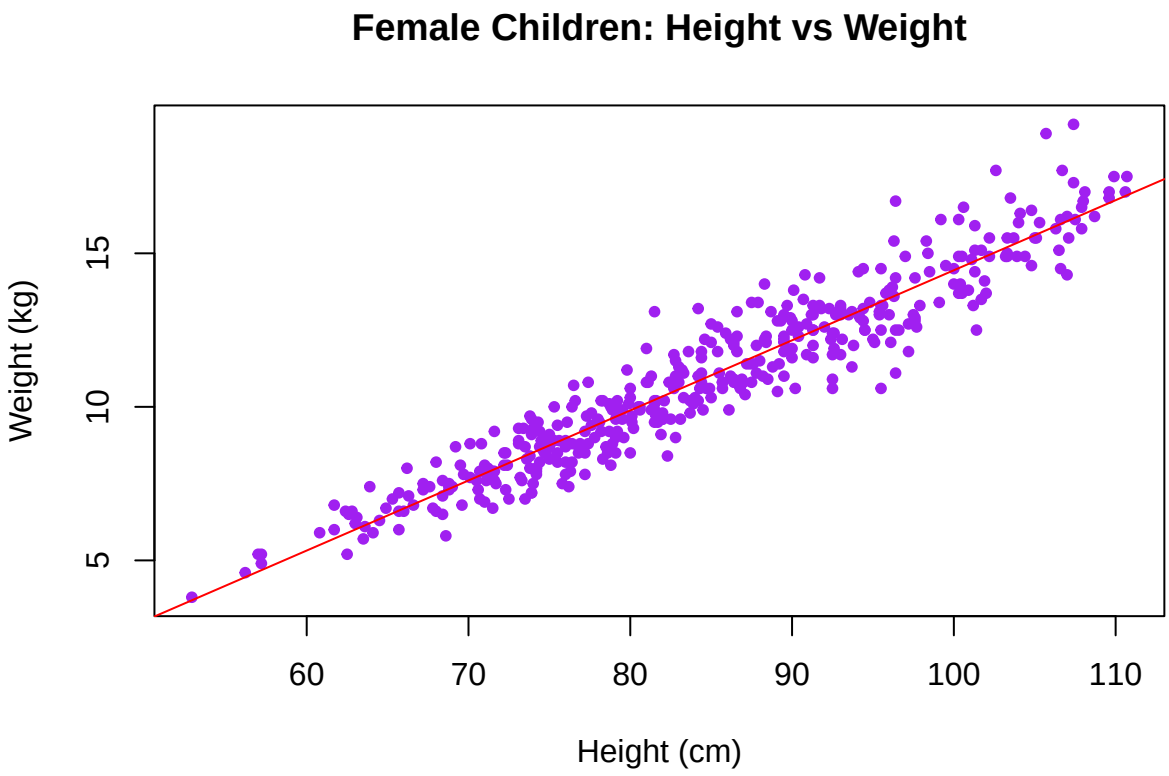
```r
print(female_coeffs)
```

```
## (Intercept)      height
##  -8.3712108   0.2281936
```

```r
# Scatter plot and regression line for males
plot(male_data$height, male_data$weight, xlab = "Height (cm)", ylab = "Weight (kg)", main = "Male Childr
abline(a = male_model$coefficients[1], b = male_model$coefficients[2], col = "red")
```




```r
# Scatter plot and regression line for females
plot(female_data$height, female_data$weight, xlab = "Height (cm)", ylab = "Weight (kg)", main = "Female
abline(a = female_model$coefficients[1], b = female_model$coefficients[2], col = "red")
```




```
# 4b: The male model appears to fit the data better.
# This is because the male model has a higher coefficient than the female model.
# This indicates height has a stronger influence on weight in male children than
# female. Therefore, the male model has a stronger relationship between height
# and weight, making it a better fit.
# The male model also has a higher R-squared value and a lower residual standard
# error (RSE) than the female model, also indicating that it fits the
# data better.
```