worksheet04_group01

Antonio Alfaro de Prado, Hyunchang Oh, Tanya Toluay

2024-05-09

Exercise 1 - Simple linear regression

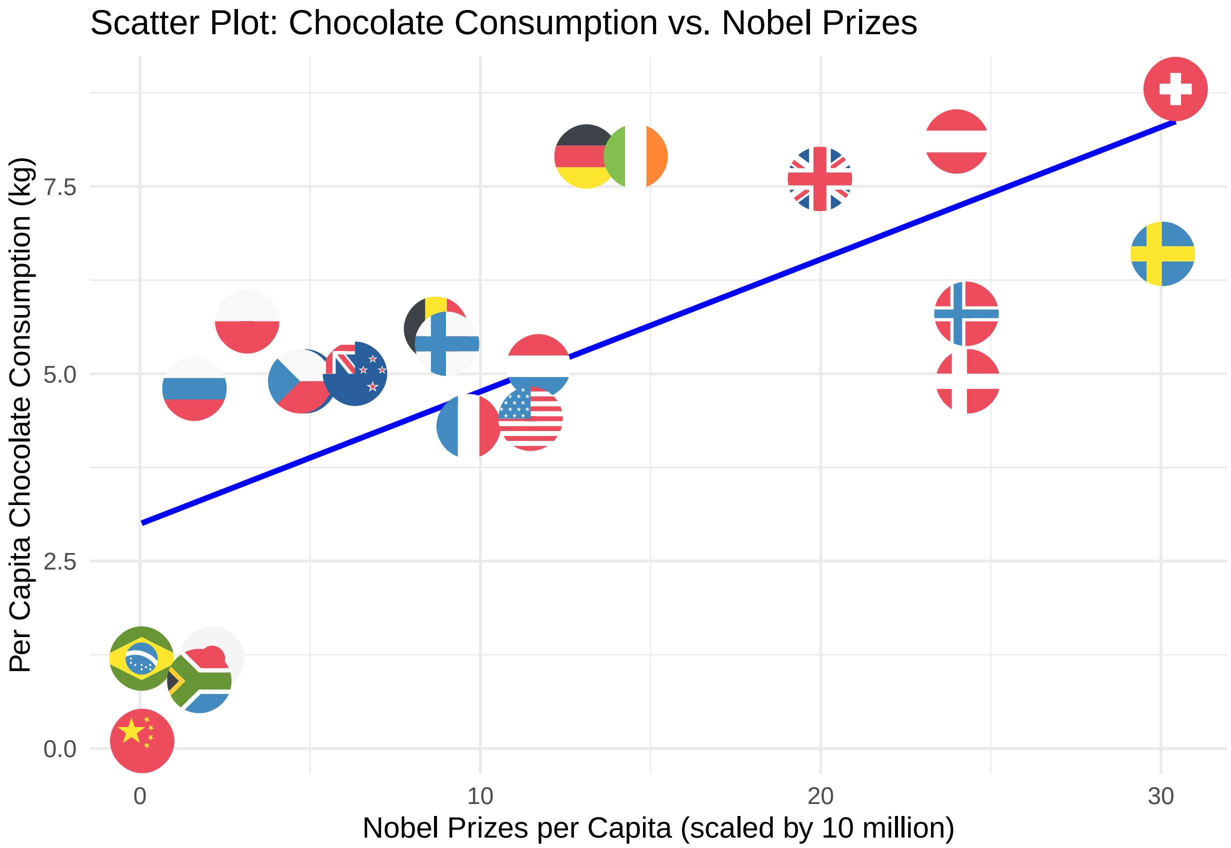
```
chocolate <- read_csv("C:/Users/tanya/Downloads/chocolate.csv")
head(chocolate)
```

```
## # A tibble: 6 x 3
##   Country        Nobel prizes per capita (scaled by 10 ~1 Per capita chocolate~2
##   <chr>                                             <dbl>                  <dbl>
## 1 Switzerland                                        30.4                    8.8
## 2 Austria                                            24.0                    8.1
## 3 Ireland                                            14.6                    7.9
## 4 Germany                                            13.1                    7.9
## 5 United Kingdom                                     20.0                    7.6
## 6 Sweden                                             30.1                    6.6
## # i abbreviated names: 1: ‘Nobel prizes per capita (scaled by 10 million)‘,
## #   2: ‘Per capita chocolate consumption (kg)‘
```

```
chocolate$code <- tolower(countrycode(chocolate$Country, "country.name", "iso2c"))

lm_model <- lm(`Per capita chocolate consumption (kg)` ~
                 `Nobel prizes per capita (scaled by 10 million)`, data = chocolate)
summary(lm_model)
```

```
##
## Call:
## lm(formula = ‘Per capita chocolate consumption (kg)‘ ~ ‘Nobel prizes per capita (scaled by 10 millio
##     data = chocolate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9083 -1.6464  0.6216  1.0714  2.5867
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                       2.99699    0.57883   5.178
## ‘Nobel prizes per capita (scaled by 10 million)‘  0.17649    0.03842   4.594
##                                                  Pr(>|t|)
## (Intercept)                                      4.58e-05 ***
## ‘Nobel prizes per capita (scaled by 10 million)‘ 0.000176 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.726 on 20 degrees of freedom
## Multiple R-squared:  0.5135, Adjusted R-squared:  0.4891
## F-statistic: 21.11 on 1 and 20 DF,  p-value: 0.0001757
```

```
ggplot(chocolate, aes(x = `Nobel prizes per capita (scaled by 10 million)`,
                      y = `Per capita chocolate consumption (kg)`, label = Country)) +
  geom_point() +  # Scatter plot of data points
  geom_smooth(method = "lm", se = FALSE, color = "blue") +  # Add regression line
  geom_flag(aes(country=chocolate$code),size = 10) +  # Add country names to points
  labs(x = "Nobel Prizes per Capita (scaled by 10 million)",
       y = "Per Capita Chocolate Consumption (kg)",
       title = "Scatter Plot: Chocolate Consumption vs. Nobel Prizes") +
    theme_minimal()
```



```
#What are the dependent and independent variables for your model?
#Visualize both using a scatterplot
intercept <- coef(lm_model)[1]
slope <- coef(lm_model)[2]
r_squared <- summary(lm_model)$r.squared

cat("Intercept:", intercept, "\n")
```

```
## Intercept: 2.996993
```

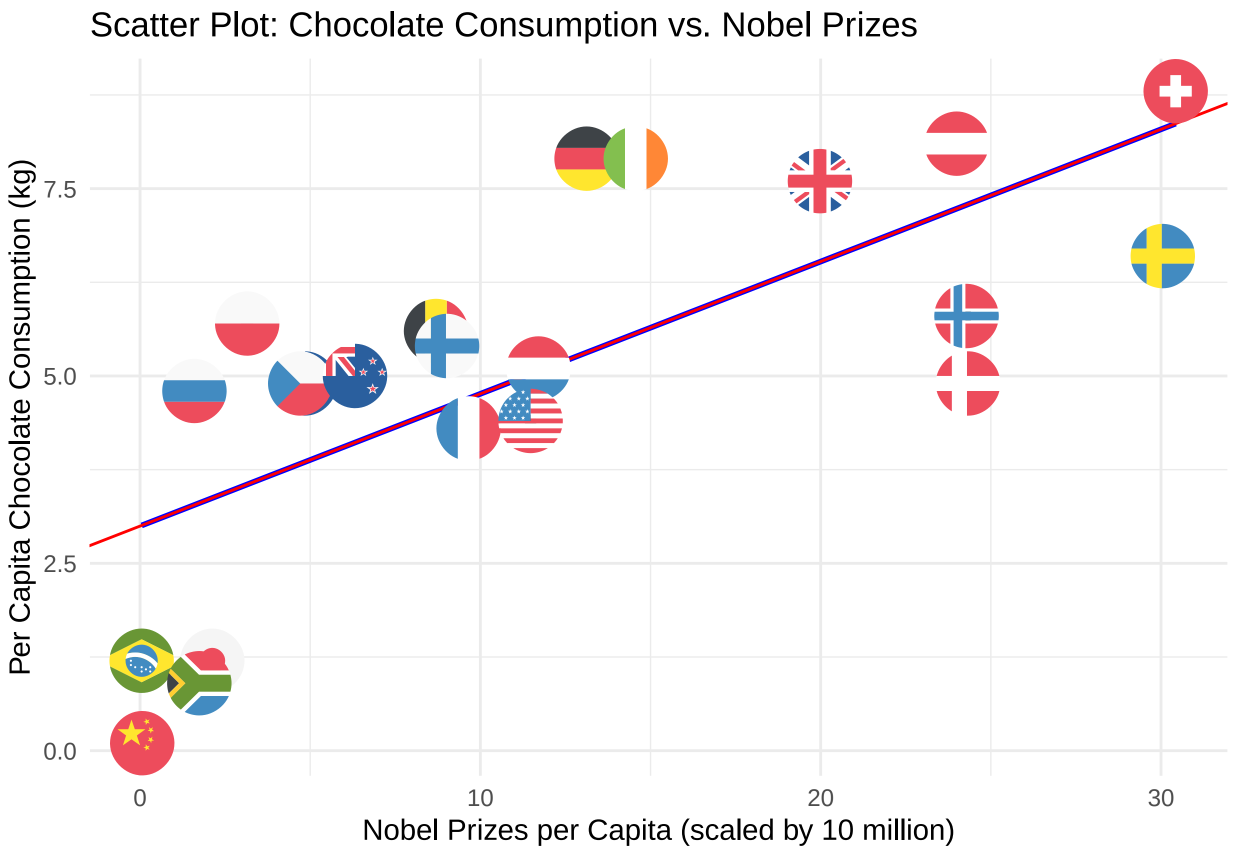
```r
cat("Slope:", slope, "\n")
```

```
## Slope: 0.1764903
```

```r
cat("Coefficient of determination (R^2):", r_squared, "\n")
```

```
## Coefficient of determination (R^2): 0.5134667
```

```r
#add geom_smooth
ggplot(chocolate, aes(x = `Nobel prizes per capita (scaled by 10 million)`,
                      y = `Per capita chocolate consumption (kg)`, label = Country)) +
  geom_point() +  # Scatter plot of data points
  # Add regression line using geom_smooth
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  # Add regression line computed manually
  geom_abline(intercept = intercept, slope = slope, color = "red") +
  # Add country names to points
  geom_flag(aes(country=chocolate$code),size = 10) +
  labs(x = "Nobel Prizes per Capita (scaled by 10 million)",
       y = "Per Capita Chocolate Consumption (kg)",
       title = "Scatter Plot: Chocolate Consumption vs. Nobel Prizes") +
  theme_minimal()
```



Discussion

The p-values associated with the coefficients suggest that these relationships are statistically significant at conventional significance levels ($p < 0.05$), indicating that the observed relationships are unlikely to have occurred by chance.

Additionally, the R-squared value of 0.5135 indicates that approximately 51.35% of the variability in per capita chocolate consumption can be explained by the linear relationship with Nobel prizes per capita.

Overall, these findings suggest a positive association between Nobel prizes per capita and per capita chocolate consumption, indicating that regions or countries with higher numbers of Nobel prizes per capita tend to have higher levels of chocolate consumption.

However, it's important to note that correlation does not imply causation! We think there are more complex reasons behind this p-value, and we should investigate this further. One example could be that it's possible that higher socioeconomic standards lead to higher chocolate consumption as wealthier individuals or countries may have greater disposable income to spend on luxury items like chocolate. Elsewise, we could argue that countries such as Brazil or Columbia where we import the chocolate from would earn the most nobel prizes.

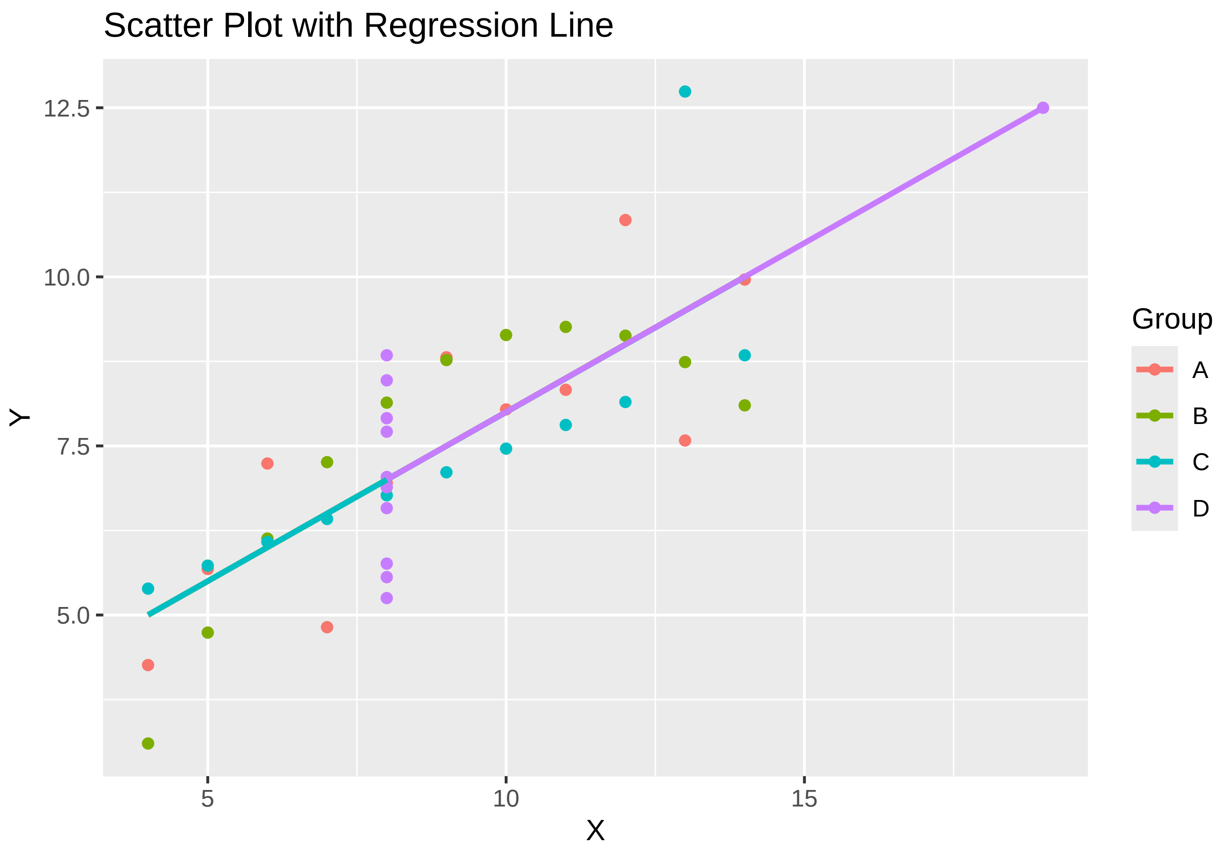
Let's note that it seems like it is almost as if the people who make chocolate are the same people who give out Nobel prizes.

Exercise 2 - Four datasets

```
data <- read_csv("C:/Users/tanya/Downloads/four_datasets.csv")

#summary(data)

summary_stats <- data %>%
  group_by(Dataset) %>%
  summarise(
    mean_x = mean(x),
    std_x = sd(x),
    mean_y = mean(y),
    std_y = sd(y),
    correlation = cor(x, y)
  )

# Fit linear regression model and visualize the data for each group
lm_model <- lm(y ~ x, data = data)
summary(lm_model)
```

```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9204 -0.7459 -0.0202  0.7592  3.2396
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.00130    0.52059   5.765 8.64e-07 ***
## x            0.49993    0.05457   9.161 1.44e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.145 on 42 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6585
## F-statistic: 83.92 on 1 and 42 DF,  p-value: 1.437e-11
```

```
ggplot(data, aes(x = x, y = y, color = Dataset)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Scatter Plot with Regression Line",
       x = "X",
       y = "Y",
       color = "Group")
```



Discussion

The dataset includes four groups labeled A, B, C, and D, with observations for two variables, x and y.

For all groups combined, the average value of x is around 9, with some variation within groups (standard deviation of about 3.32). Similarly, the average value of y is approximately 7.50, with some variability (standard deviation around 2.03).

There's a strong positive relationship between x and y in each group. This means that as x values increase, y values tend to increase as well, and vice versa. The correlation between x and y is consistently high across all groups, with values ranging from 0.816 to 0.817.

Using linear regression models, we found that there's a significant linear relationship between x and y in each group. This is supported by the high correlation values and the fitted regression lines in the scatter plots.

In conclusion, we see consistent pattern across all groups, indicating a strong association between variables x and y, characterized by a positive linear trend.

Exercise 3 - Regression to the mean

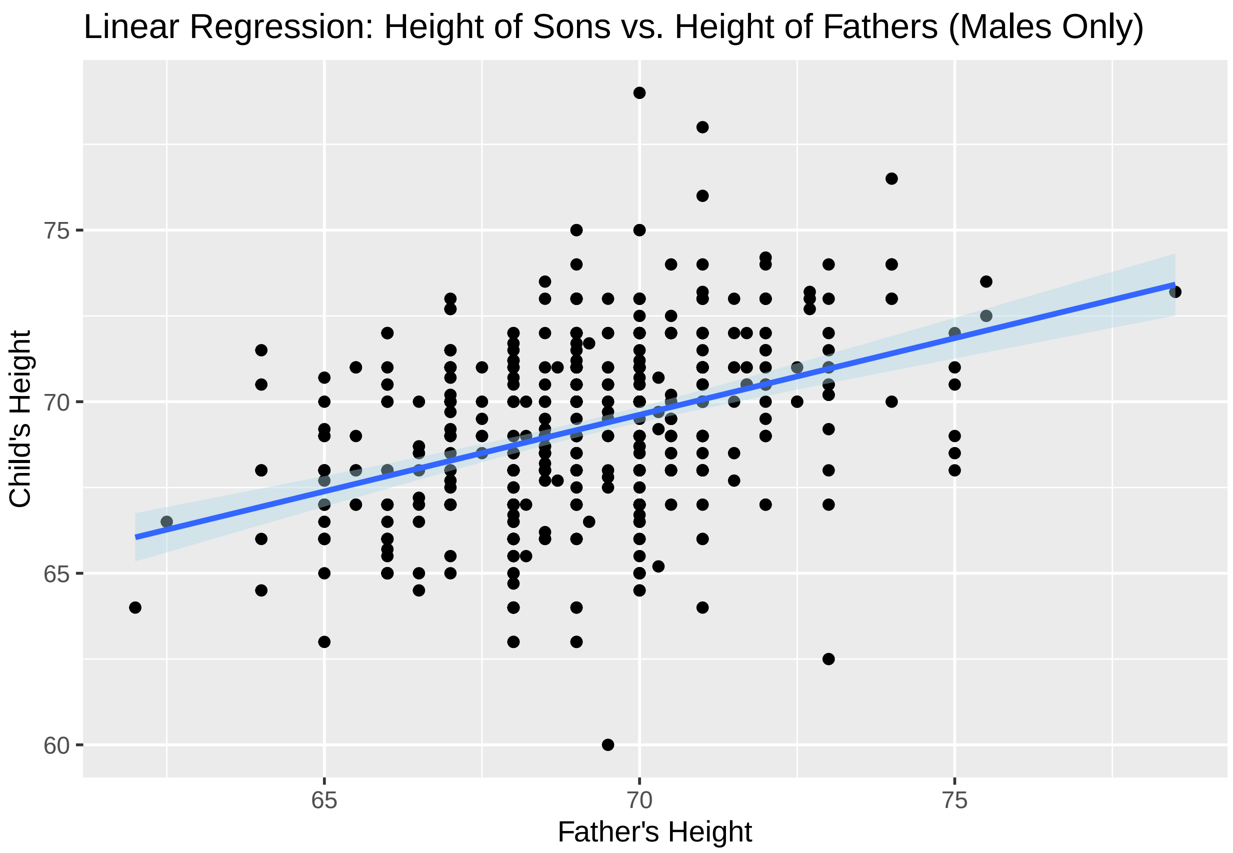
```
galton <- GaltonFamilies
head(galton)
```

```
##   family father mother midparentHeight children childNum gender childHeight
## 1    001   78.5   67.0           75.43        4        1   male        73.2
## 2    001   78.5   67.0           75.43        4        2 female        69.2
## 3    001   78.5   67.0           75.43        4        3 female        69.0
## 4    001   78.5   67.0           75.43        4        4 female        69.0
## 5    002   75.5   66.5           73.66        4        1   male        73.5
## 6    002   75.5   66.5           73.66        4        2   male        72.5
```

```
# Filter data for sons
galton_males <- subset(galton, gender == "male")

lm_model <- lm(childHeight ~ father, data = galton_males)
summary(lm_model)
```

```
##
## Call:
## lm(formula = childHeight ~ father, data = galton_males)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3959 -1.5122  0.0413  1.6217  9.3808
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.36258    3.30837  11.596   <2e-16 ***
## father       0.44652    0.04783   9.337   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.416 on 479 degrees of freedom
## Multiple R-squared:  0.154,  Adjusted R-squared:  0.1522
## F-statistic: 87.17 on 1 and 479 DF,  p-value: < 2.2e-16
```

```
# I dont think this looks right, check it again
ggplot(galton_males, aes(x = father, y = childHeight)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, fill = "lightblue") +
#   geom_line() +
  labs(title = "Linear Regression: Height of Sons vs. Height of Fathers (Males Only)",
       x = "Father's Height",
       y = "Child's Height")
```



```
# Regress the height of fathers on the height of sons
lm_model_father_on_son <- lm(father ~ childHeight, data = galton_males)
slope_father_on_son <- coef(lm_model_father_on_son)[2]
lm_model_son_on_father <- lm(childHeight ~ father, data = galton_males)
slope_son_on_father <- coef(lm_model_son_on_father)[2]

slope_father_on_son
```

```
## childHeight
##   0.3448085
```

```
slope_son_on_father
```

```
##    father
## 0.4465226
```

```
# Relationship between the slopes
inverse_slope_son_on_father <- 1 / slope_father_on_son
inverse_slope_father_on_son <- 1 / slope_son_on_father
```

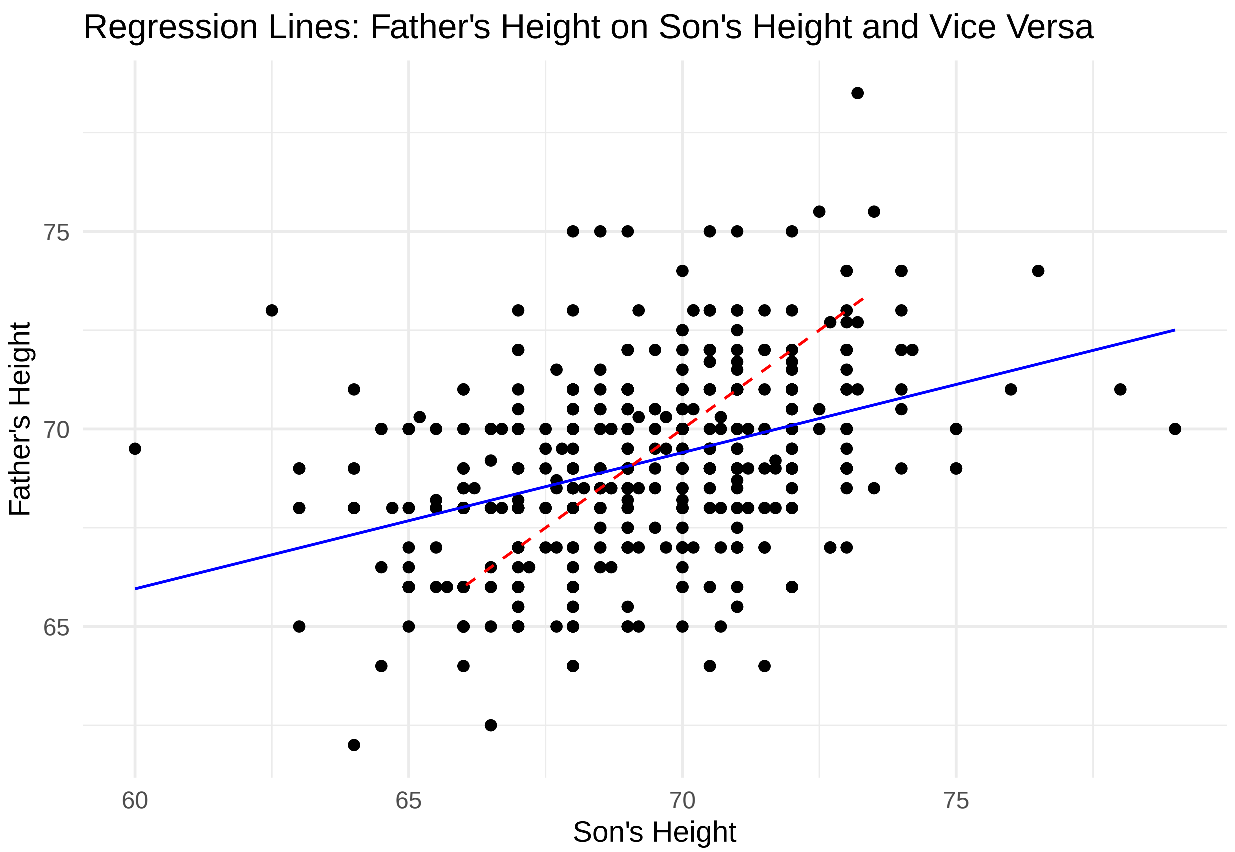
```r
# Output the relationship
inverse_slope_son_on_father
```

```
## childHeight
##     2.90016
```

```r
inverse_slope_father_on_son
```

```
##   father
## 2.239528
```

```r
lm_model_father_on_son <- lm(father ~ childHeight, data = galton_males)
lm_model_son_on_father <- lm(childHeight ~ father, data = galton_males)

# Create prediction data frames for both models
prediction_data_father_on_son <- data.frame(childHeight =
                                              seq(min(galton_males$childHeight),
                                                  max(galton_males$childHeight),
                                                  length.out = 100))
prediction_data_son_on_father <- data.frame(father =
                                              seq(min(galton_males$father),
                                                  max(galton_males$father),
                                                  length.out = 100))

# Predict father's height and son's height for both models
prediction_data_father_on_son$father <- predict(lm_model_father_on_son,
                                                newdata = prediction_data_father_on_son)
prediction_data_son_on_father$childHeight <- predict(lm_model_son_on_father,
                                                     newdata =
                                                       prediction_data_son_on_father)

ggplot(galton_males, aes(x = childHeight, y = father)) +
  geom_point() +
  geom_line(data = prediction_data_father_on_son,
            aes(x = childHeight, y = father),
            color = "blue") +
  geom_line(data = prediction_data_son_on_father,
            aes(x = childHeight, y = childHeight),
            color = "red", linetype = "dashed") +
  labs(title = "Regression Lines: Father's Height on Son's Height and Vice Versa",
       x = "Son's Height",
       y = "Father's Height") +
  theme_minimal()
```



Discussion

The concept of regression to the mean, or reversion to mediocrity, is a phenomenon observed when extreme values in a sample tend to move closer to the average or mean value upon subsequent measurements. In the context of the Galton dataset and the relationship between the height of fathers and sons, as follows:

If a father has an extreme height (either exceptionally tall or short), his son's predicted height, according to the regression model, will tend to be closer to the average height (mean) of the population. This is because extreme heights are less likely to be maintained across generations due to genetic recombination and environmental factors. Conversely, if a father has a height close to the average, his son's predicted height will also tend to be close to the average, but with less extreme deviation.

So the regression to the mean phenomenon suggests that extreme values observed in one generation are likely to be less extreme in the next generation, tending towards the population average. This is reflected in the relationship between the heights of fathers and sons, where extreme heights in fathers tend to be moderated towards the average height in their sons.