Worksheet4

Jingwen Luo (5597340) Daeho Lee (5597689) Gyeongsil kim (5597789)

2025-05-19

Exercise 1 - Four datasets

For every group (A-D) in four_datasets.csv

- compute means, standard deviations and correlations,

```
data <- read.csv("../Data/four_datasets.csv")

stats <- data %>%
  group_by(Dataset) %>%
  summarise(
    mean_x = mean(x), mean_y = mean(y),
    sd_x = sd(x),sd_y = sd(y),
    correlation = cor(x, y),
    intercept = lm(y ~ x)$coefficients[1],
    slope = lm(y ~ x)$coefficients[2]
  )

stats
```

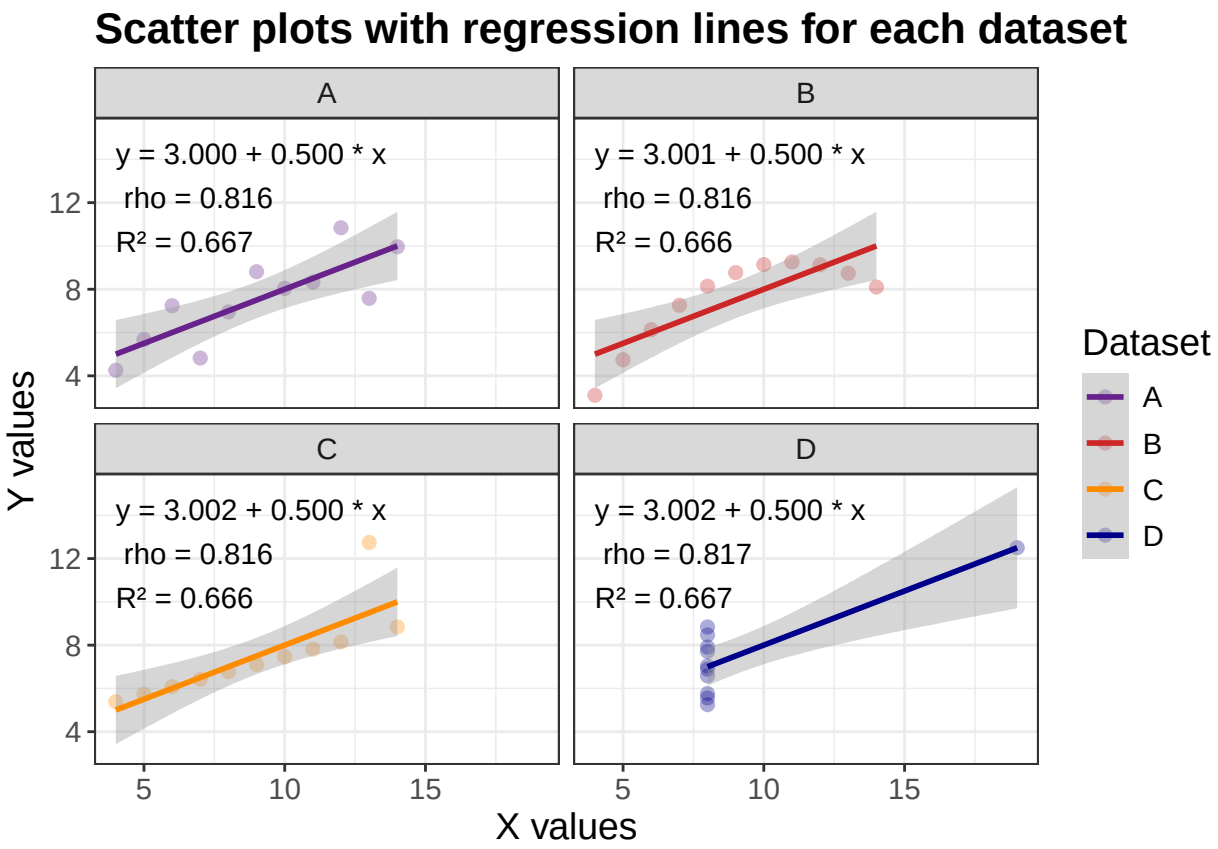
```
## # A tibble: 4 x 8
##   Dataset mean_x mean_y  sd_x  sd_y correlation intercept slope
##   <chr>    <dbl>  <dbl> <dbl> <dbl>       <dbl>     <dbl> <dbl>
## 1 A            9   7.50  3.32  2.03       0.816      3.00 0.500
## 2 B            9   7.50  3.32  2.03       0.816      3.00 0.5
## 3 C            9   7.5   3.32  2.03       0.816      3.00 0.500
## 4 D            9   7.50  3.32  2.03       0.817      3.00 0.500
```

- fit a linear regression model and visualize the data using scatter plots and regression lines,

```r
regression_stats <- data %>%
  group_by(Dataset) %>%
  summarise(
    mean_x = mean(x),
    mean_y = mean(y),
    rho = cor(x, y),                      # Pearson correlation coefficient
    r_squared = summary(lm(y ~ x))$r.squared,  # Coefficient of determination
    intercept = lm(y ~ x)$coefficients[1],     # alpha
    slope = lm(y ~ x)$coefficients[2],         # beta
    # construct lr function (y = a + b*x)
    equation = sprintf("y = %.3f + %.3f * x", intercept, slope),
    # compute rho and R²
    stats_label = sprintf("rho = %.3f\nR² = %.3f", rho,r_squared))

regression_stats
```

```
## # A tibble: 4 x 9
##   Dataset mean_x mean_y   rho r_squared intercept slope equation     stats_label
##   <chr>    <dbl>  <dbl> <dbl>     <dbl>     <dbl> <dbl> <chr>        <chr>
## 1 A            9   7.50 0.816     0.667      3.00 0.500 y = 3.000 +~ "rho = 0.8~
## 2 B            9   7.50 0.816     0.666      3.00 0.5   y = 3.001 +~ "rho = 0.8~
## 3 C            9   7.5  0.816     0.666      3.00 0.500 y = 3.002 +~ "rho = 0.8~
## 4 D            9   7.50 0.817     0.667      3.00 0.500 y = 3.002 +~ "rho = 0.8~
```

```r
ggplot(data, aes(x = x, y = y, color = Dataset)) +
  geom_point(size=2,alpha=0.33) +
  geom_smooth(formula = 'y ~ x',method = "lm") +
  facet_wrap(~ Dataset) +
  scale_color_manual(
    values = c("darkorchid4","firebrick3","darkorange","darkblue"))+
  geom_text(data = regression_stats,
            aes(x = min(data$x), y = max(data$y)+2,
                label = paste(equation, "\n", stats_label)),
            hjust = 0,vjust = 1,color = "black",size = 4) +
  labs(title = "Scatter plots with regression lines for each dataset",
       x = "X values",y = "Y values") +
  theme_bw() +
  theme(text = element_text(size = 14),
        plot.title = element_text(size = 16, face = "bold"))
```



- discuss what you find.

These four datasets shares the same mean values for both x and y, nearly identical correlation coefficients ranging from 0.816 to 0.817.When performing linear regression on these four datasets, we obtain nearly identical regression equations and correlation coefficients.

However, from the scatter plots, we can observe significant differences: Dataset A closely follows a linear relationship; Dataset B appears to follow a nonlinear pattern; Dataset C shows increasing deviation from the regression line toward the end, with a clear outlier at the far end; and in Dataset D, all points except one stand near the vertical line x = 7.5. This example demonstrates that even when linear regression results are the same, the degree to which the data fits a linear model varies. Visual inspection through scatter plots and subjective insight is always helpful for accurate data interpretation. Also after fitting a linear model, we should check our residuals.

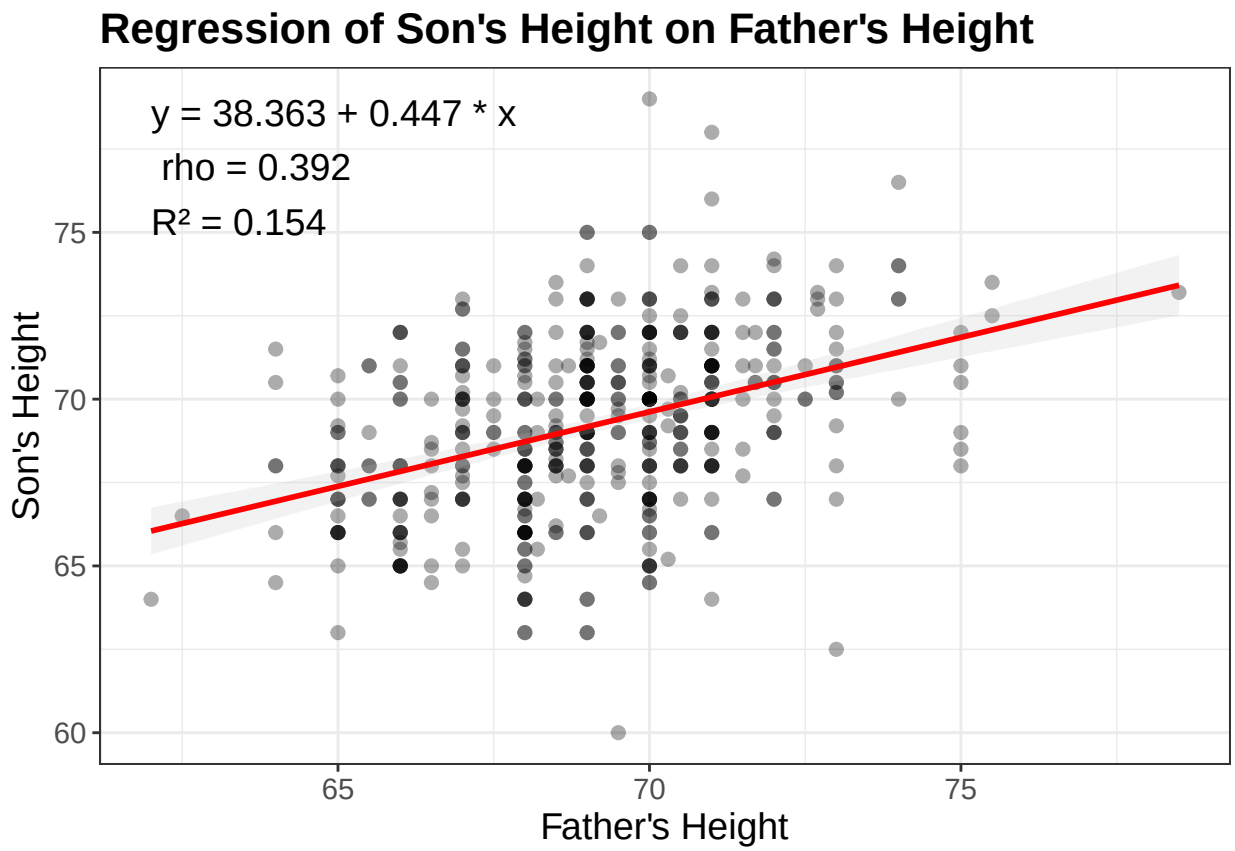
Exercise 2 - Regression to the mean

- Use the GaltonFamilies dataset from the HistData package fit a simple linear regression model of the height of sons on the height of fathers, plotting the data points, the regression line and the uncertainty in the regression line.

```r
library(HistData)
data(GaltonFamilies)
head(GaltonFamilies)
```

```
##   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
```

```r
data_filtered <- GaltonFamilies %>% filter(gender=="male")
lm_sons <- data_filtered%>%
  group_by(gender) %>%
  summarise(intercept = lm(childHeight ~ father)$coefficients[1],
            slope = lm(childHeight ~ father)$coefficients[2],
            rho = cor(father, childHeight),
            r_squared = summary(lm(childHeight ~ father))$r.squared,
            equation = sprintf("y = %.3f + %.3f * x", intercept, slope),
            stats_label = sprintf("rho = %.3f\nR² = %.3f", rho,r_squared))

plot1 <- data_filtered %>%
  ggplot(aes(x = father, y = childHeight)) +
  geom_point(size = 2, alpha = 0.33) +
  geom_smooth(formula='y ~ x',method="lm",color="red",fill="grey",alpha = 0.2)+
  geom_text(data = lm_sons,
            aes(x = min(data_filtered$father),
                y = max(data_filtered$childHeight),
                label = paste(equation, "\n", stats_label)),
            hjust = 0,vjust = 1,color = "black",size = 5) +
  labs(x = "Father's Height",
       y = "Son's Height",
       title = "Regression of Son's Height on Father's Height") +
  theme_bw() +
  theme(text = element_text(size = 14),
        plot.title = element_text(size = 16, face = "bold"))

plot1
```



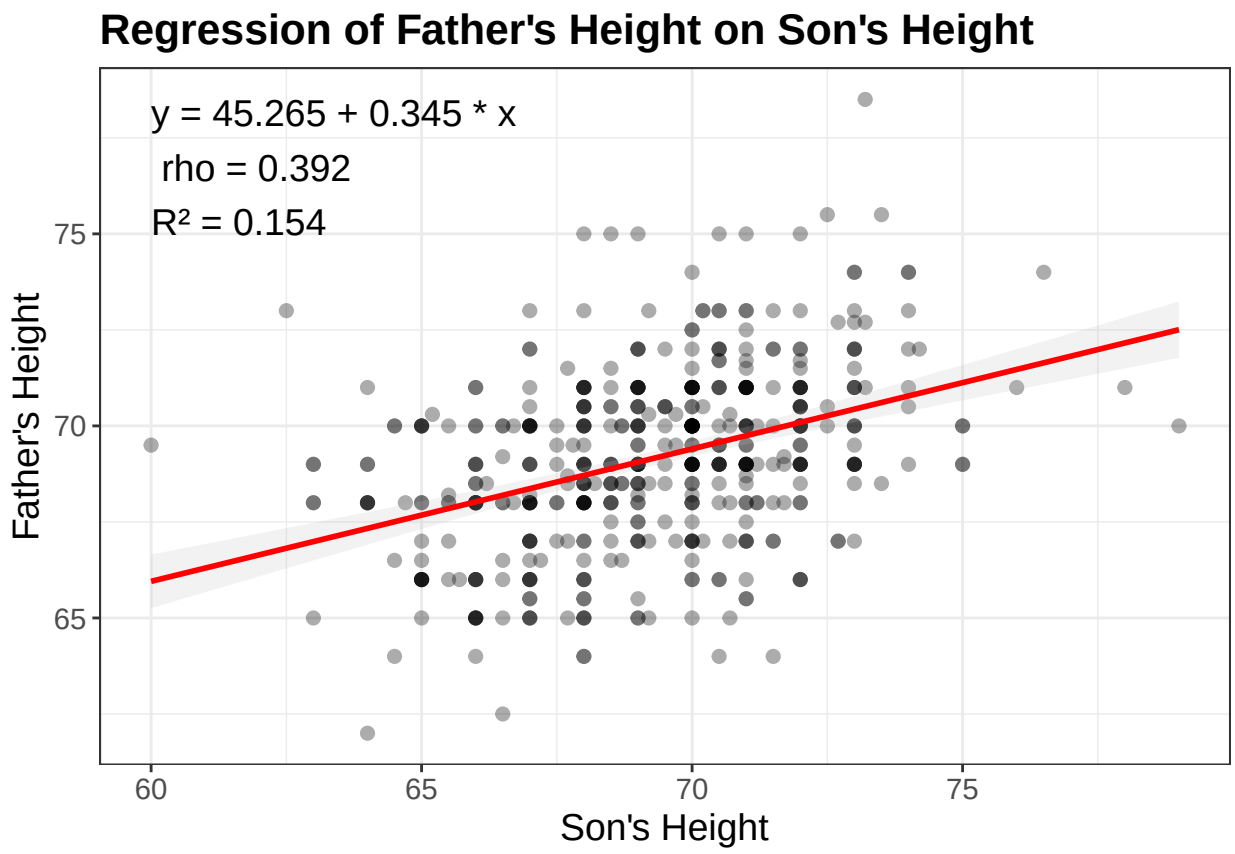
- How does the predicted height of the son depend on the height of the father? Discuss how this related to the concept of regression to the mean (or as Galton more unpleasantly put it: reversion to mediocrity).

The regression function shows that son's height $= 38.363 + 0.447 \times$ father's height. This reflects Galton's 19th-century observation of regression to the mean: extreme traits in parents tend to be less extreme in their children. In regression analysis, a slope $\beta < 1$ means that increases in X lead to smaller, attenuated changes in Y. This regression effect isn't due to error, but to natural variation and the model's tendency to pull extreme values toward the mean.

- Now regress the height of fathers on the height of sons.

```
data_filtered <- GaltonFamilies %>% filter(gender == "male")
lm_fathers <- data_filtered%>%
  group_by(gender) %>%
  summarise(intercept = lm(father ~ childHeight)$coefficients[1],
            slope = lm(father ~ childHeight)$coefficients[2],
            rho = cor(father, childHeight),
            r_squared = summary(lm(father ~ childHeight))$r.squared,
            equation = sprintf("y = %.3f + %.3f * x", intercept, slope),
            stats_label = sprintf("rho = %.3f\nR² = %.3f", rho,r_squared))
```

```r
plot2 <- data_filtered %>%
  ggplot(aes(x = childHeight, y = father)) +  # Swapped x and y
  geom_point(size = 2, alpha = 0.33) +
  geom_smooth(formula='y ~ x',method="lm",color="red",fill="grey",alpha = 0.2)+
  geom_text(data = lm_fathers,
            aes(x = min(data_filtered$childHeight),
                y = max(data_filtered$father),
                label = paste(equation, "\n", stats_label)),
            hjust = 0,vjust = 1,color = "black",size = 5) +
  labs(x = "Son's Height",y = "Father's Height",
    title = "Regression of Father's Height on Son's Height") +
  theme_bw() +
  theme(text = element_text(size = 14),
        plot.title = element_text(size = 16, face = "bold"))

plot2
```



- What is the relationship between the slope of the regression of the height of fathers on the height of sons and the slope of the regression of the height of sons on the height of fathers?

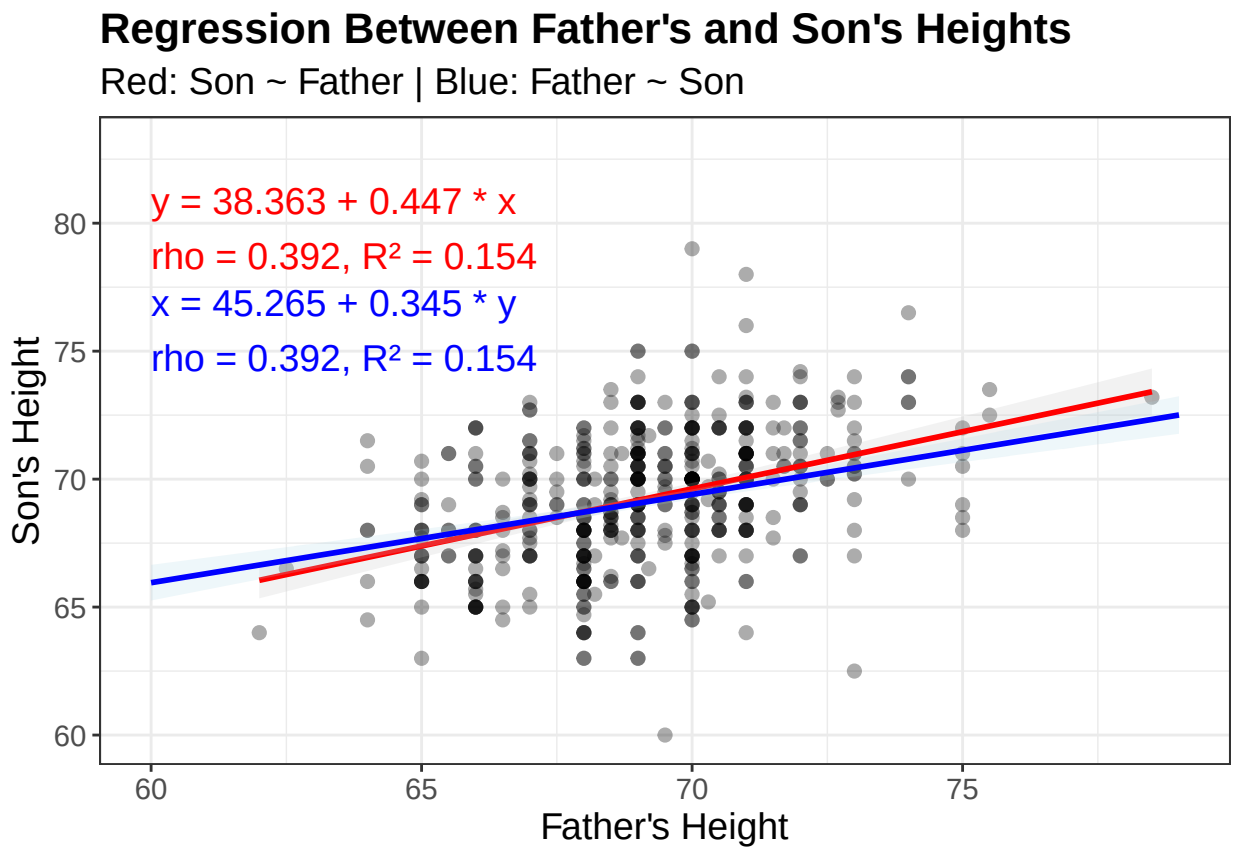
The slope of the regression of father's height on son's height is not the same as the slope of the regression of son's height on father's height. In this non-standardized case. $\beta_{Y|X} = r \cdot \frac{\sigma_X}{\sigma_Y}$, while $\beta_{X|Y} = r \cdot \frac{\sigma_Y}{\sigma_X}$.This shows that each regression captures a different relationship, depending on which variable is treated as the predictor.

- Overlay both regression lines on the same scatter plot. What does this tell us about regression to the mean?

```
# regression 1: son ~ father
lm_sons <- lm(childHeight ~ father, data_filtered)
intercept1 <- coef(lm_sons)[1]
slope1 <- coef(lm_sons)[2]
r1 <- cor(data_filtered$father, data_filtered$childHeight)
r2_1 <- summary(lm_sons)$r.squared
label1 <- sprintf("y = %.3f + %.3f * x\nrho = %.3f, R² = %.3f",
                  intercept1, slope1, r1, r2_1)

# regression 2: father ~ son
lm_fathers <- lm(father ~ childHeight, data_filtered)
intercept2 <- coef(lm_fathers)[1]
slope2 <- coef(lm_fathers)[2]
r2 <- cor(data_filtered$childHeight, data_filtered$father)
r2_2 <- summary(lm_fathers)$r.squared
label2 <- sprintf("x = %.3f + %.3f * y\nrho = %.3f, R² = %.3f",
                  intercept2, slope2, r2, r2_2)

plot3 <- data_filtered %>%
  ggplot(aes(x = father, y = childHeight)) +
  geom_point(size = 2, alpha = 0.33) +
  # regression 1: son ~ father
  geom_smooth(formula = 'y ~ x', method = "lm",
              color = "red", fill = "grey", alpha = 0.2) +
  # regression 2: father ~ son
  geom_smooth(aes(x = childHeight, y = father), method = "lm",
              color = "blue", fill = "lightblue", alpha = 0.2) +
  annotate("text", # regression 1: son ~ father
           x = 60, y = 83, label = label1,
           hjust = 0, vjust = 1.5, size = 5, color = "red") +
  annotate("text",  # regression 2: father ~ son
           x = 60, y = 79, label = label2,
```

```
           hjust = 0, vjust = 1.5, size = 5, color = "blue") +
  labs(x = "Father's Height",
       y = "Son's Height",
       title = "Regression Between Father's and Son's Heights",
       subtitle = "Red: Son ~ Father | Blue: Father ~ Son") +
  theme_bw() +
  theme(text = element_text(size = 14),
        plot.title = element_text(size = 16, face = "bold"))

plot3
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



When predicting father's height based on a son's height, the linear functions are different And the interpretation changes: now we're predicting father's height from son's height. Regression to the mean is reflected in a regression coefficient with a slope less than 1. Extreme predicted values are 'shrunk' toward the population average. This is a natural adjustment in statistical models due to imperfect correlation between variables, not a result of biological mechanisms.