Introduction to Statistical Modeling

Case Western Reserve University, Spring 2026

Unit 03-01: Simple Linear Regression: Discovery & Model

Instructor: Md Mutasim Billah

Example 1

Suppose we have a dataset of $n = 30$ CWRU women's tennis players, specifically their heights and weights. Let us use this information to draw a scatterplot!

Example 2

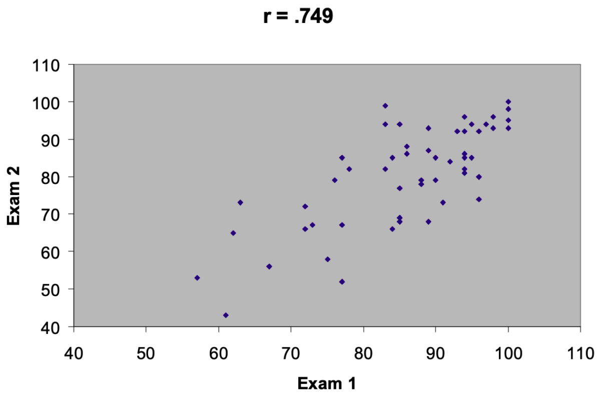
Suppose 25 students from the Department of Mathematics, Applied Mathematics, and Statistics at CWRU are randomly selected, and we measure two variables: X=foot length (inches) and Y=shoe length (inches). What kind of association would you be expecting between these two variables? Positive or Negative?

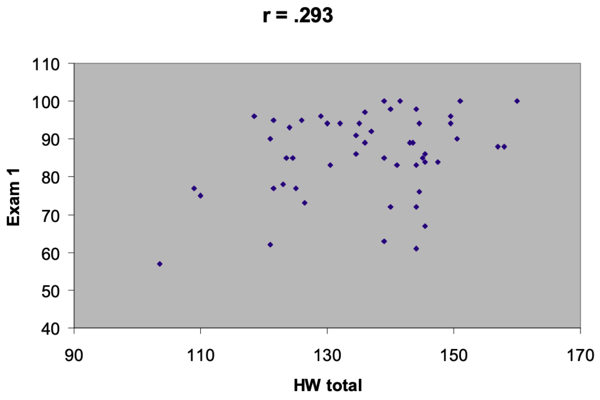
Example 3

Now, suppose 25 male students from the department of mathematics, applied mathematics and statistics are randomly selected. Let's say X=last three digits of student ID# and Y=height (inches). What kind of association are you guessing in this situation?

Example 4

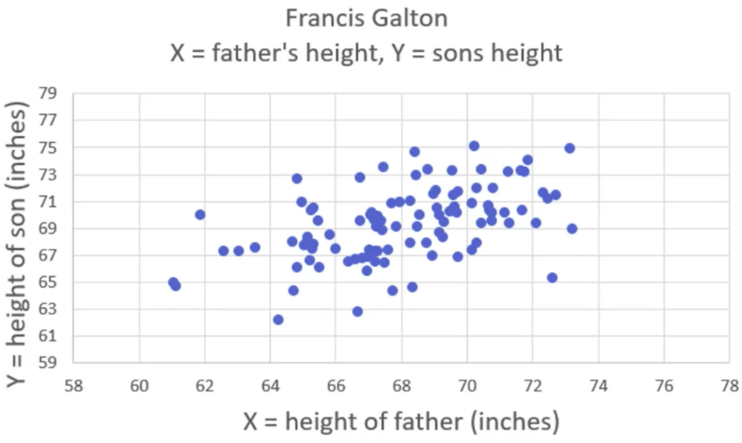
Regression by eye!





Example 5 (Galton's Heights Example)

A famous example: Francis Galton noticed a positive linear association between X=father's height (inches) and Y=son's height (inches). Here is the scatterplot and $r = 0.50$.



(1) What does it mean by $r = 0.50$?

A correlation of $r = 0.50$ *indicates a* ***moderate positive linear association*** *between father's height and son's height. This means that, in general, taller fathers tend to have taller sons and shorter fathers tend to have shorter sons, though there is still substantial variability.* ***Note:*** *This describes an association, not a causal relationship.*

(2) What will happen to r if we switch X and Y? In other words, what if Y=father's height (inches) and X=son's height (inches)?

(3) What if we change the units of measurement? For example, what if X=father's height (cm) and Y=son's height (cm)?

Example 6 (Continuation of Galton's Data)

Suppose that we are given the following summary statistics from Galton's data: $\bar{X} = 68$, $S_x = 2.7$, $\bar{Y} = 69$, $S_y = 2.7$, and $r = 0.50$.

If a group of fathers are all 70.7 inches tall, (1 S_x above $\bar{X}$), what is our prediction for the mean height among their sons?
Guess: If an X value is 1 S_x above $\bar{X}$, then the corresponding Y values will on average be 1 S_y above $\bar{Y}$.

Fit a simple linear regression model using the information above and interpret the coefficients.

SLR model: $Y = \beta_0 + \beta_1 X + \varepsilon$

Fitted (least square) regression line: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.

Step 1: Compute the estimated slope, $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{rS_y}{S_x} = \frac{0.50 \times 2.7}{2.7} = 0.50.$$

Step 2: Compute the estimated intercept, $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 69 - (0.50)(68) = 69 - 34 = 35.$$

Step 3: Fitted (least square) regression line:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = 35 + 0.50X.$$

Interpretation of the coefficients:

- **Intercept** ($\hat{\beta}_0 = 35$): When the father's height is 0 inches, the model predicts that the expected (average) height of a son would be 35 inches.

  Note: Is this interpretation meaningful in practice? A height of 0 inches is far outside the observed father-height range. In general, we should not **extrapolate** a model far beyond the range of observed data.

- **Slope** ($\hat{\beta}_1 = 0.50$): For every one-inch increase in the father's height, the model predicts an expected (average) increase of 0.50 inches in the son's height.

Using the fitted regression to make predictions:

If $X = 70.7$, then $\hat{Y} = 35 + 0.50(70.7) = \boxed{70.35}$ inches

Comments: This means that when a father's height is 70.7 inches, the model predicts that the expected (average) height of sons with fathers of that height is 70.35 inches.

Note: We use the regression line to predict expected (average) values of Y, not specific individual observations.

Exercise

A student at Case Western Reserve University has designed a new electric clothes dryer. Maybe there is something special about this dryer (for example, it is energy efficient). However, the student is having a problem with the temperature setting and the actual temperature inside the dryer. Let, X: Temperature setting (F) and Y: Actual interior temperature during cycle (F).

The student has carefully measured several readings for X along with the values of Y. The dataset is below:

X	**Y**
140	170
145	174
150	172
155	178
160	176

$$\begin{aligned} \bar{X} &= 150 \\ \bar{Y} &= 174 \\ S_x &= 7.91 \\ S_y &= 3.16 \\ SS_{xx} &= 250 \\ SS_{yy} &= 40 \\ SS_{xy} &= 80 \end{aligned}$$

Answer the following questions:

(1) Draw a scatterplot of the data with X on the horizontal axis and Y on the vertical axis. Describe the overall direction and form of the relationship.

(2) Compute the correlation coefficient r between X and Y using the summary statistics provided. Interpret the value of r in the context of this problem.

(3) What happens to r if we switch the roles of the variables, that is, if X: Actual interior temperature during cycle (F), and Y: Temperature setting (F)

(4) What happens to r if we change the units of measurement to Celsius, that is, if X: Temperature setting (C) and Y: Actual interior temperature during cycle (C).

(5) Fit a simple linear regression model of Y on X.

(a) Find the estimated intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$.

(b) Write down the fitted (least square) regression line.

(c) Interpret the estimated slope in the context of the dryer problem.

(d) Interpret the estimated intercept in the context of the dryer problem, and comment on whether this interpretation is meaningful.

(e) On average, we expect the average actual interior temperature to be _____________ °F if we set the dryer at $X = 152$°F.