Simple Linear Regression

Grinnell College

February 21, 2025

Review

- Scatterplot descriptions
  - form, strength, direction
- Pearson's correlation (r)
  - strength and direction of linear relationship for 2 quant. variables
- Spearman's correlation (ρ)
  - strength and direction of *monotone* relationship
  - more robust to outliers

Basic Idea

Regression is a technique that we can use when there is a linear relationship between 2 quantitative variables.

Regression = creating a line on the scatterplot that best represents the linear relationship we see.

Goal: use the explanatory variable to predict values for the response variable.

- the variable being predicted is the response
- the variable we are using to predict is the explanatory variable ('predictor')

Basic Idea

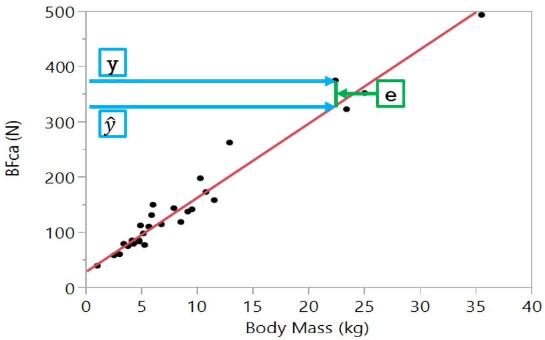
We are going to create a line on the scatterplot that best represents the linear relationship we see.

Algebra
y = mx + b
m = slope: change in y over the change in x (rise / run)
b = intercept: value where the line cross the y-axis
All points fall exactly on the line

Statistics
$\hat{y} = \beta_0 + \beta_1 X$
β_1 = slope
β_0 = intercept
Not all of our data points will exactly on the line $\rightarrow$ variability

How it works

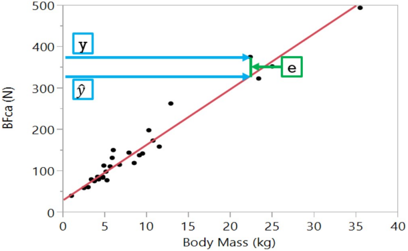
Canidae data set (predicting bite force using body mass)



The **regression line** is the line that fits through the data points.

- y's denote the values of the datapoints for the response variable
- points on the line are predicted values for the y's, denoted as $\hat{y}$
- **residual**: difference between data and predictions ($\mathbf{e} = y - \hat{y}$)

How it works



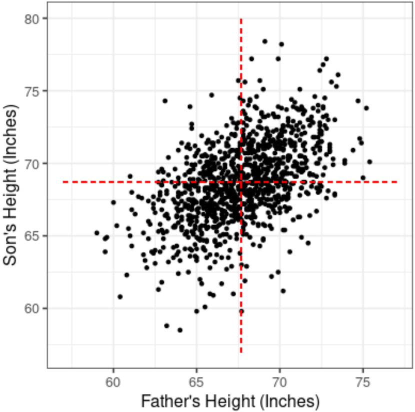
The **regression line** is the line that best fits through the data

- critera: minimizes sum of squared residuals $\sum e_i^2$
- $\hat{y} = b_0 + b_1 X$ (regression equation)
- $b_1 = (\frac{s_x}{s_y})r$ (slope)
- $b_0 = \overline{y} - b_1\overline{x}$ (intercept)

Pearson's Height Data

	Mean	Std.Dev.	Correlation (r_{xy})
Father	67.68	2.74	0.501
Son	68.68	2.81	

Father	Son
65.0	59.8
63.3	63.2
65.0	63.3
65.8	62.8
61.1	64.3
63.0	64.2
⋮	⋮



Pearson's Height Data

We could calculate our regression line using info from this table.

	Mean	Std.Dev.	Correlation (r_{xy})
Father	67.68	2.74	0.501
Son	68.68	2.81	

Regression equation:
$\hat{y} = b_0 + b_1 X$

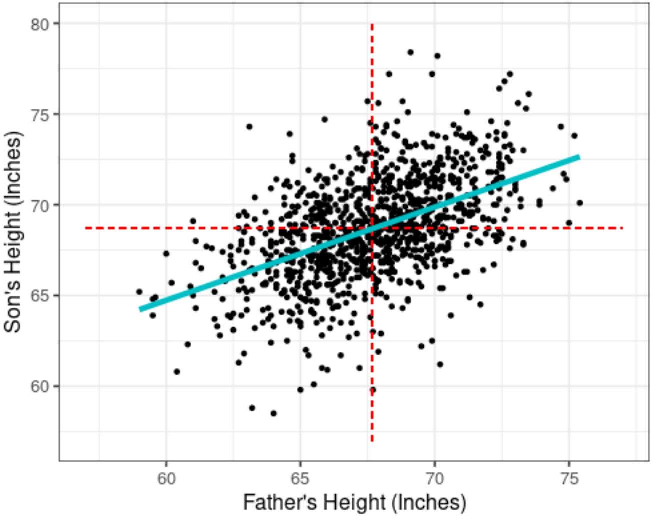
$$b_0 = (\frac{s_x}{s_y})r$$
$$= (\frac{2.81}{2.74})0.501 = 0.514$$

```
> heights <- read.csv("Pearson.tsv", sep = "\t")
> fit <- lm(Son ~ Father, heights)
> fit

Call:
lm(formula = Son ~ Father, data = heights)

Coefficients:
(Intercept)       Father
     33.893        0.514
```

$$b_1 = \overline{y} - b_1\overline{x}$$
$$= 68.68 - 0.514 * 67.68 = 33.893$$

Pearson's Height Data – Plot Line

We can make R graph the line on our scatterplot.



Pearson's Height Data – Prediction

The formula for the regression line

$$\hat{y} = b_0 + Xb_1$$

can be expressed in terms our our original variables and what we wish to predict

$$\widehat{\text{Son's Height}} = 33.9 + 0.51 \times \text{Father's Height}$$

Given the Father's height, we can predict the son's height using this equation by plugging in a value for the father's height

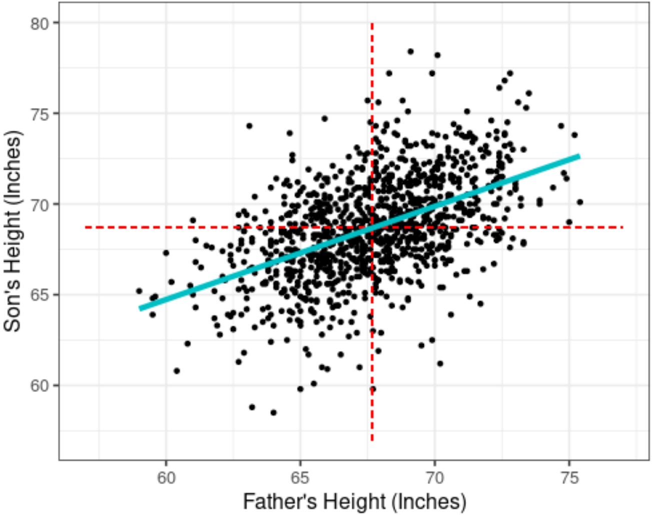
Example: Predict the height of the son for a father with a height of 65in.

$$\widehat{\text{Son's Height}} = 33.9 + 0.51 \times 65.0 = 67.30 in.$$

Pearson's Height Data – Prediction

Predicted Son's Height = 67.30 inches for a father with height = 65in

- Check to see if our prediction makes sense on the graph



Residual

A **Residual** is the difference between an observed value and a prediction

- often labeled as **e** ("error", r is taken)
- e = y - $\hat{y}$

Interpretation: the residual tells us whether we have over- or under-predicted the values for the response variable in our data (and by how much)

- positive value $\rightarrow$ under-predicted
- negative value $\rightarrow$ over-predicted

Pearson's Height Data – Residual

In our data set, the first father had a height of 65 inches. We can calculate the residual for this father. We predicted the son's height to be 67.30 inches.

$$\begin{aligned} e &= y - \hat{y} \\ &= \text{observed value - predicted value} \\ &= 59.8in. - 67.30in. = -7.5in. \end{aligned}$$

Interpretation: We overpredicted the height of this particular son by 7.5 inches

Father	Son
65.0	59.8
63.3	63.2
65.0	63.3
65.8	62.8
61.1	64.3
63.0	64.2
⋮	⋮

Slope Interpretation

Regression equation: $\hat{y} = b_0 + b_1 X$

The **slope** (b_1) tells us how our predictions change when we use different values for the explanatory variable.

Interpretation 1:
For each 1 unit change in the explanatory variable (x), the predicted value of the response variable (y) will change by [value of slope].

Interpretation 2:
For each 1 unit change in the explanatory variable (x), the value of the response variable (y) will change by the [value of slope], on average.

Intercept Interpretation

Regression equation: $\hat{y} = b_0 + b_1 X$

The **intercept** (b_0) is the value where our line crosses the y-axis.

Interpretation: When the explanatory variable (x) is zero, we predict the response variable (y) to have a value of [intercept value].

Ask yourself: Does the intercept interpretation make sense?

- Is the intercept value actually possible for our response variable?
- Does it make sense to make a prediction using zero for the explanatory variable?

Pearson's Height Data – Interpretations

$$\widehat{\text{Son's Height}} = 33.9 + 0.51 \times \text{Father's Height}$$

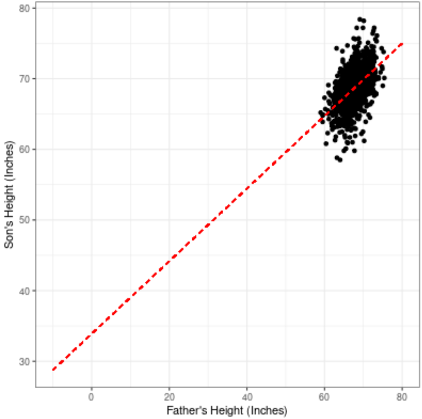
Slope Interpretation:
For each 1 inch change in Father's height, the prediction for son's height changes by 0.51 inches.
-OR-
For each 1 inch change in Father's height, the son's height changes by 0.51 inches, *on average*.

Intercept Interpretation:
When the father's height is zero inches, the predicted height for the son is 33.9 inches.

- does this make sense?

Intercept and Extrapolation

$$\widehat{\text{Son's Height}} = 33.9 + 0.51 \times \text{Father's Height}$$



Extrapolation means making predictions for values outside of our data

- These predictions are unreliable, since we don't know if the relationship is true for these values

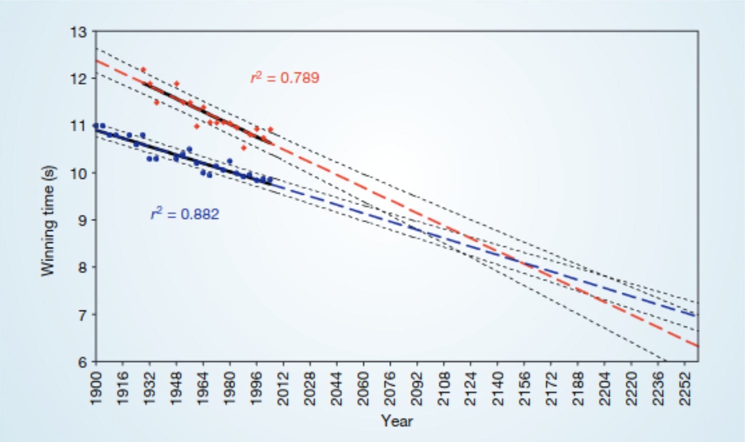
Extrapolation

In 2004, an article was published in *Nature* titled "Momentous sprint at the 2156 Olympics." The authors plotted the winning times of men's and women's 100m dash in every Olympic contest, fitting separate regression lines to each; they found that the two lines will intersect at the 2156 Olympics. Here are a few of the headlines:

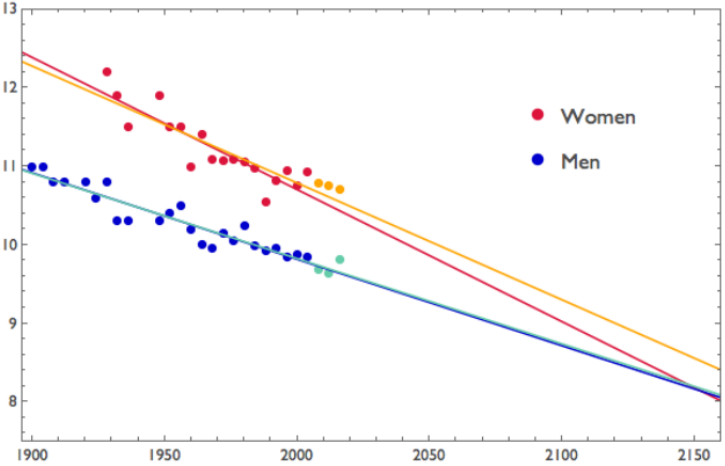
- "Women 'may outsprint men by 2156'" – BBC News
- "Data Trends Suggest Women Will Outrun Men in 2156" – Scientific American
- "Women athletes will one day out-sprint men" – The Telegraph
- "Why women could be faster than men within 150 years" – The Guardian

Momentous sprint at the 2156 Olympics?

Women sprinters are closing the gap on men and may one day overtake them.

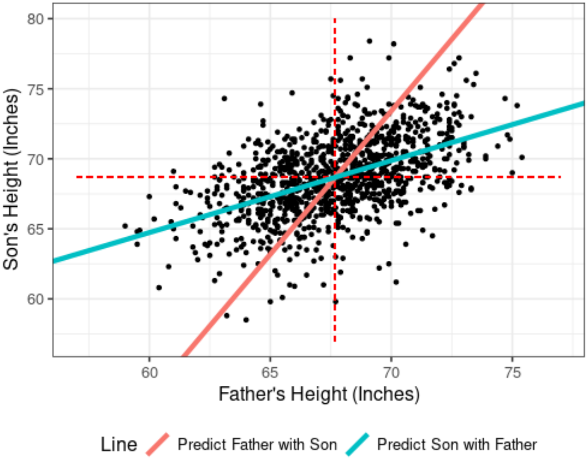


12 years of data later



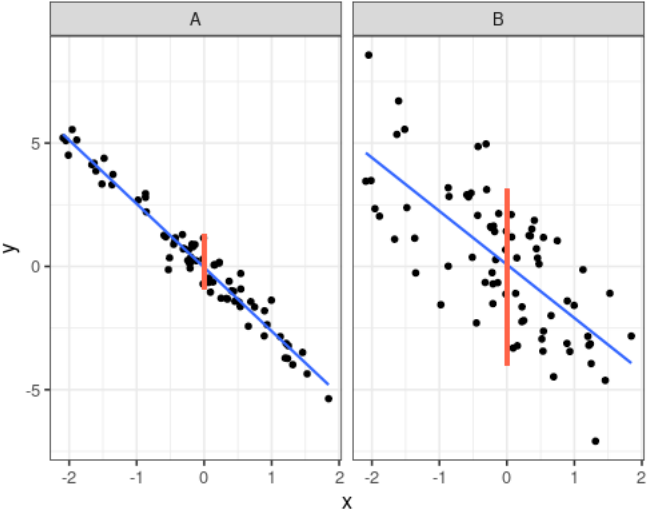
Asymmetry

Unlike correlation, where $r_{xy} = r_{yx}$ (whether you put the variables on the x- or y-axes doesn't matter) regression is *asymmetrical*: the choice of explanatory and response variables matter for the line



Assessing Quality of Fit

The less variability there is for the points around the regression line, the better the line fits the data. (More variability $\rightarrow$ worse fit)



Assessing Quality of Fit

With this in mind, we can quantify how well the line fits the data using:

Coefficient of determination (R^2)

- measures how close the observations match the predictions

$$R^2 = \frac{\text{variance of predicted y's}}{\text{variance of observed y's}} = \frac{s_{\hat{y}}^2}{s_y^2}$$

- ratio written as decimal or percentage between 0% and 100%
- larger values indicate better fit, stronger linear relationship between the variables

Interpretation:
R^2 is the percentage of variation in the observed values of the response variable (x) that can be explained with the linear regression model including the explanatory variable (y). [include context]

Assessing Quality of Fit

We also saw that the **correlation coefficient (r)** can be used to quantify the strength of the linear relationship.

There is a connection between r and R^2.

- $r^2 = R^2$
- $\mathrm{r} = \pm\sqrt{R^2}$ (need to find the correct sign using scatterplot / slope)

R^2 Interpretation

The correlation coefficient for the Pearson Height data is r $= 0.501$

$$R^2 = r^2 = .501^2 = 0.251$$

Interpretation: "25.1% of the variation in son's height can be explained using our linear regression with father's height as the predictor."

$\rightarrow$ 25.1% of the differences in height for sons is because of the their father's height. 74.9% of their differences in height is because of other stuff

Review

We should be able to

- Use a line to describe a linear relationship
- Be able to predict an outcome, given a predictor
- Interpret the slope (and intercept if applicable)
- Assess the quality of a fitted line using R^2