Simple Linear Regression

Grinnell College

September 20, 2024

Z-scores and Correlation

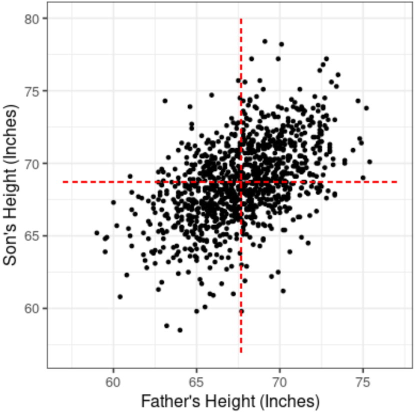
Recall that:

- **Z-scores** or **standardized scores** relate each observation to the mean and standard deviation of the variable
- **Correlation** specifies the *linear* relationship between two quantitative variables

Pearson's Height Data

	Mean (μ)	SD (σ)	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
⋮	⋮



Regression towards the mean

	Mean (μ)	SD (σ)	Correlation (r_{xy})
Father	67.68	2.74	0.501
Son	68.68	2.81	

The correlation coefficient tells us how much "regression" we expect to observe in terms of standardized values:

$$z_S = r \times z_F$$

If the father is one and a half standard deviations aove average ($z_F = 1.5$), and the correlation between heights is 0.501, we have:

$$\begin{aligned} z_S &= r \times z_F \\ &= 0.501 \times 1.5 \\ &= 0.752 \end{aligned}$$

Correlation and Prediction

	Mean (μ)	SD (σ)	Correlation (r_{xy})
Father	67.68	2.74	0.501
Son	68.68	2.81	

From here, we can back substitute the value for z_S to get our unstandardized predictions:

$$
\begin{aligned}
z_S &= 0.752 \\
\left(\frac{\hat{y} - 68.68}{2.81}\right) &= 0.752 \\
\hat{y} &= 0.752 \times 2.81 + 68.68 \\
\hat{y} &= 70.793
\end{aligned}
$$

Regression Line

The relationship $z_y = r \times z_x$ can always be manipulated to rewrite the relationship between the variables X and y so they fit the formula

$$\hat{y} = \hat{\beta}_0 + X\hat{\beta}_1$$

We interpret these as follows:

- $\hat{\beta}_0$ represents the *intercept*, or the estimated value of y when $X = 0$
- $\hat{\beta}_1$ represents the *slope*, indicating the magnitude of change in y given a unit change in X

Predictions

The formula for the regression line

$$\hat{y} = \beta_0 + X\beta_1$$

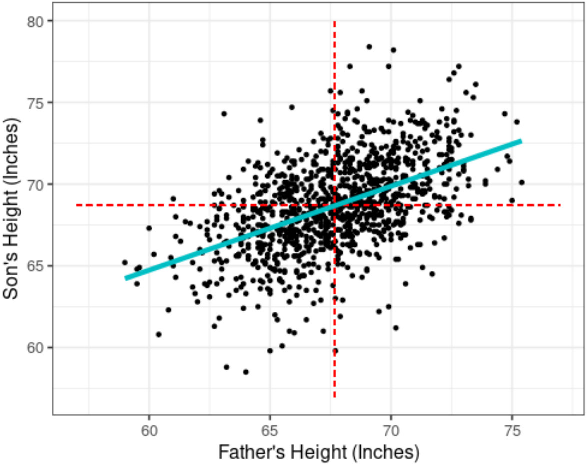
can be expressed in terms our our original variables and what we wish to predict

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

From this, there are a few things about lines we can observe:

- Using this line, *given* the Father's height, we can predict the son's height using this line by plugging in a value for the father's height
- "For each 1 inch change in Father's height, we expect to see a 0.51 inch change in Son's height"
- Intercept interpretation

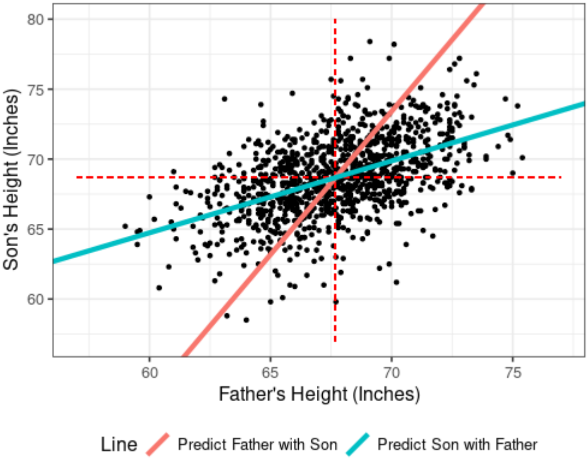
Using Correlation to Make Predictions



"*Given father's height*, the average height of the son is..."

Symmetry

Unlike correlation, where $r_{xy} = r_{yx}$, regression is *asymmetrical*: the choice of explanatory and response variables matter



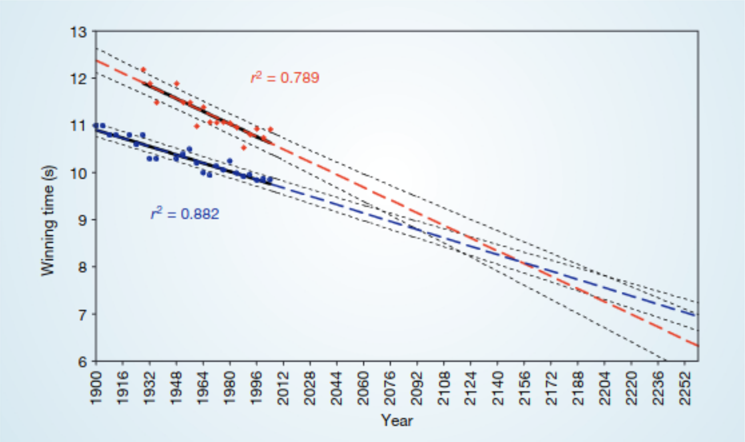
Extrapolation

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:

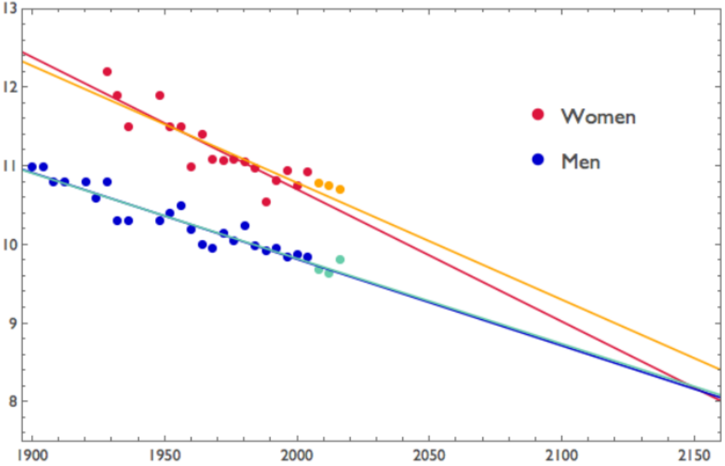
- "Women 'may outsprint men by 2156'" – BBC News
- "Data Trends Suggest Women Will Outrun Men in 2156" – Scientific American
- "Women athletes will one day out-sprint men" – The Telegraph
- "Why women could be faster than men within 150 years" – The Guardian

Momentous sprint at the 2156 Olympics?

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

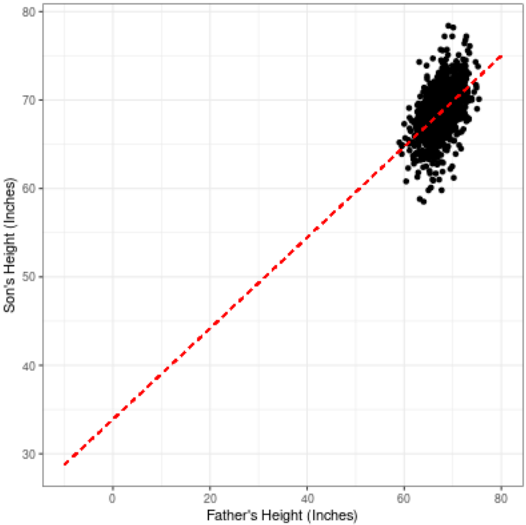


12 years of data later



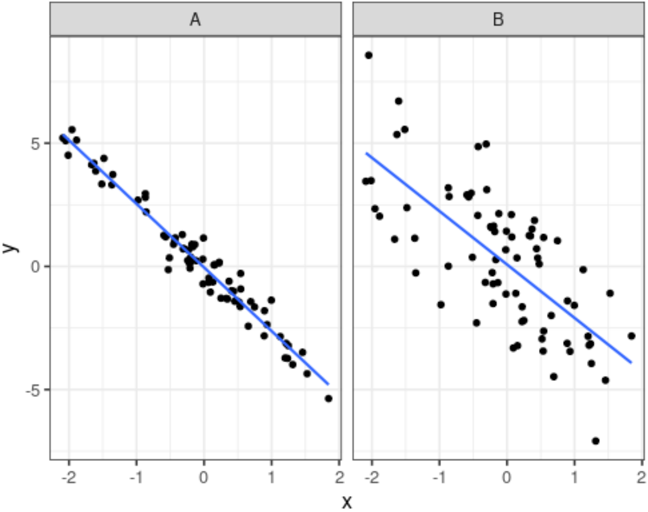
Intercept Interpretation/Extrapolation

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



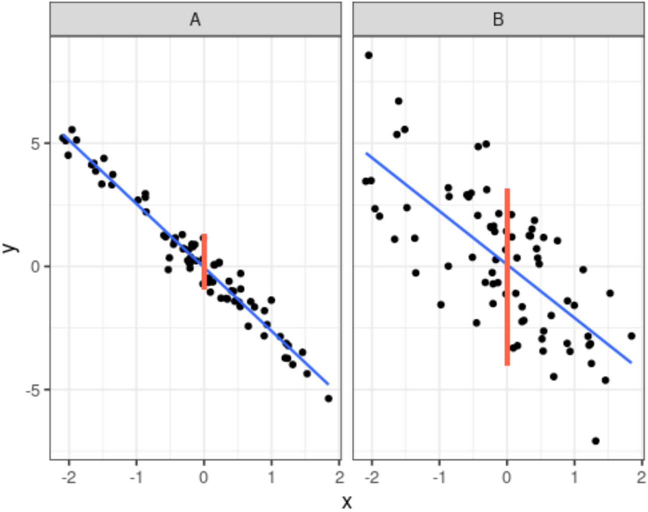
Assessing Quality of Fit

"How much variability is left once I have selected my prediction on the line?"



Assessing Quality of Fit

"How much variability is left once I have selected my prediction on the line?"



Total Sum of Squares

If we had an outcome y and no predictor variable x, our best guess for an estimate of y would simply by the mean, $\overline{y}$

From this, we get a sense of the *total variance* by taking the *sum of squares*:

$$\text{Total Sum of Squares} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

We can think of this as our baseline: this is how much variability we see with no other predictors

Regression Sum of Squares

Now assume for each y_i we used a variable x_i, along with their correlation, to create an estimated value $\hat{y}_i$, with

$$\hat{y}_i = \beta_0 + \beta_1 x_1$$

We could then ask ourselves: how much variability is left once I have used my predictor to make $\hat{y}_i$? This gives us the *residual sum of squares*:

$$\text{Residual Sum of Squares} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Coefficient of Determination

Now consider the ratio of variance explained in model against variance without model:

$$\frac{\text{Residual SS (SSR)}}{\text{Total SS (SST)}} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

If our model is no better than guessing the average (i.e., if $\hat{y} = \overline{y}$), this ratio would be 1; if we are able to perfectly predict each value y_i, this ratio would be 0

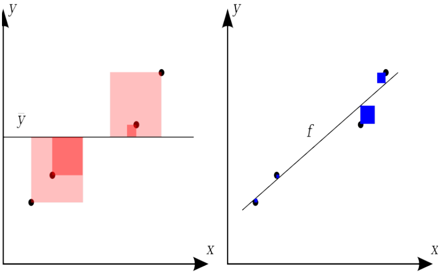
Our **coefficient of determination** or R^2 (R-squared) is defined as

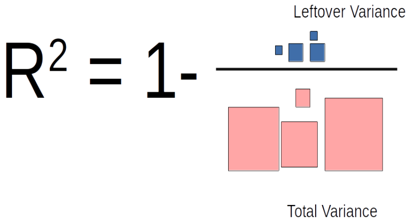
$$R^2 = 1 - \frac{SSR}{SST}$$

Somewhat surprisingly, in the case with a single predictor variable we have that the coefficient of determination is simply the squared correlation

$$R^2 = r^2$$

$$\frac{\text{Residual SS (SSR)}}{\text{Total SS (SST)}} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$





Review

We should be able to

- Describe how correlation and regression related
- Be able to predict an outcome, given a predictor
- Interpret the slope and intercept (if applicable)
- Assess the quality of a fitted line