Simple Linear Regression

Grinnell College

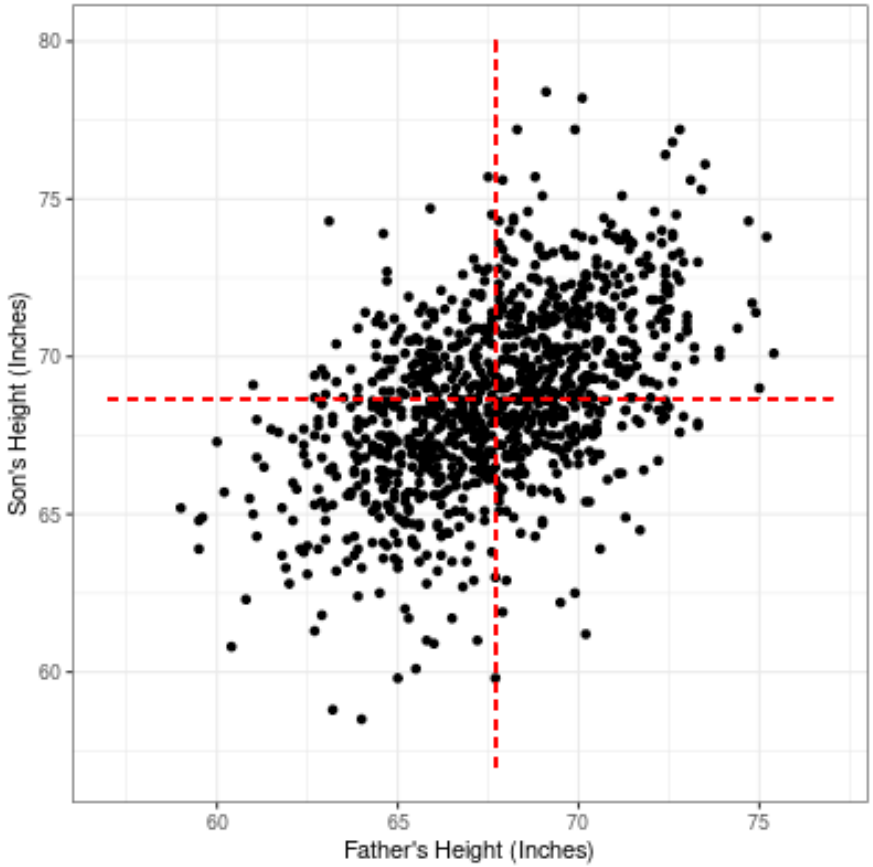
February 14, 2024

Review

- Measures of centrality
- Measures of spread
- Robust statistics
- Conditional Tables
- Standardization

Pearson's Height Data

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
⋮	⋮



z-scores and corrleation

Recall the relationship between standardized scores and the correlation coefficient:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (z_{x_i})(z_{y_i})$$

Correlation and Prediction

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

If the father's height was exactly average at 67.68 inches, it would be perfectly reasonable to predict the son to be exactly average as well

What if the father was one standard deviation below average?

Regression towards the mean

Correlation and Prediction

	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 standard deviation below average ($z_F = -1$), and the correlation between heights is 0.5, we have:

$$\begin{aligned} z_S &= r \times z_F \\ &= -0.5 \end{aligned}$$

Correlation and Prediction

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

$$z_S = -0.5$$

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

$$\begin{aligned} z_S &= -0.5 \\ \left(\frac{\hat{y} - 68.68}{2.81}\right) &= -0.5 \\ \hat{y} &= -0.5 \times 2.81 + 68.68 \\ \hat{y} &= 67.275 \end{aligned}$$

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

$$y = \beta_0 + X\beta_1$$

where the slope is the correlation multiplied by the ratio of standard deviations:

$$\beta_1 = r_{xy}\frac{s_y}{s_x}$$

and the intercept is a difference in means:

$$\beta_0 = \overline{y} - \overline{x}\beta_1$$

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:

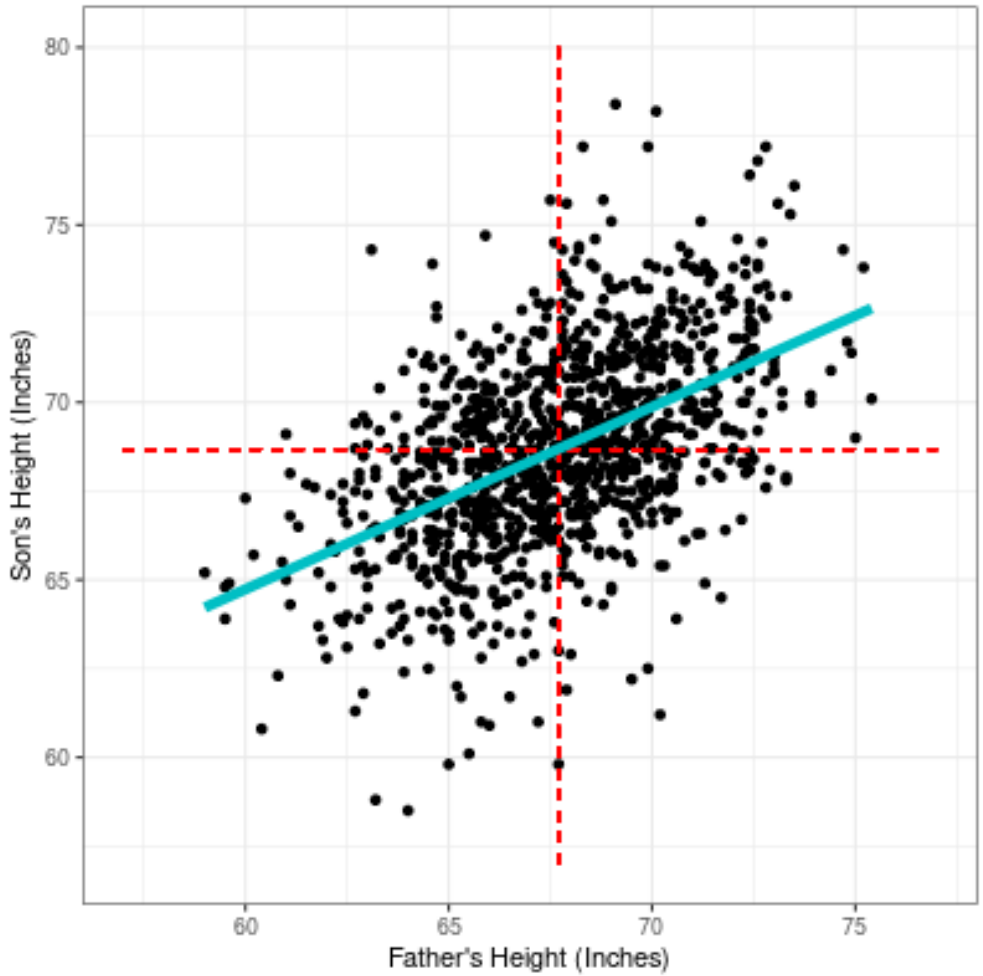
- Using this line, *given* the Father's height, we can predict the son's height using this line by plugging in a value for the father's height
- "For each 1 inch change in Father's height, we expect to see a 0.51 inch change in Son's height"
- Intercept interpretation

Wikipedia Quote

If $-1 < r_{xy} < 1$, then we say that the data points exhibit regression toward the mean. In other words, if linear regression is the appropriate model for a set of data points whose sample correlation coefficient is not perfect, then there is regression toward the mean. The predicted (or fitted) standardized value of y is closer to its mean than the standardized value of x is to its mean.

Using Correlation to Make Predictions

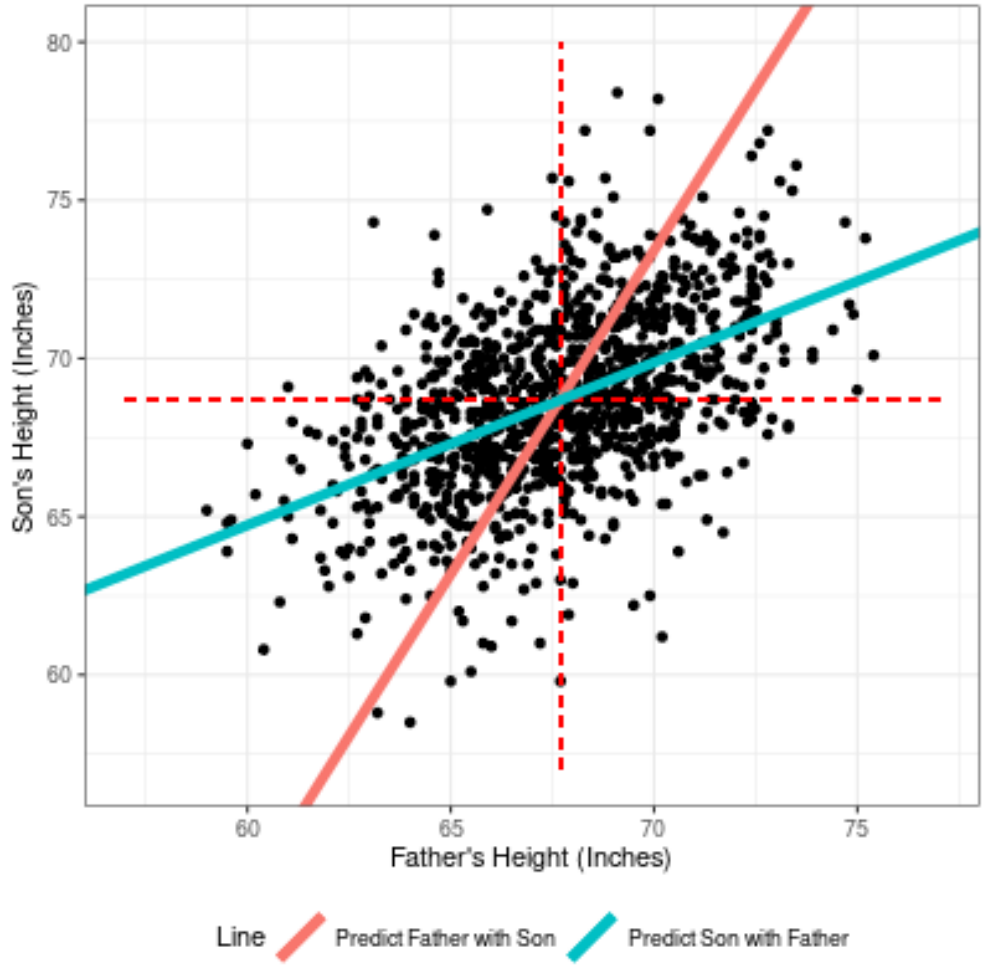
We can use this to make predictions for *any* father's height:



"*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

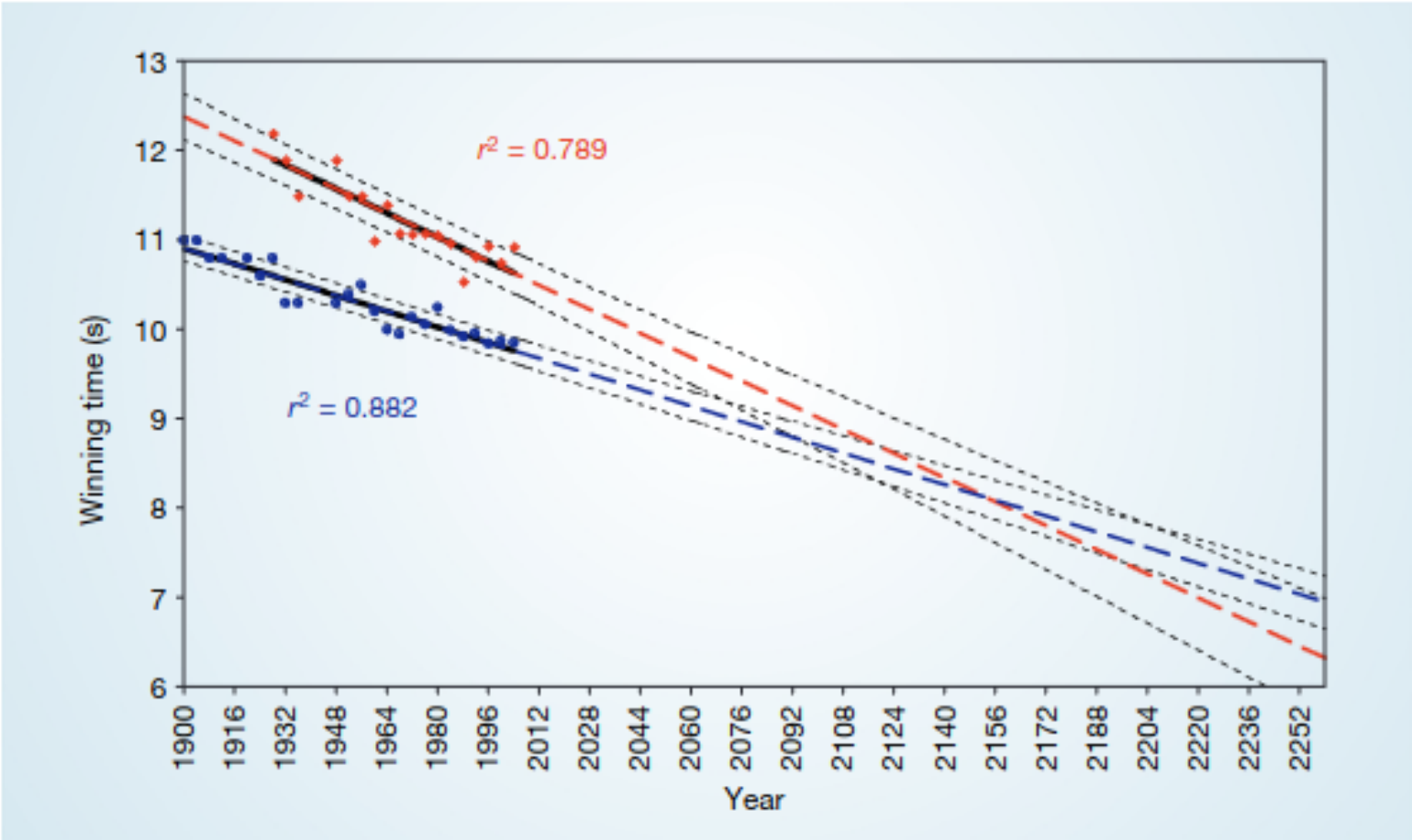
In 2004, an article was published in *Nature* titled "Momentous spring at the 2156 Olympics." The authors plotted the winning times of men's and women's 100m dash in every Olympic contest, fitting separate regression lines to each; they found that the two lines will intersect at the 2156 Olympics. Here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

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

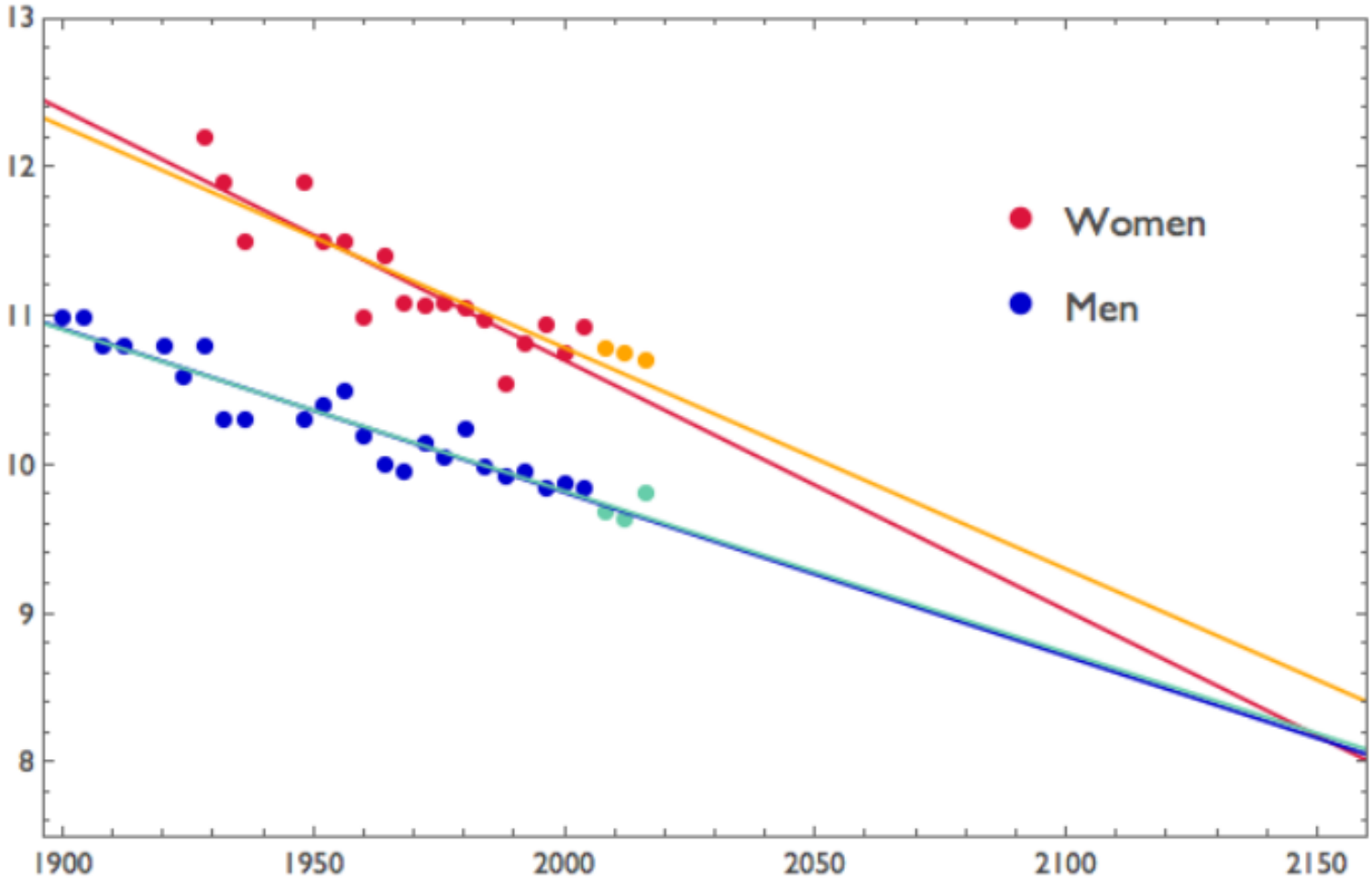
Do you see any issues with these conclusions?

Momentous sprint at the 2156 Olympics?

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

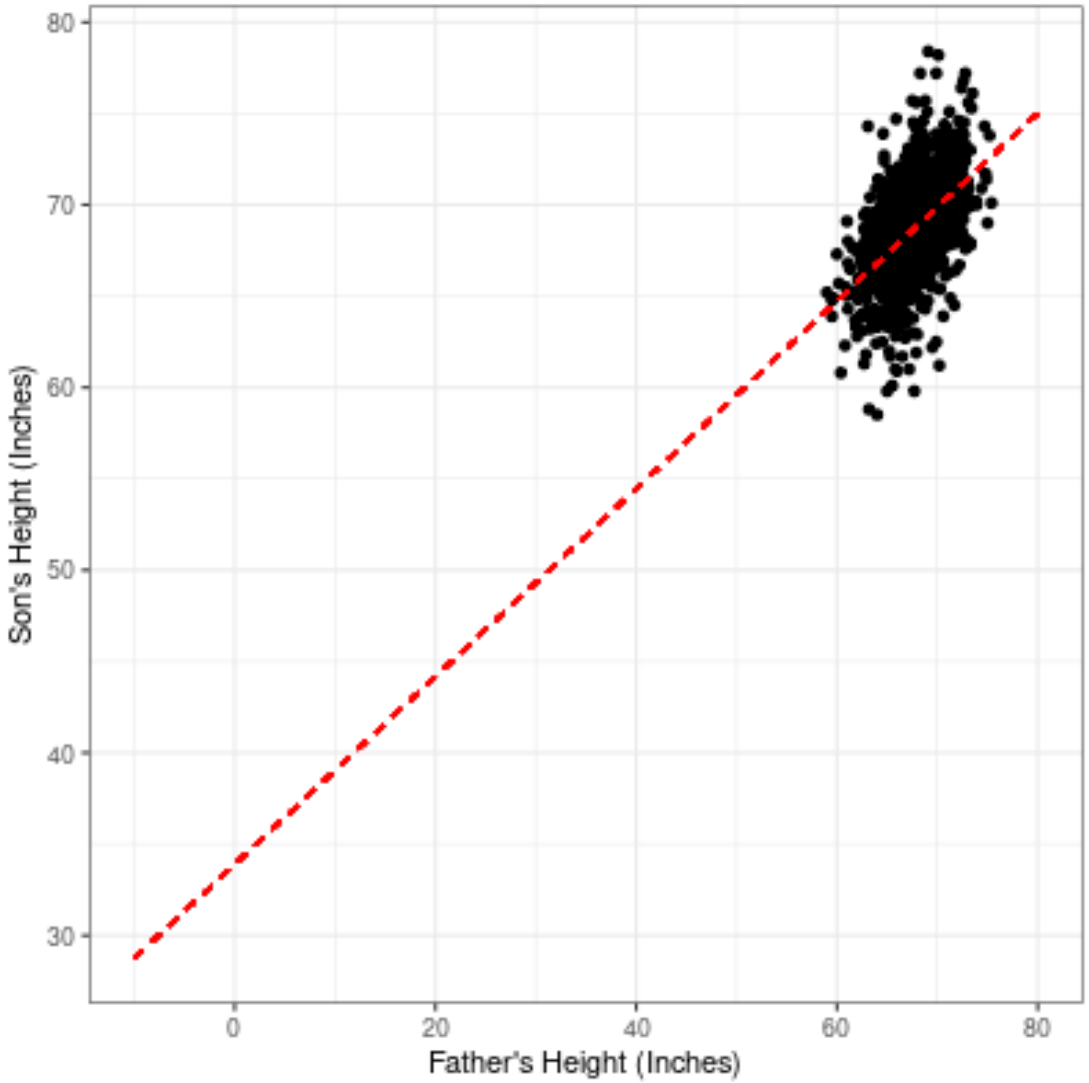


12 years of data later

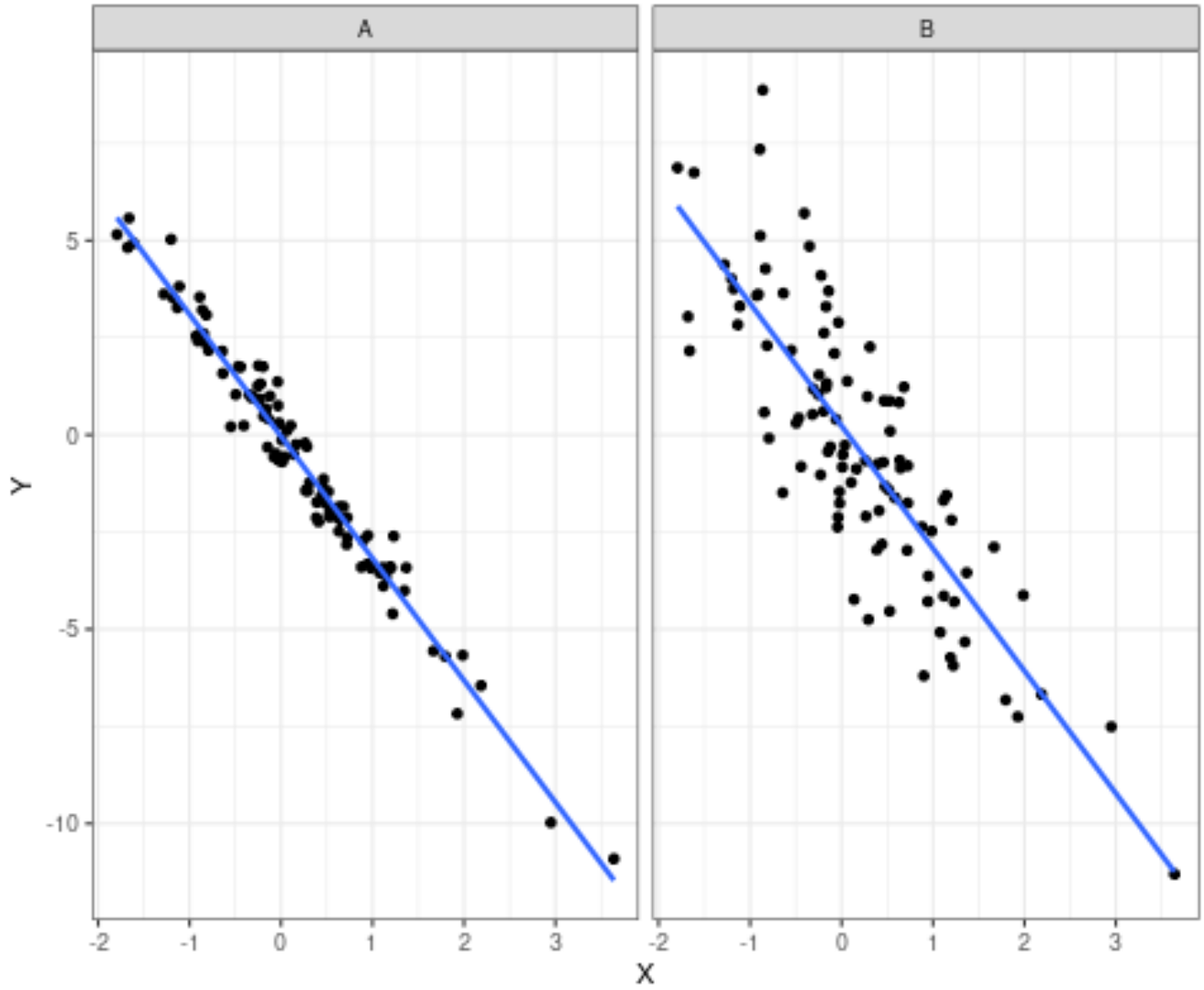


Intercept Interpretation/Extrapolation

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

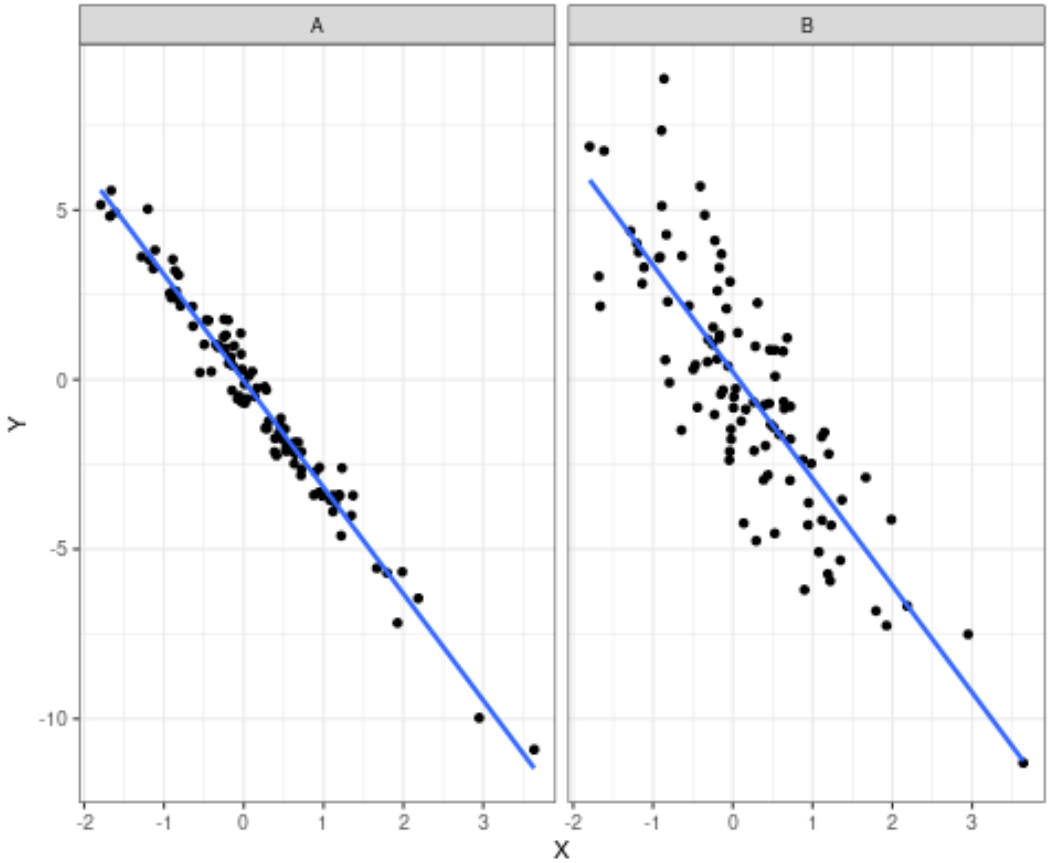


Assessing Quality of Fit



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

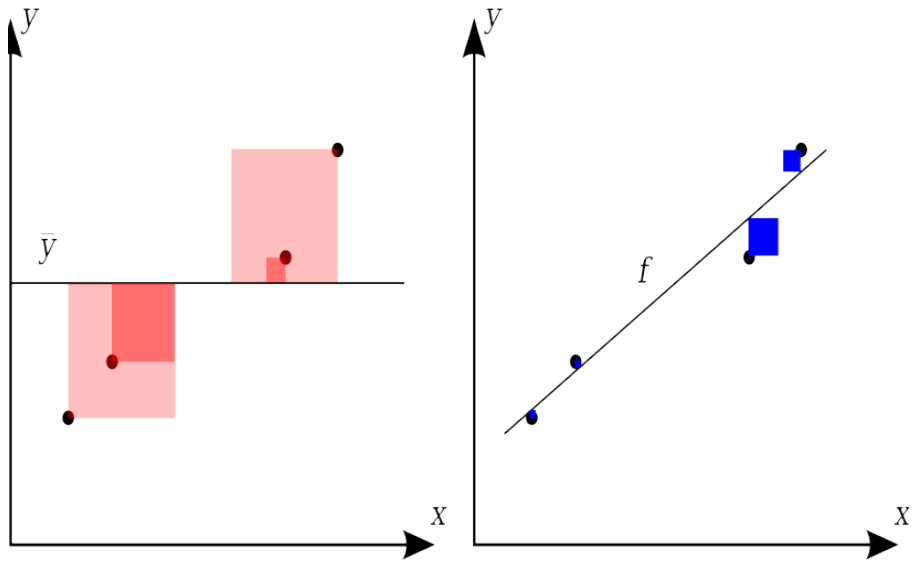
Our **coefficient of determination** or R^2 (R-squared) is defined as

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

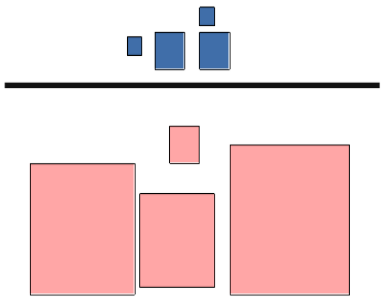
Somewhat surprisingly, in the case with a single predictor variable we have that the coefficient of determination is simply the squared correlation

$$R^2 = r^2$$

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



$$R^2 = 1 - \frac{\text{Leftover Variance}}{\text{Total Variance}}$$



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