Simple Linear Regression

Ryan Miller

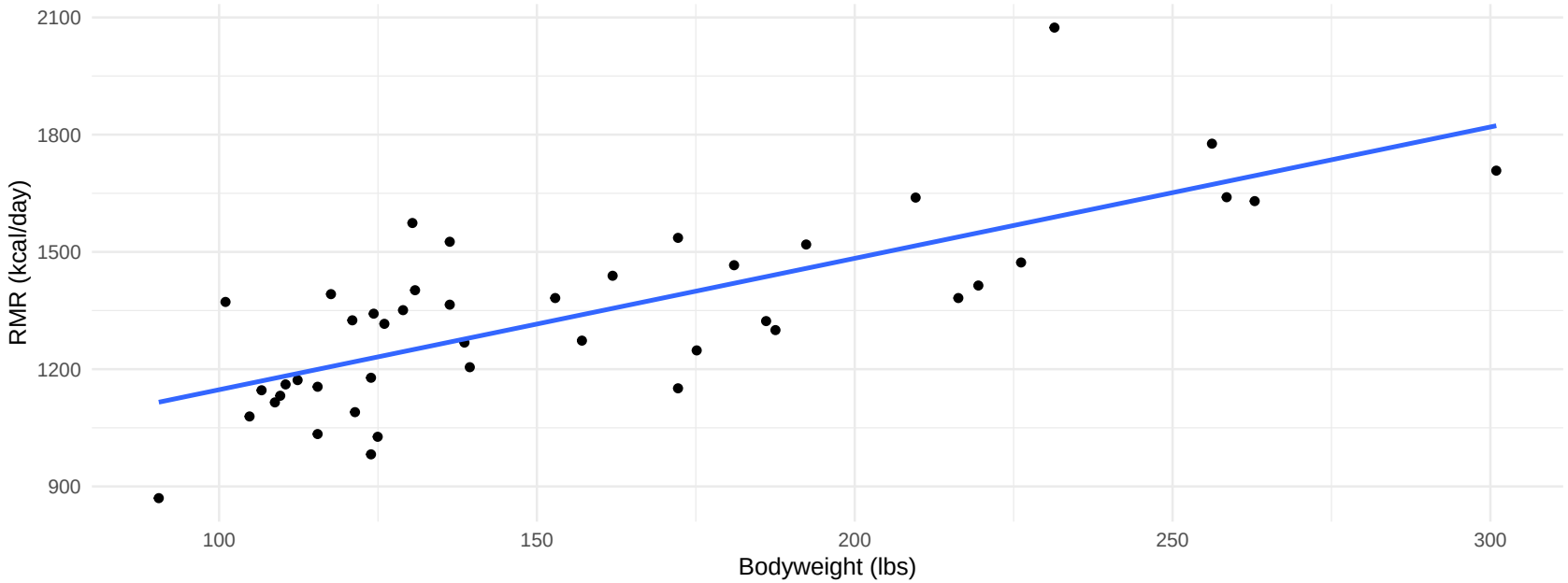
Grinnell College
Statistics

Motivation

- Pearson's correlation coefficient allows us to quantify the strength of a linear association between two variables
  - But in situations with a clear explanatory and response variable, correlation doesn't tell us *how* a change in the explanatory variable impacts the response variable
  - For example, if $r = -0.7$ we know an increase in the explanatory variable should lead to a decrease in the response, but without more information we do not know *how much* of a decrease to expect

Simple Linear Regression

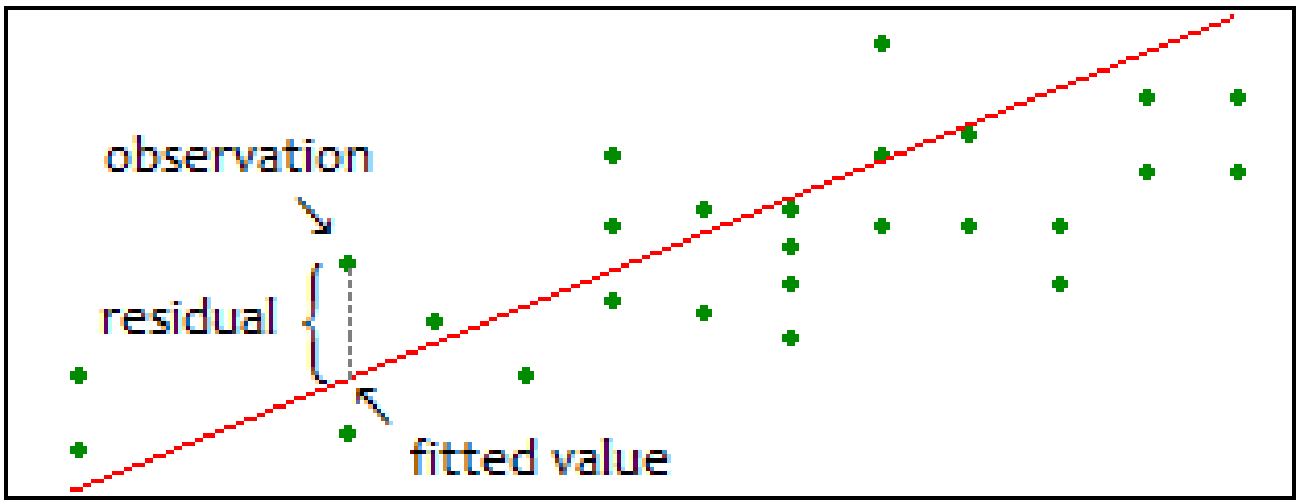
Simple linear regression is a *model* used to represent a *linear relationship* between a quantitative explanatory variable and a quantitative response variable



As a line, the model is defined by a **slope**, b_1, and an **intercept**, b_0

Simple Linear Regression

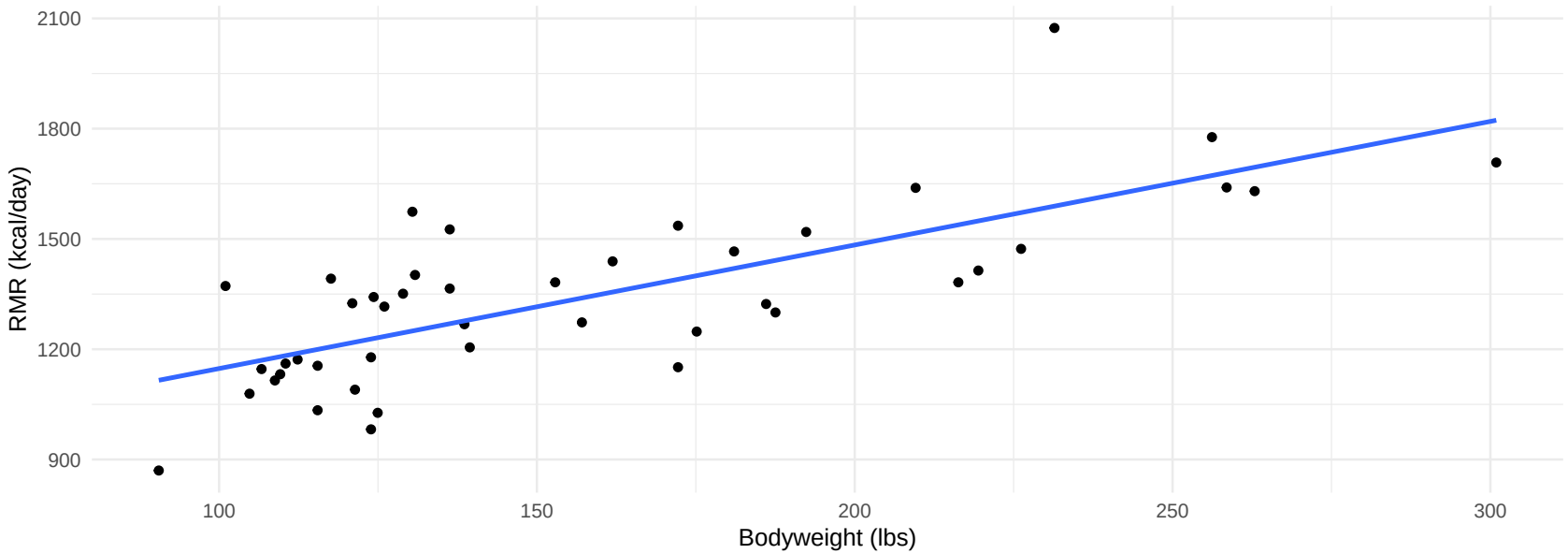
There are infinitely many lines, we want the one with the *smallest sum of squared residuals*



- The **residual** of the i^{th} data-point is the difference: $y_i - \hat{y}_i$, where $\hat{y}_i = b_0 + b_1 x_i$ is the model's prediction
  - Residuals reflect the "errors" made by the model, we want the model with the smallest overall amount of error

Simple Linear Regression

We'll rely upon R to find the slope and intercept corresponding to the smallest amount of error:



- Here the fitted regression equation is: $\hat{y} = 881.2 + 3.4 * \text{Weight}$
  - What does the estimated *intercept*, 881.2, tell us?
  - What does the estimated *slope*, 3.4, tell us?

Coefficient of Variation

The **coefficient of variation**, R^2, is a popular measure of how much "information" an explanatory variable (or set of multiple explanatory variables) contains about a response variable:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

- If the explanatory variable provides very little information, its predictions, $\hat{y}_i$, will be similar to the average value of the response variable, $\overline{y}$, regardless of the value of X
  - In this scenario, the line's slope is approximately zero and $R^2 \approx 0$
- If the regression line perfectly coincides with the observed data, $\hat{y}_i = y_i$ for all cases, and $R^2 = 1$

Correlation vs. Regression and R^2

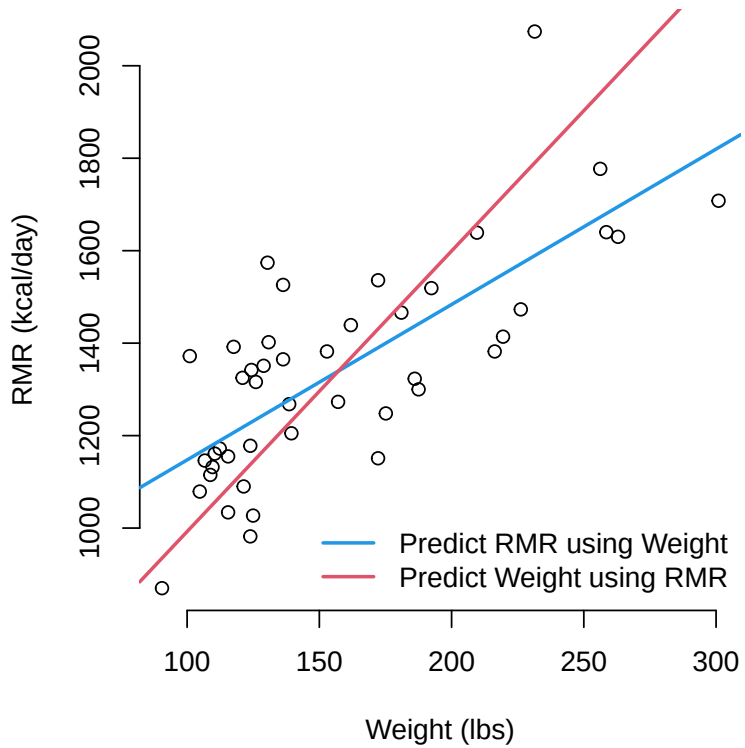
- For simple linear regression, R^2 equals Pearson's correlation coefficient squared (ie: $R^2 = r^2$)
- Correlation is *symmetric*
  - The correlation between RMR and Weight is the *same* as the correlation between Weight and RMR

Correlation vs. Regression and R^2

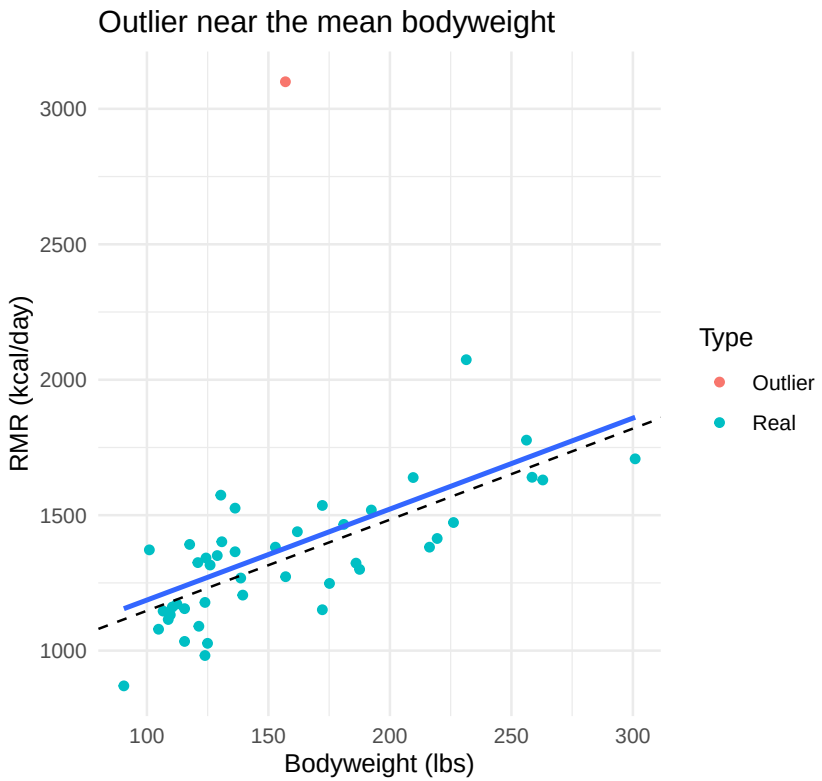
- For simple linear regression, R^2 equals Pearson's correlation coefficient squared (ie: $R^2 = r^2$)
- Correlation is *symmetric*
  - The correlation between RMR and Weight is the *same* as the correlation between Weight and RMR
- Regression is *asymmetric*
  - The regression model that uses RMR to predict Weight is *different* from the model that uses Weight to predict RMR
  - The choice of explanatory and response variable matter for regression! They do not for correlation!

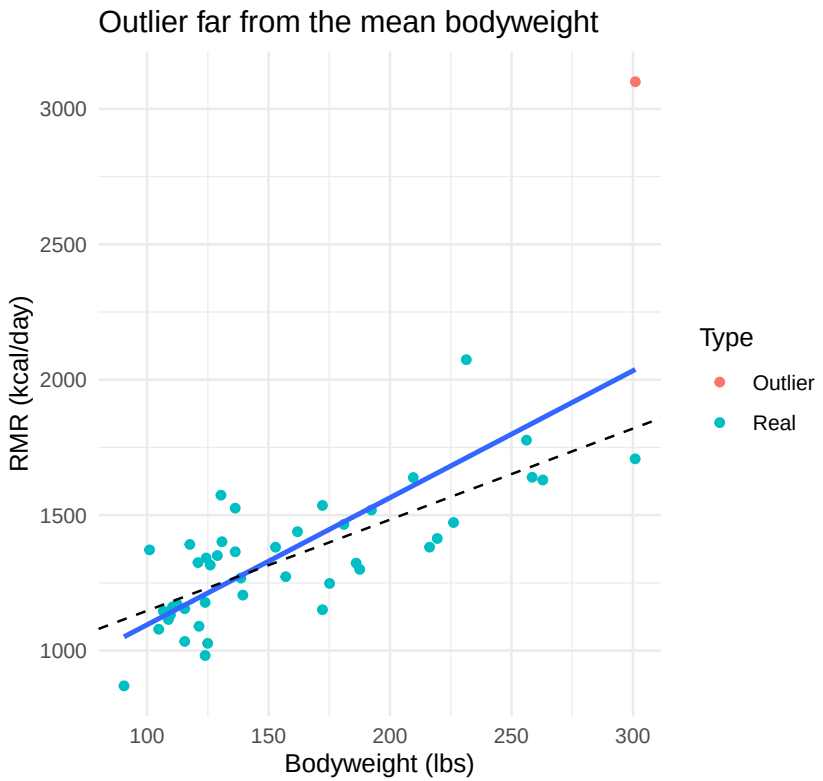
Caution #1 - Two Regression Lines

Below are the two choices of explanatory and response variables for the RMR dataset:



Caution #2 - Outliers and Influence





Outliers in the response variable only exert a disproportionate impact on the regression line if they are also far from the average value of the explanatory variable

Caution #3 - Extrapolation

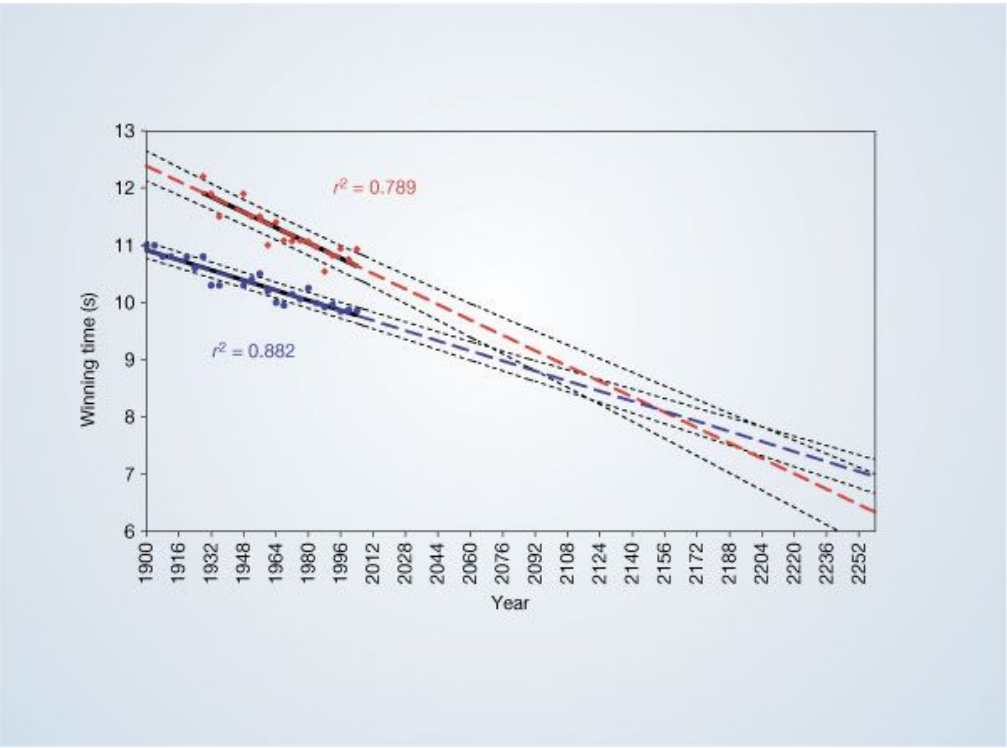
In 2004, an article was published in *Nature* titled "Momentous sprint at the 2156 Olympics". The authors plotted the winning times of the men's and women's 100m dash in every Olympics, fitting separate regression lines to each. They found that the lines will intersect at the 2156 Olympics, here are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

Do you have any problems with these conclusions?

Extrapolation

Here's a figure from the original publication in Nature:

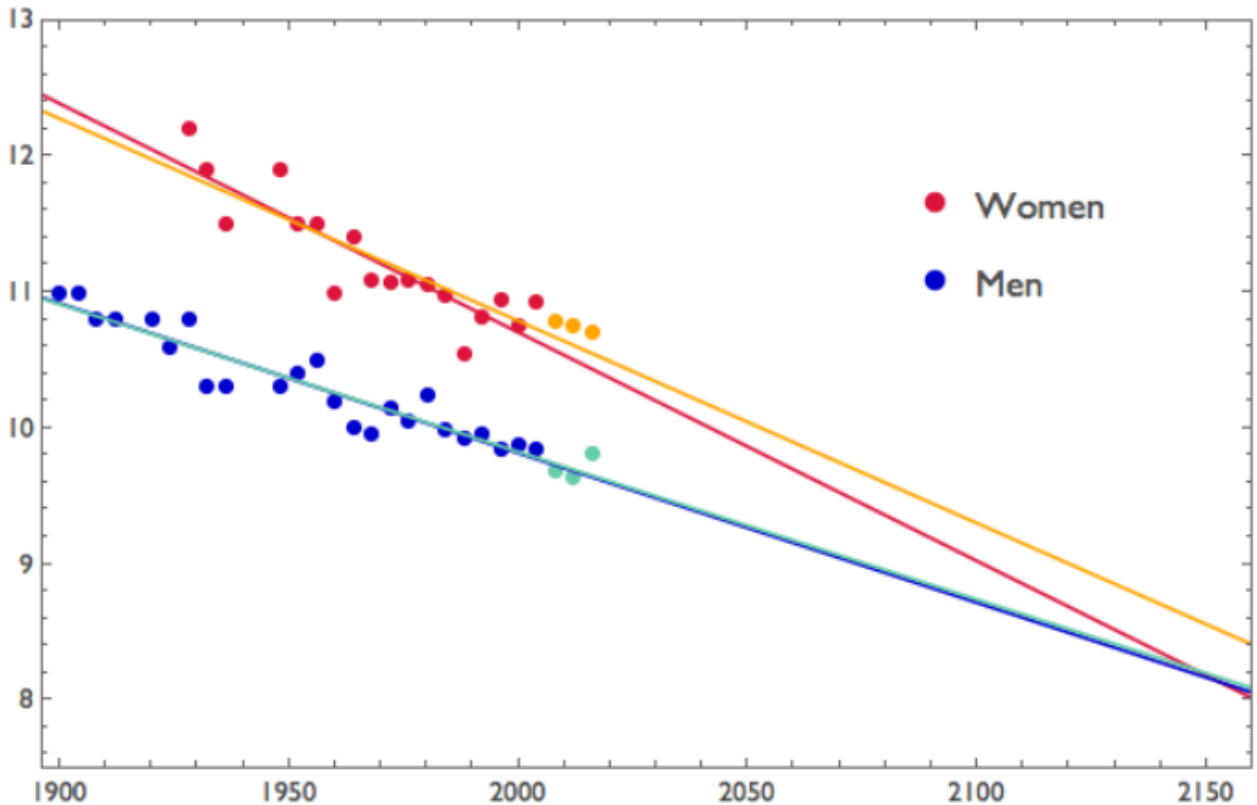


The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

Extrapolation

source: https://callingbullshit.org/case_studies/case_study_gender_gap_running.html

Since the *Nature* paper was published, we've had three additional Olympic games. It is interesting to add the results from those three games (yellow and green points below) and see how the model has performed.



Conclusion

- Simple linear regression provides us an asymmetric tool for describing the relationship between two quantitative variables
  - We'll soon see that this framework can be extended to encompass more complex models involving several variables
- Simple linear regression assumes a straight-line relationship, with software like R estimating the slope and intercept that produce the best-fitting line for the observed data
  - The sum of squared residuals determines the best fitting model
- When using simple linear regression we should be careful about our choice of explanatory variable, aware of the role of outliers, and we should avoid extrapolation