Regression

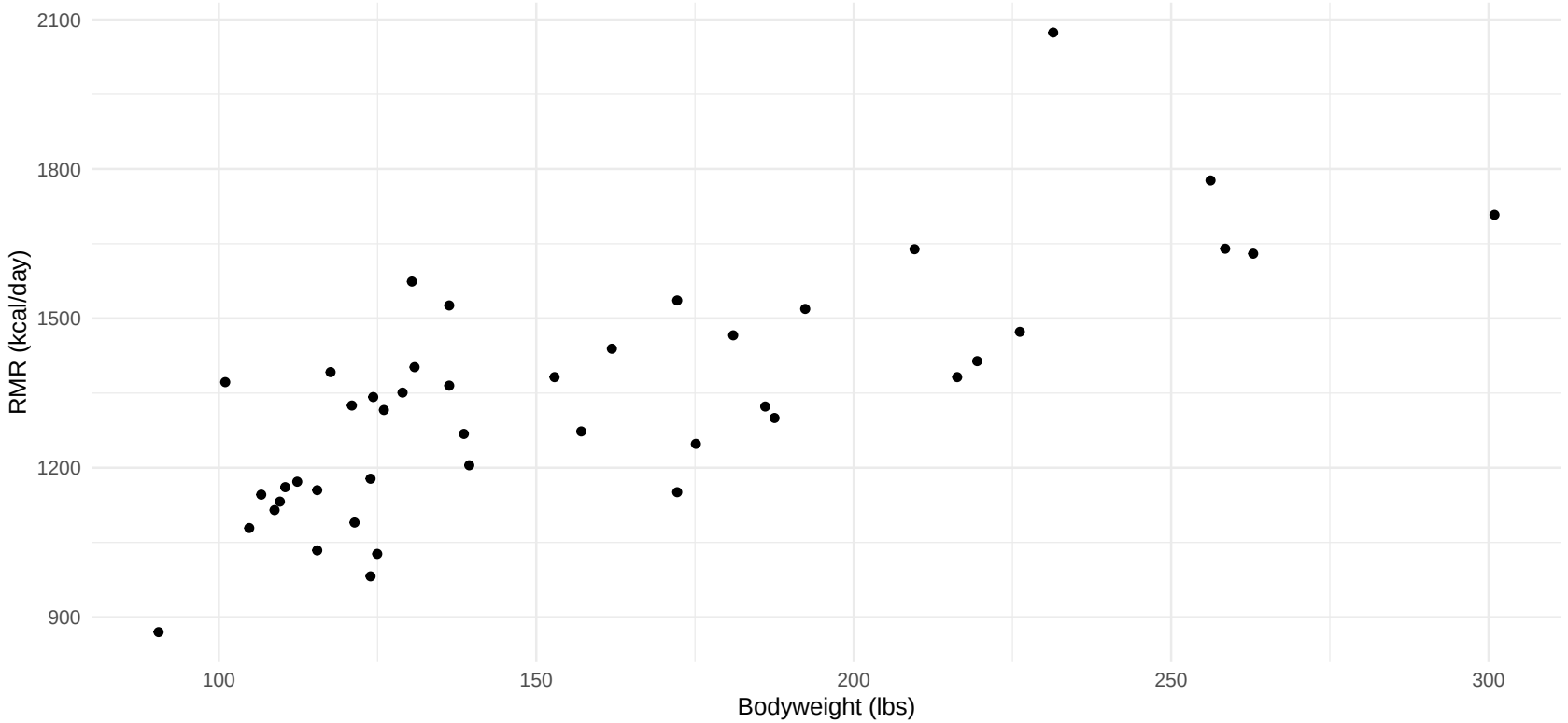
Ryan Miller

Outline

1. Simple linear regression
   - Model coefficients and interpretations, R^2
2. Relationship with Pearson's correlation coefficient
3. Common pitfalls
   - Highly influential data-points, extrapolation

Bodyweight and resting metabolism

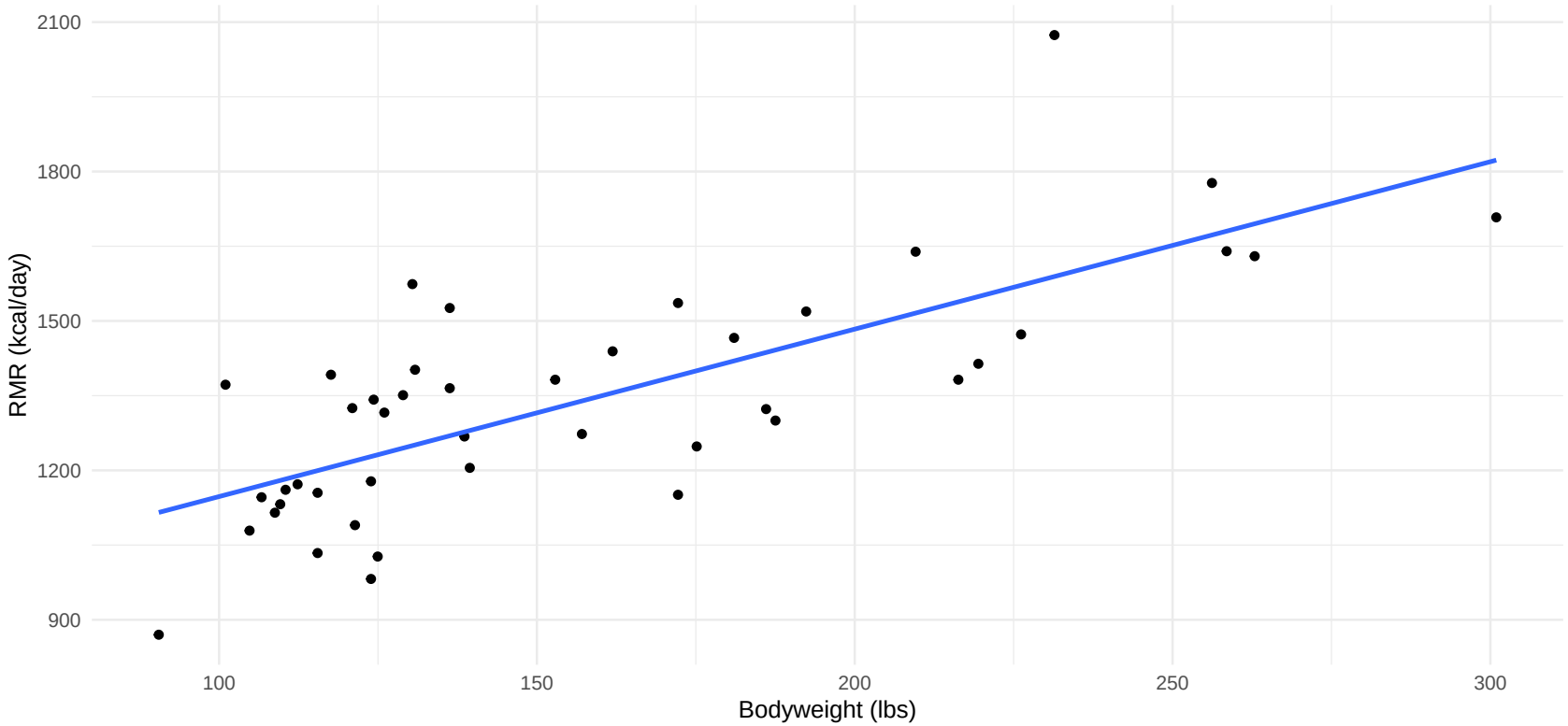
Depicted below are the bodyweight and resting metabolic rate of 44 adult women:



1) Describe the relationship between these variables
2) How much higher do expect the RMR of a woman weighing 175 lbs to be relative to a woman weighing 125 lbs?

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



Being a line, the model is defined by its **slope** and its **intercept**

Simple linear regression

For the RMR data, the *fitted* or *estimated* simple linear regression line equation is:

$$\widehat{\text{RMR}} = 881.2 + 3.4 * \text{Weight}$$

- This equation is the straight-line that is the *best fit* for *these data*

Simple linear regression

For the RMR data, the *fitted* or *estimated* simple linear regression line equation is:

$$\widehat{\text{RMR}} = 881.2 + 3.4 * \text{Weight}$$

- This equation is the straight-line that is the *best fit* for *these data*
- The estimated slope, $b_1 = 3.4$, suggests that every 1 pound increase in weight is expected to result in a 3.4 kcal/day increase in resting metabolic rate

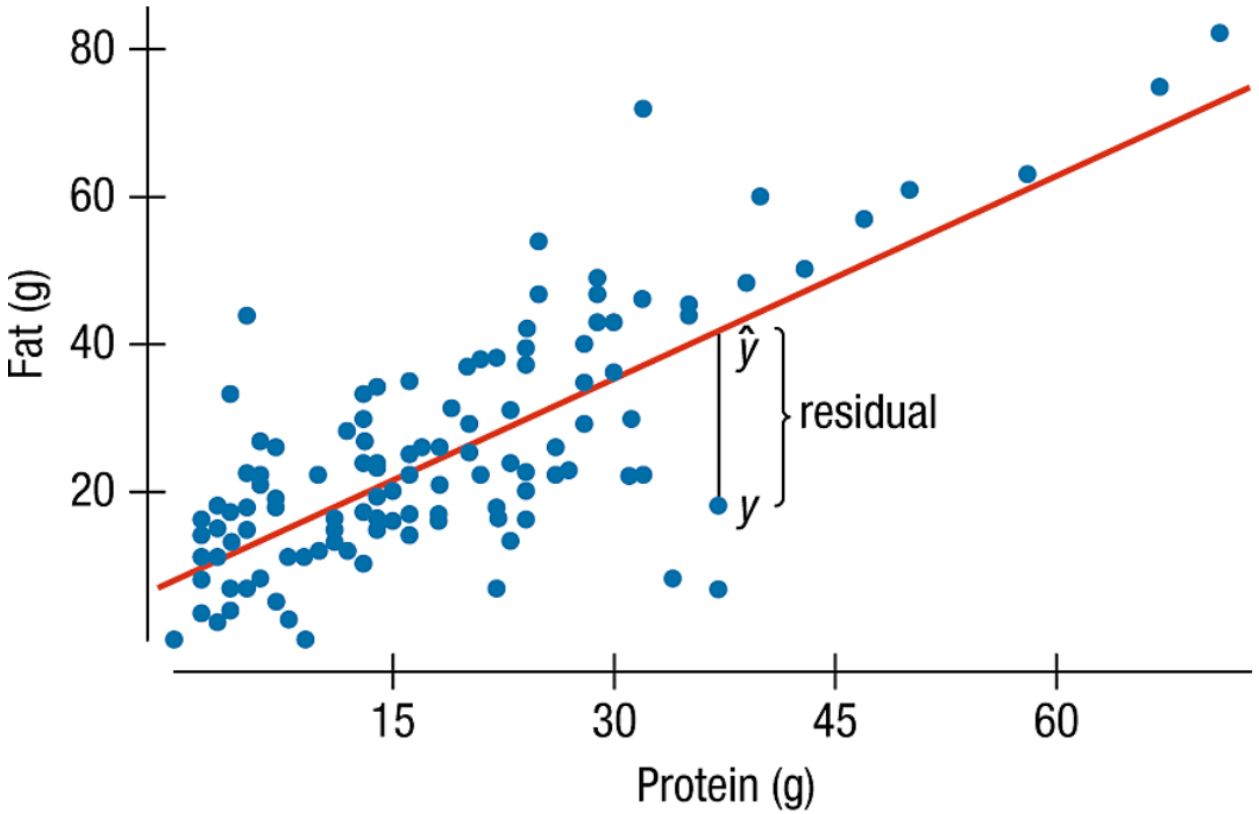
Simple linear regression

For the RMR data, the *fitted* or *estimated* simple linear regression line equation is:

$$\widehat{\text{RMR}} = 881.2 + 3.4 * \text{Weight}$$

- This equation is the straight-line that is the *best fit* for *these data*
- The estimated slope, $b_1 = 3.4$, suggests that every 1 pound increase in weight is expected to result in a 3.4 kcal/day increase in resting metabolic rate
- The intercept, $b_0 = 881.2$, is the expected resting metabolic rate of woman who weights zero lbs (ie: meaningless), but we need it to define the line

Least squares estimation



- For each case, its **residual** is defined as $y - \hat{y}$, or the model's prediction minus the observed y-value
- *Least squares estimation* finds the best fitting regression line by finding the combination of slope and intercept that minimize the *sum of squared residuals*

Prediction

Simple linear regression models can also be used to generate predictions

$$\widehat{\text{RMR}} = 881.2 + 3.4 * \text{Weight}$$

- For example, a woman who weights 125 lbs is expected to have a resting metabolic rate equal to $881.2 + 3.4 * 125 = 1306.2$ kcals

Prediction

Simple linear regression models can also be used to generate predictions

$$\widehat{\text{RMR}} = 881.2 + 3.4 * \text{Weight}$$

- For example, a woman who weights 125 lbs is expected to have a resting metabolic rate equal to $881.2 + 3.4 * 125 = 1306.2$ kcals
- A woman who weighs 175 lbs is expected to have a resting metabolic rate equal to $881.2 + 3.4 * 175 = 1476.2$ kcals

Practice

Using the Colleges 2019 Complete dataset:

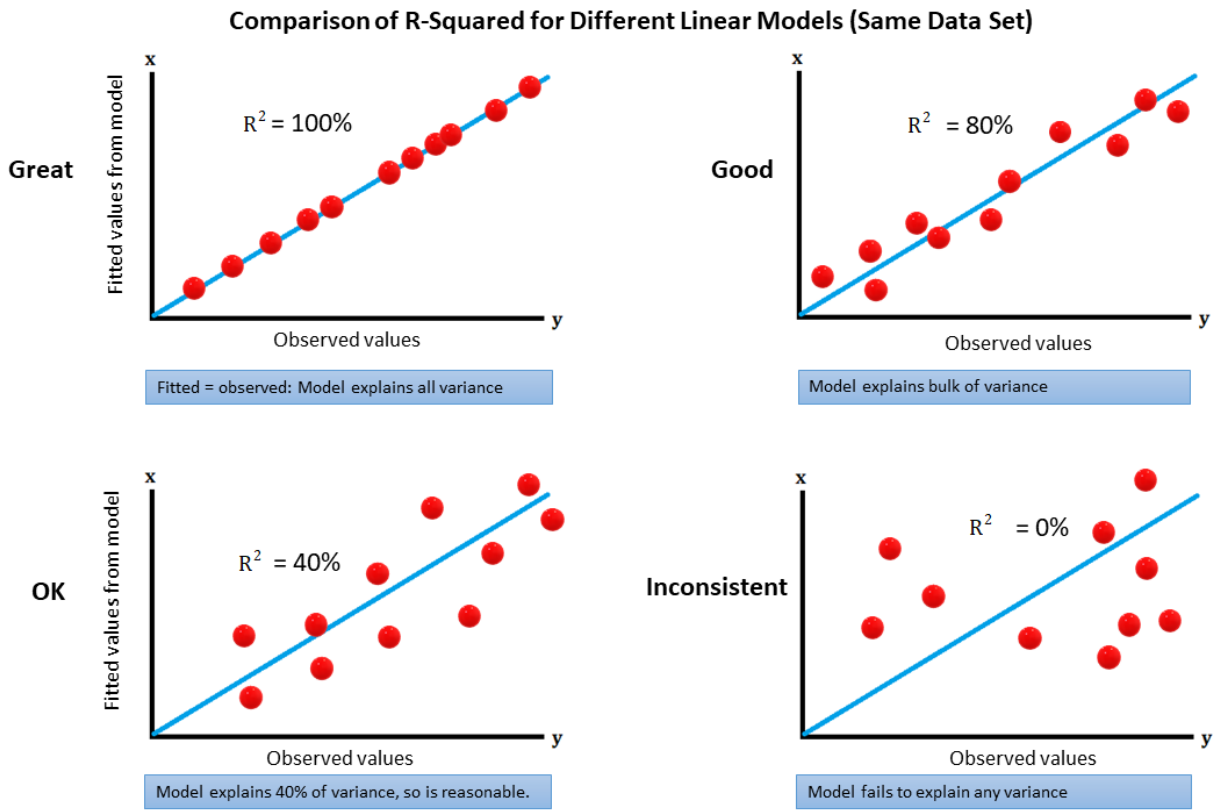
1) For the variables "Cost" (explanatory) and "Net_Tuition" (response), use StatKey to find the fitted simple linear regression model.
2) Interpret the slope and intercept of this model.
3) Predict the net tuition of a college with a full price cost of $50000

Practice (solution)

1) $\widehat{\text{Net Tuition}} = -949.6 + 0.395 * \text{Cost}$
2) The intercept is meaningless (since it's the expected net tuition of a college that costs nothing). The slope suggests that each \$1 increase in cost is accompanied by an expected increase in net tuition of \$0.395 (implying net tuition is generally around 40% of full cost for this set of colleges)
3) $-949.6 + 0.395 * 50000 = \$18800.40$

The coefficient of variation

The **coefficient of variation**, R^2, is a way of expressing how tightly a regression model resembles the data it was fit to:



R^2 is the proportion of variation in the outcome variable that can be explained by the regression line

Connections between regression and correlation

- For simple linear regression, R^2 is equal to the Pearson's correlation coefficient squared (ie: $R^2 = r^2$)

Connections between regression and correlation

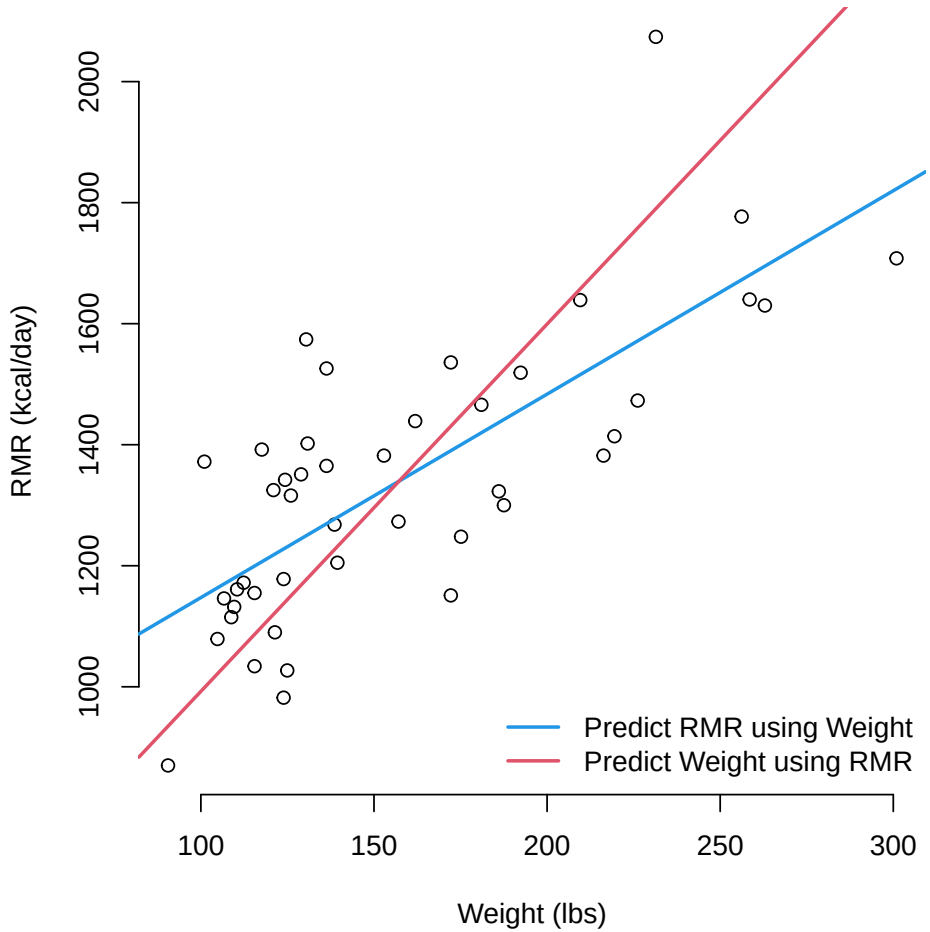
- For simple linear regression, R^2 is equal to the Pearson's correlation coefficient squared (ie: $R^2 = r^2$)
- Correlation is *symmetric*
  - The correlation between RMR and Weight is the *same* as the correlation between Weight and RMR

Connections between regression and correlation

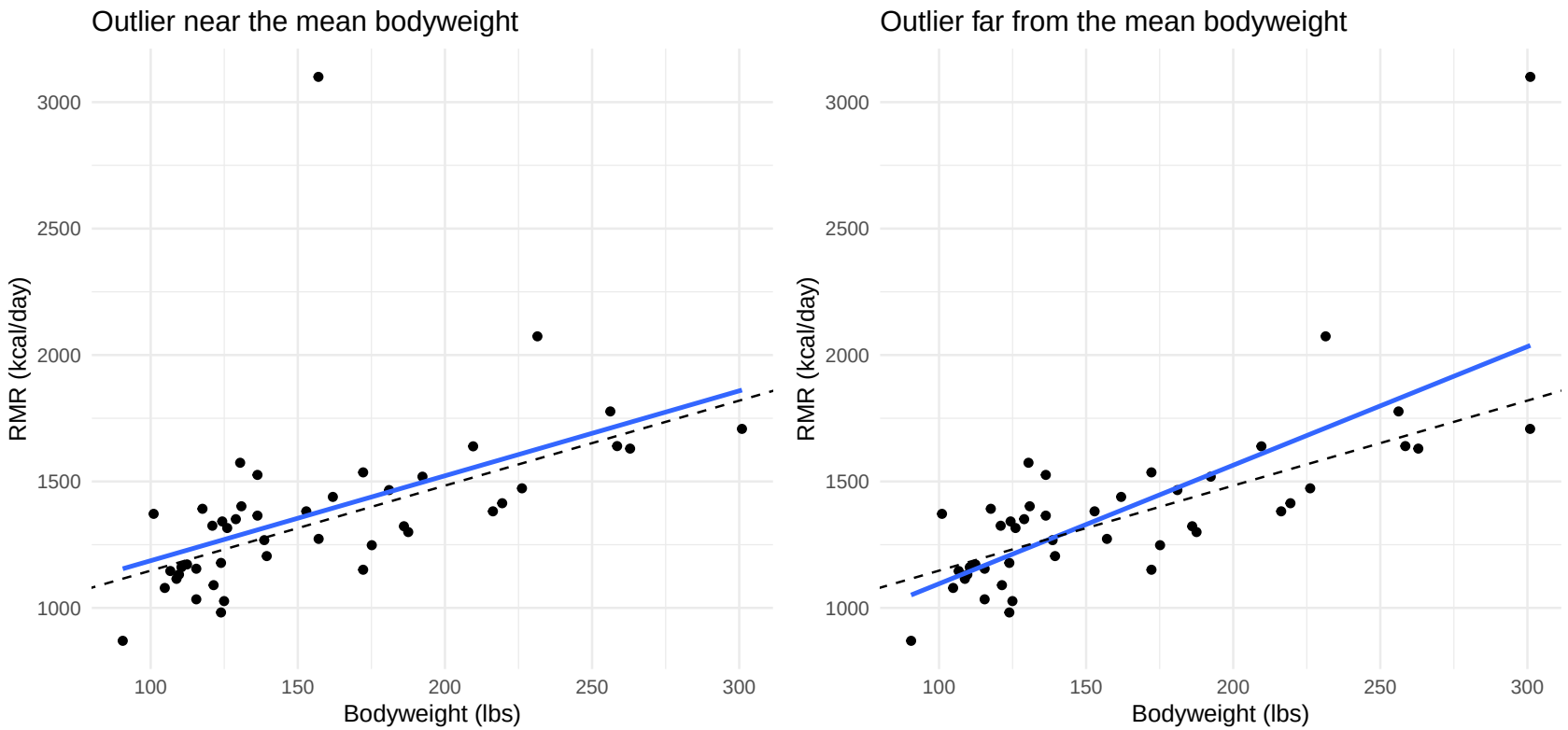
- For simple linear regression, R^2 is equal to the Pearson's correlation coefficient squared (ie: $R^2 = r^2$)
- Correlation is *symmetric*
  - The correlation between RMR and Weight is the *same* as the correlation between Weight and RMR
- Regression is *asymmetric*
  - The regression model that uses 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! But do not for correlation!

Mistake #1 - two regression lines

Shown below are the two combinations of explanatory and response variables for the RMR dataset:



Mistake #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

Mistake #3 - extrapolation

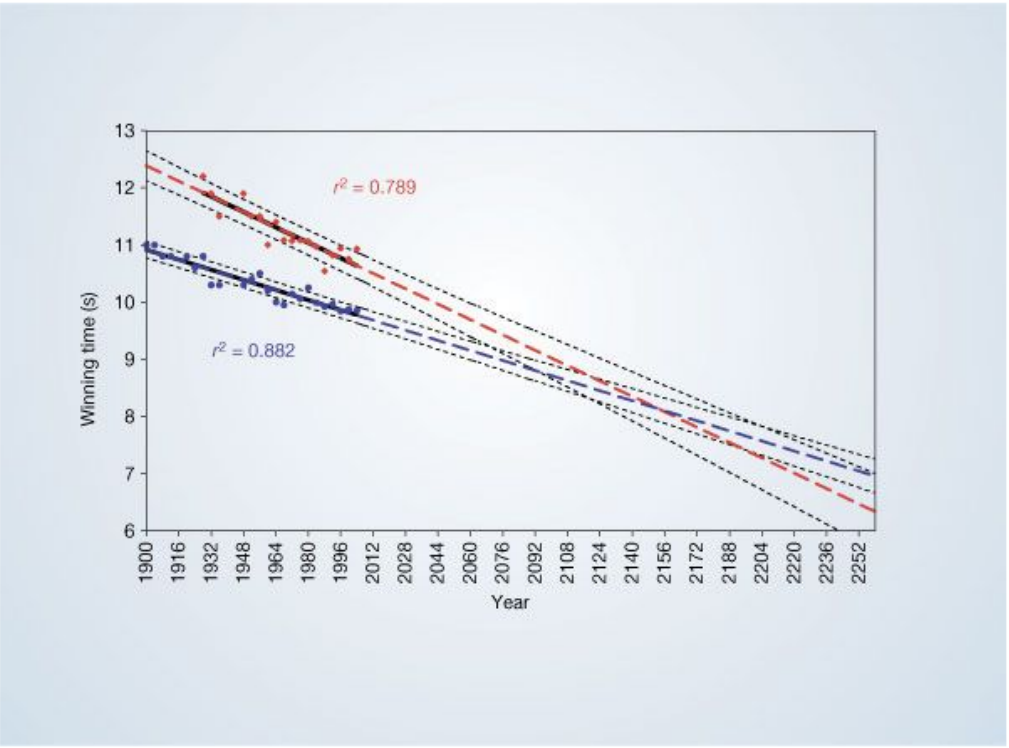
In 2004, an article was published in *Nature* titled "Momentous sprint at the 2156 Olympics". The authors plotted the winning times of the men's and women's 100m dash in every Olympics, fitting separate regression lines to each. They found that the lines will intersect at the 2156 Olympics, here are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

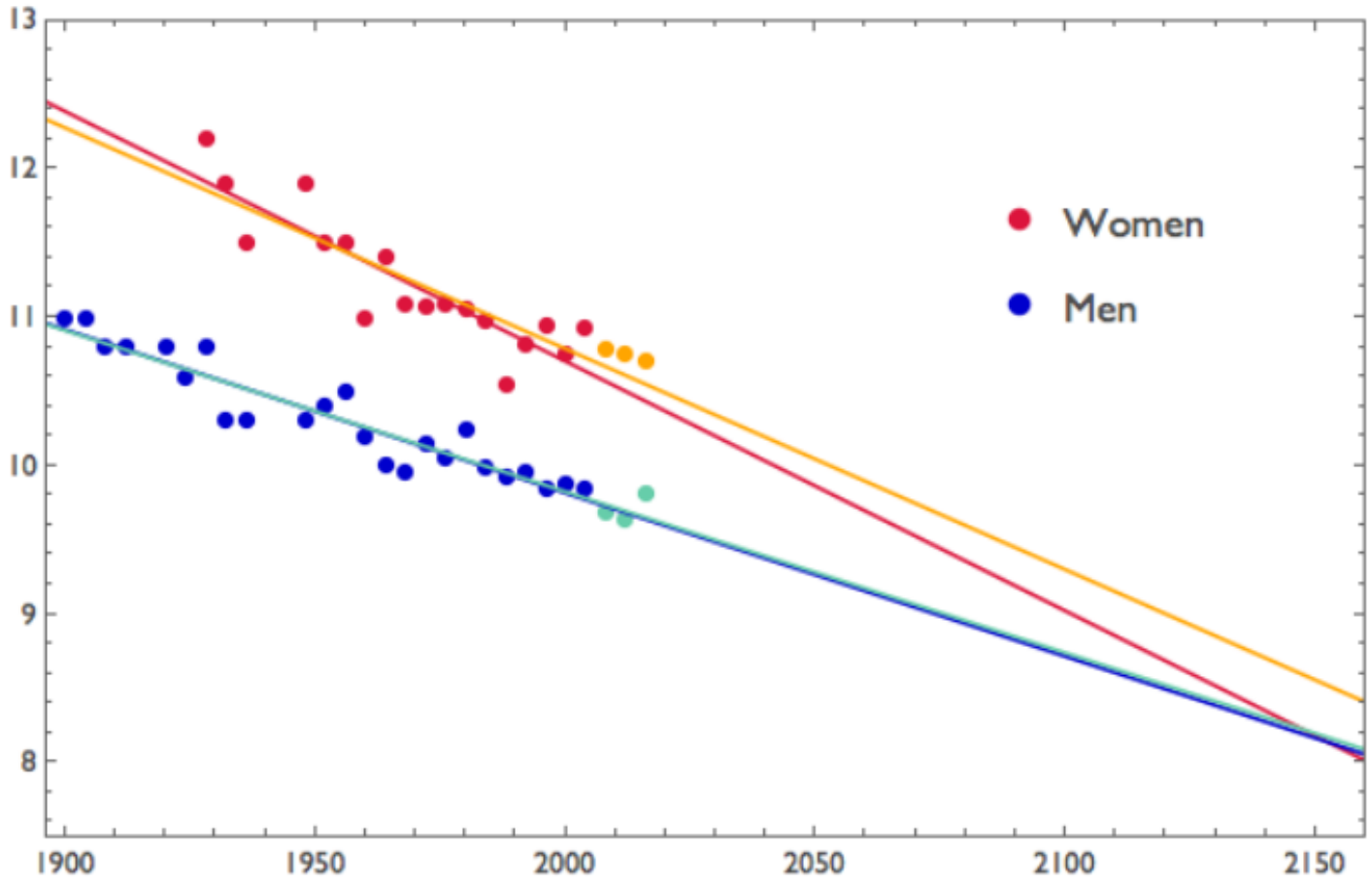
Here is a figure from the original publication in Nature:



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

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.



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

Extrapolation (summary)

Extrapolation describes the use of a model to make predictions in areas that are outside the range of the observed data

1) A linear trend in the observed range of the explanatory variable doesn't suggest that trend will persist outside of that range
2) The further you deviate from the average of the explanatory variable, the more sensitive predictions are to minor changes in the estimated slope and intercept

Summary

- Regression is an *asymmetric* method for describing the relationship between two quantitative variables
  - A fitted simple linear regression model is defined by its slope and intercept
- Outliers and influence
  - Outliers can be problematic if they are far from the average of *both variables* involved in the model
- Extrapolation
  - It is important not to predict beyond the observed range of your explanatory variable, your data tells you nothing about what is happening outside of its range!