Regression (part 2)

Ryan Miller

Introduction

- The last presentation introduced the **linear regression line** as summary measure used to describe the relationship between two quantitative variables
- This presentation will briefly cover a few misconceptions and common mistakes that often occur applying the regression to real data

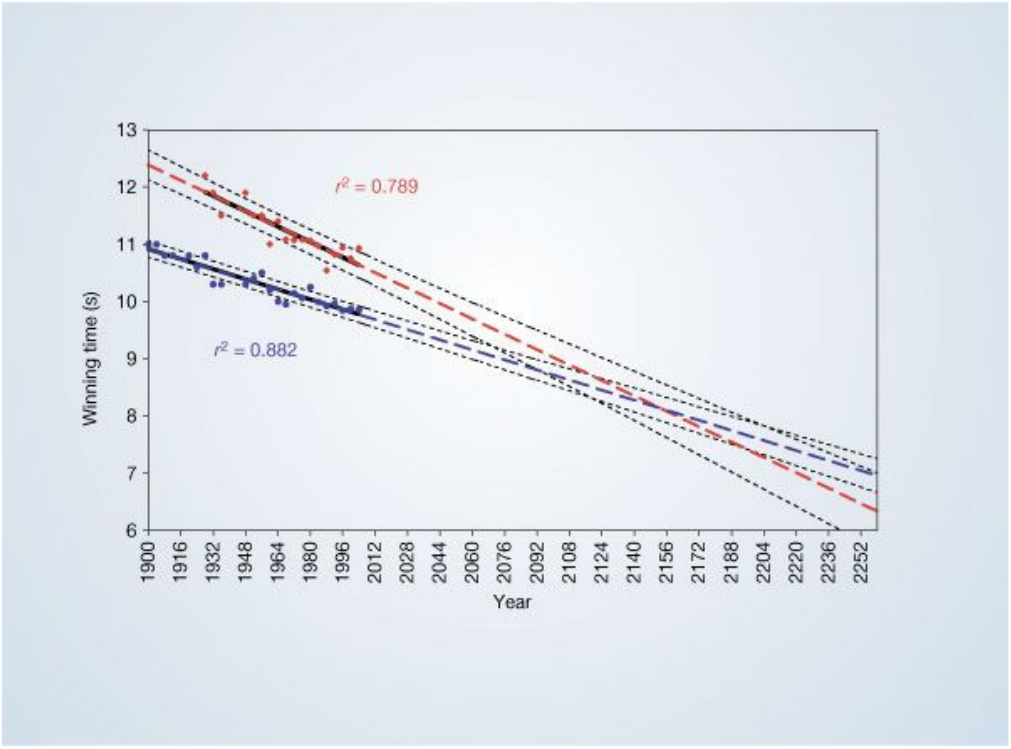
Mistake #1 - Extrapolation

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

Mistake #1 - Extrapolation

Here is a figure from the original publication in Nature:



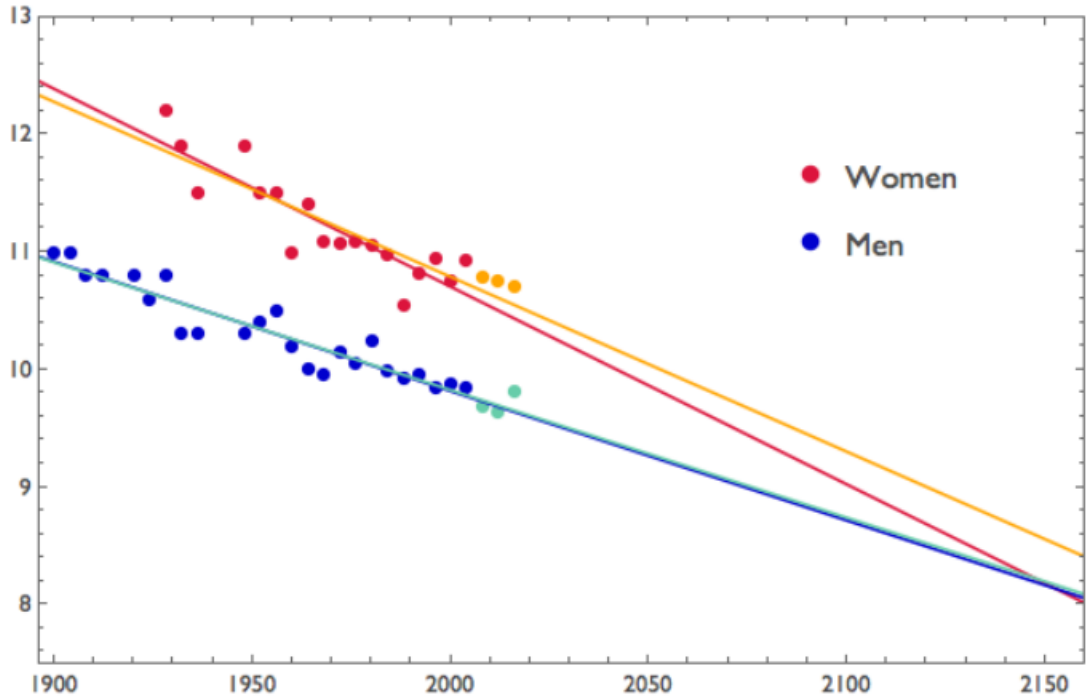
The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of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

Do you have any problems with the headlines on previous slide?

Advice

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!

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

Mistake #2 - Non-linear Relationships and Outliers

- Like the correlation coefficient, regression is only suitable for summarizing *linear relationships*
- It can also be heavily influenced by outliers

Mistake #3 - Inverting the Explanatory and Response Variables

- Recall that regression (unlike correlation) is **asymmetric**, so your choice of X and Y matters
- In the Burger King menu example, if we used protein to predict fat: $\widehat{\text{Fat}} = 8.4 + 0.91 * \text{Protein}$
  - A meal with 20g protein is predicted to have 26.6g of fat
- But if we used fat to predict protein: $\widehat{\text{Protein}} = 2.3 + 0.62 * \text{Fat}$
  - A meal with 26.6g of fat is predicted to have 18.8g of protein

Advice

- Before applying regression, carefully chose and explore your explanatory and response variables
- Do not apply the method if your exploration suggests a non-linear relationship, or if you see extreme outliers