Linear Regression - Predictions

Ryan Miller

Introduction

- A major strength of linear regression (relative to other types of models that we'll discuss next week) is that is a *statistical model*
  - This allows us to statistically assess our estimate's of the slope and intercept using confidence intervals and hypothesis tests
  - It also allows us to make *predictions* that incorporate uncertainty

Estimation vs. Prediction

- For a given value of x, simple linear regression stipulates $E(y) = \beta_0 + \beta_1 x$
  - Thus, our model provides an estimate of the *average y-value* when $x = c$ via $\hat{\beta}_0 + \hat{\beta}_1 * c$

Estimation vs. Prediction

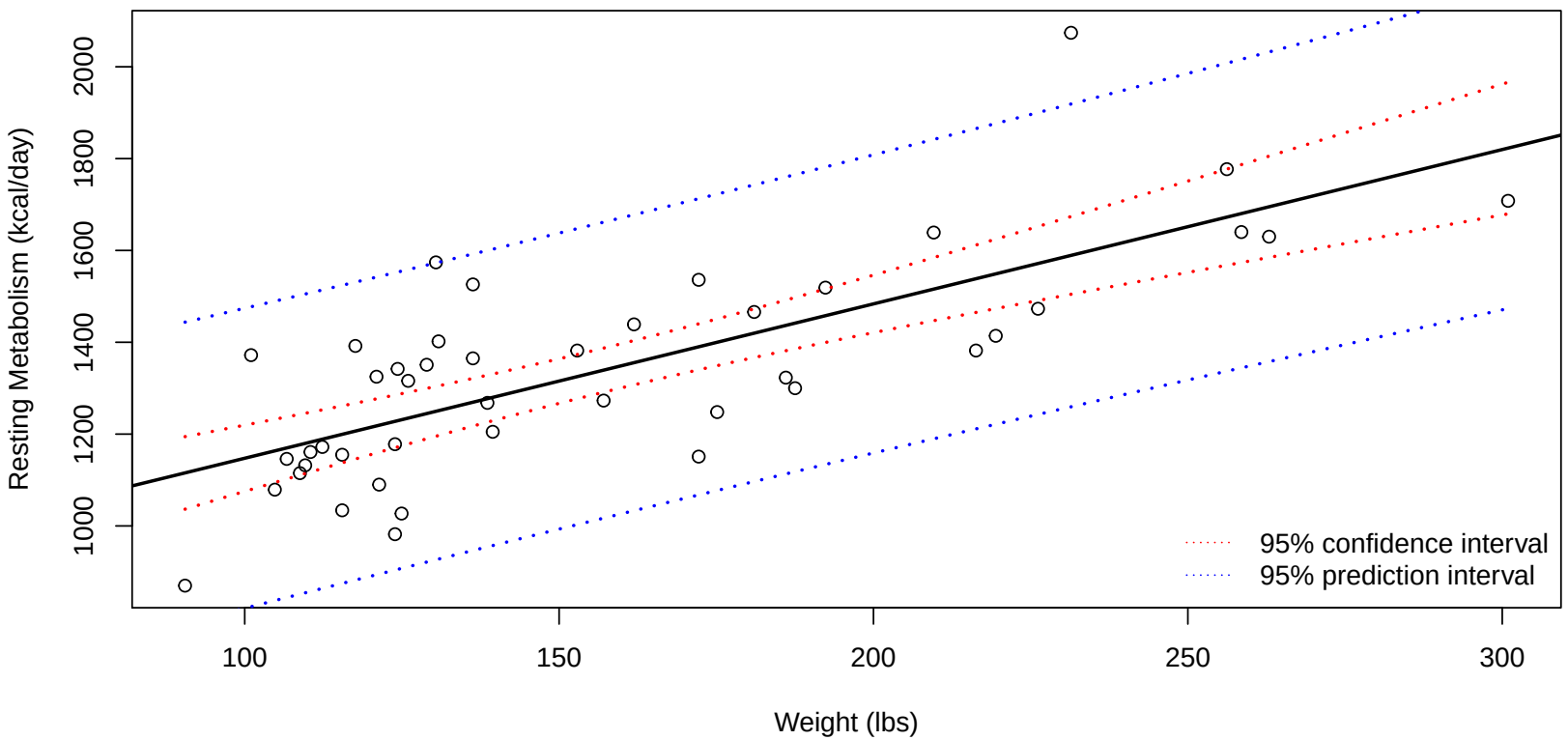
- For a given value of x, simple linear regression stipulates $E(y) = \beta_0 + \beta_1 x$
  - Thus, our model provides an estimate of the *average y-value* when $x = c$ via $\hat{\beta}_0 + \hat{\beta}_1 * c$
- Unfortunately, the average by itself doesn't tell us much about the variability in possible y-values that might be observed when when $x = c$

Estimation vs. Prediction

- Thinking about our home sales example, we might want to know what types of sale prices we'd expect for a home with an assessed value of $200,000
  - Our model suggests these homes should have *average* sale prices at (or slightly above) $200,000
  - But is a sale price of $300,000 something we might expect? Or what about $150,000?
- Could a *confidence interval* help us?

Estimation vs. Prediction

- A confidence interval will express the statistical variability of the *average y-value*, but sometimes we're more interested in variability of *individual y-values*
  - This can be assessed using a **prediction interval**



A Simple Analogy

- **Prediction interval**: applying the 68-95-99 to the *cases in a population*
- **Confidence interval**: applying the 68-95-99 to the *sampling distribution*

Some Theoretical Details

- Consider the *random component* of the simple linear regression model, $\epsilon \sim N(0, \sigma^2)$
  - The residuals in our fitted model can be seen as realizations of these random errors

Some Theoretical Details

- Consider the *random component* of the simple linear regression model, $\epsilon \sim N(0, \sigma^2)$
  - The residuals in our fitted model can be seen as realizations of these random errors
- The *residual sum or squares* (SSE) summarizes the *total variation* in these errors
  - Dividing SSE by its *degrees of freedom*, $n - 2$, yields an unbiased estimator of σ^2
  - The resulting estimate, s^2, expresses the total variability (in individual errors) that exists around the regression line

Some Theoretical Details

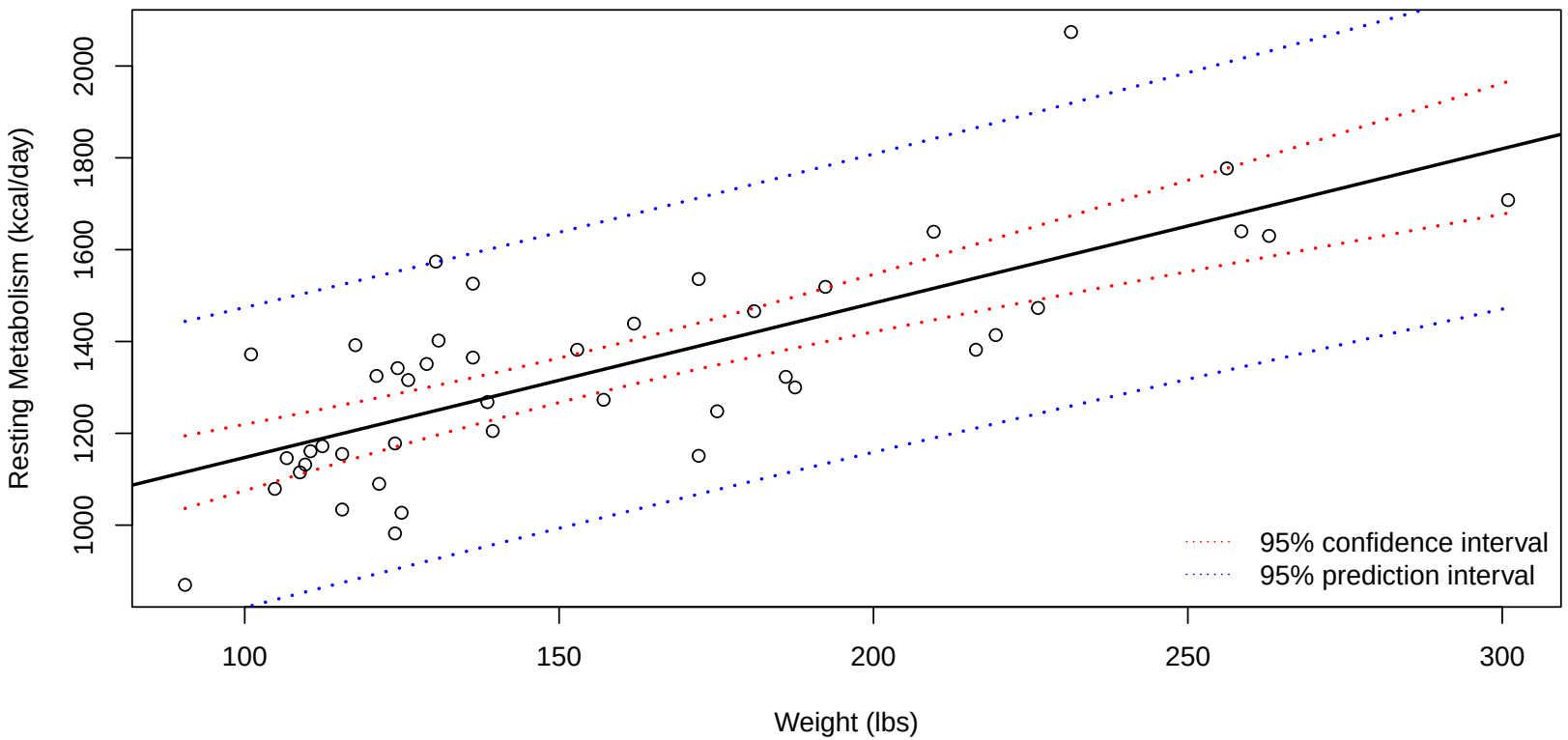
- Based upon the code below, we can estimate σ as $s = 157.9$ for the RMR dataset
  - Thus, the average distance between an observed resting metabolic rate and our line is around 158 kcal

```
rmr <- read.csv("https://remiller1450.github.io/data/RMR.csv")
m <- lm(rate_kcal ~ weight_lbs, data = rmr)
s2 <- sum(m$residuals^2)/m$df.residual
s <- sqrt(s2)
s
```

```
## [1] 157.9052
```

s^2 and Prediction Intervals

- Notice the relationship between $s = 158$ and the 95% prediction interval



More Theoretical Details

- You might notice that confidence intervals and prediction intervals tend to grow wider, or "fan out", near the edges of the regression line. This is due to two factors:
  - The slope and intercept each have their own statistical uncertainty
  - The fitted regression line must always pass through the point $\{\bar{x}, \bar{y}\}$

More Theoretical Details

- You might notice that confidence intervals and prediction intervals tend to grow wider, or "fan out", near the edges of the regression line. This is due to two factors:
  - The slope and intercept each have their own statistical uncertainty
  - The fitted regression line must always pass through the point $\{\bar{x}, \bar{y}\}$
- Conceptually, if you image a bunch of different samples you'd expect $\{\bar{x}, \bar{y}\}$ to be fairly similar in each, at least compared to the variability in other data-points (since the variability of a sample mean is Std Dev$/\sqrt{n}$)
  - We won't cover the mathematical details, but you can find them in Ch 3.9 of our textbook (A Second Course in Statistics)

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.

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.

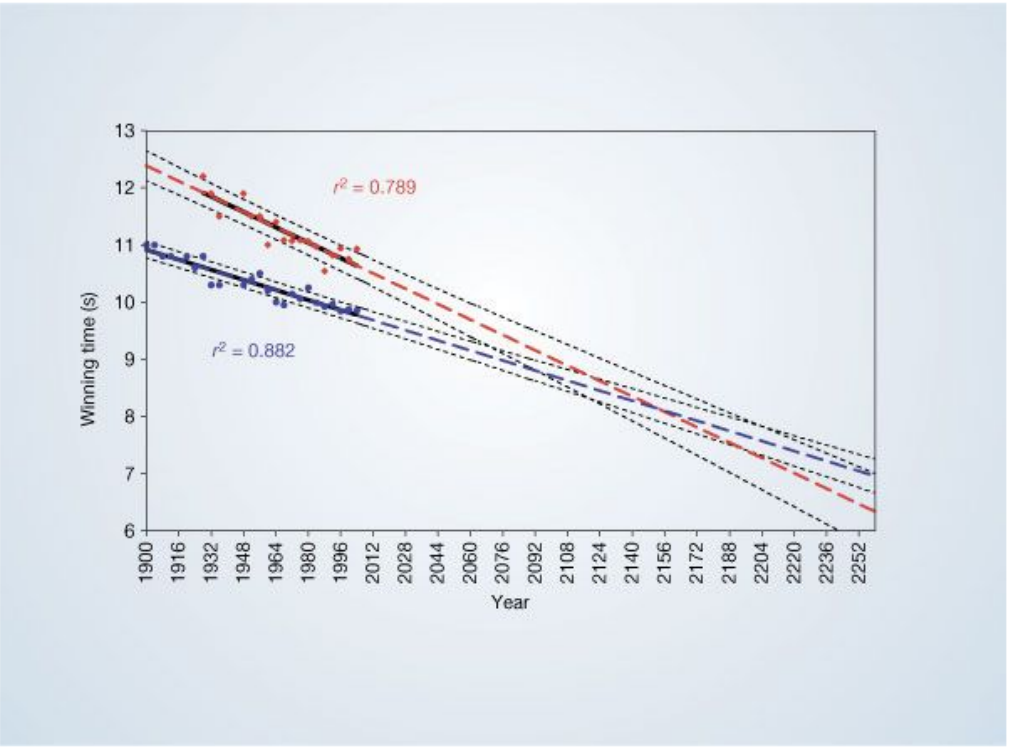
They found that the lines will intersect at the 2156 Olympics, here are a few media headlines:

- "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 the figure from the original publication in Nature:

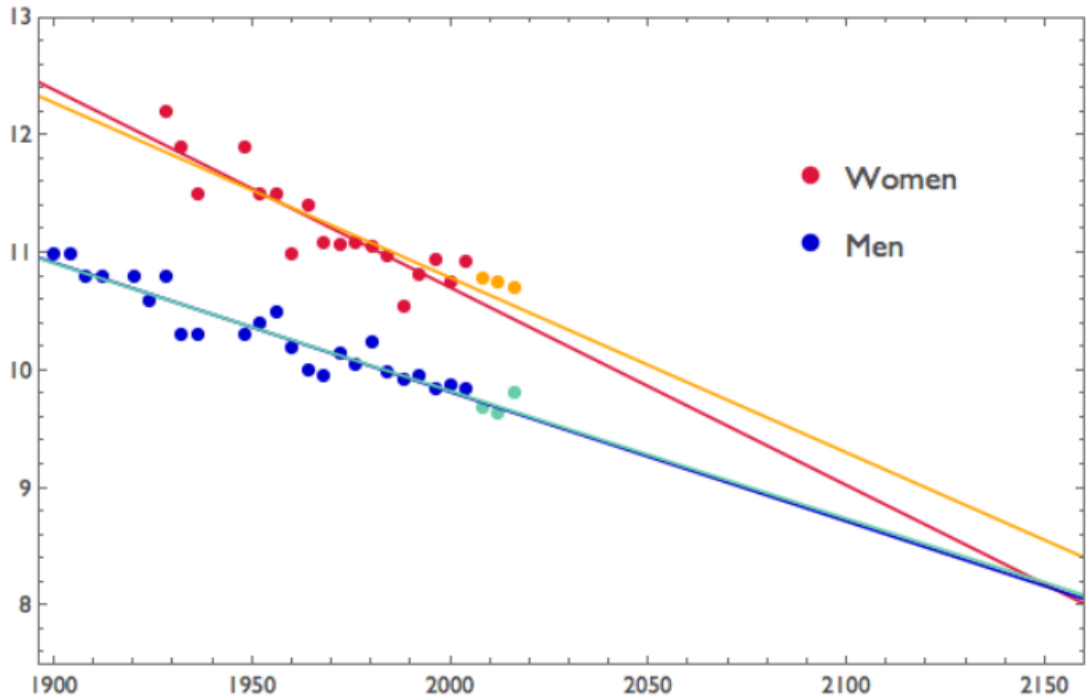


The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of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

Extrapolation

A few more Olympics have happened since 2004, so new data-points can be added:

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

Closing Remarks

- The statistical aspects of linear regression make it a very attractive model
  - As we'll soon see, this is particularly useful for models with multiple predictors, as it allows us to make "what-if" predictions that account for uncertainty using our model