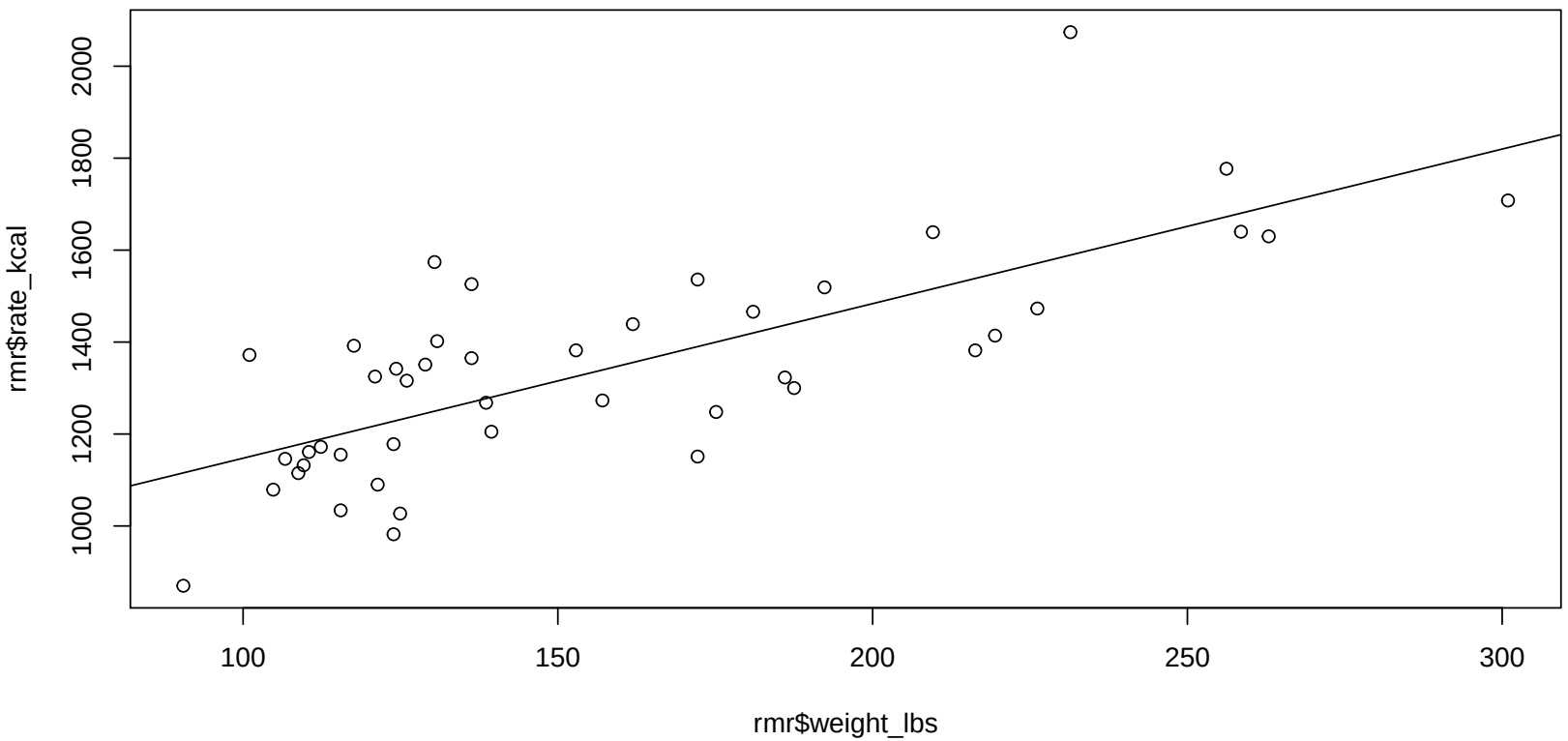
Multiple Linear Regression - Outliers and Influence

Ryan Miller

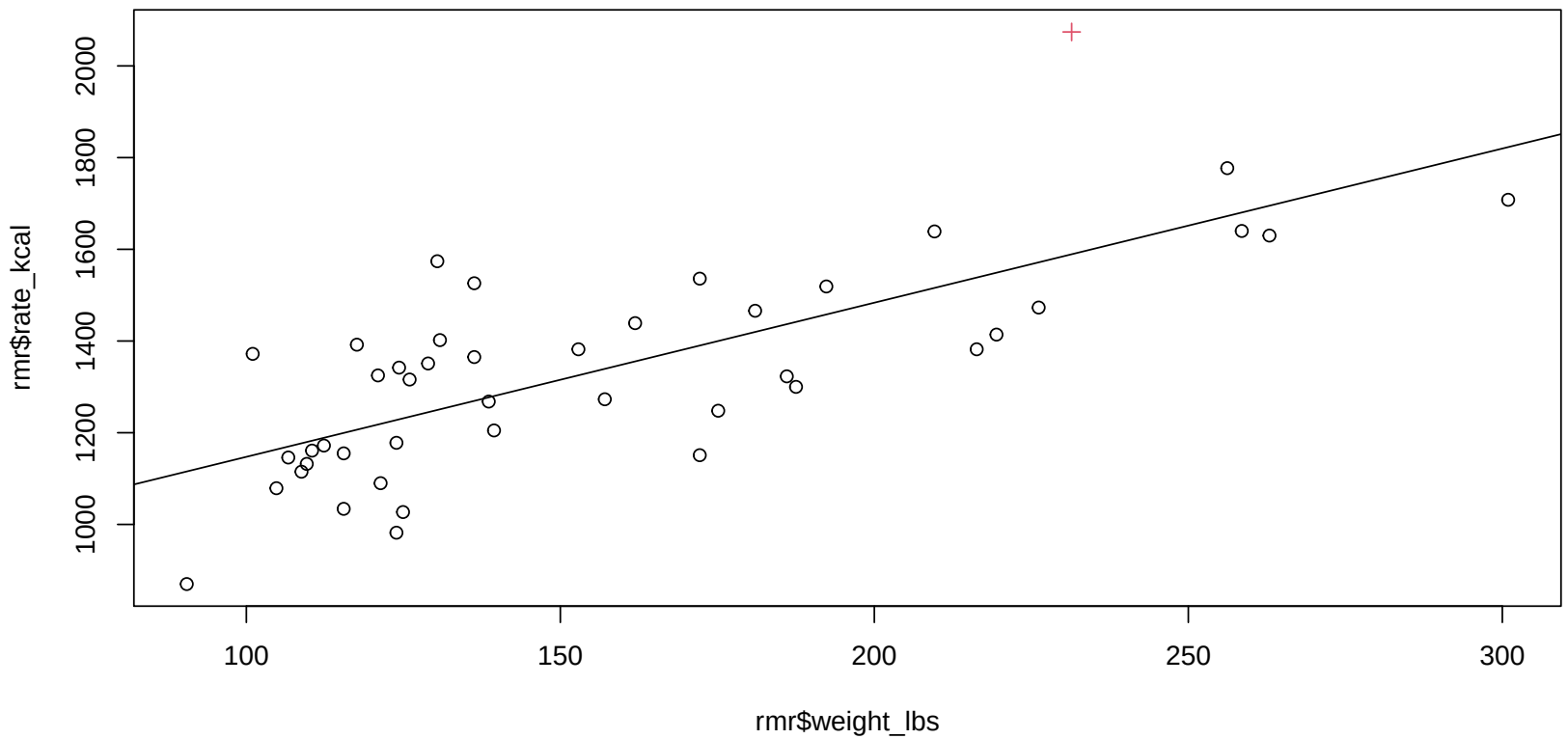
Introduction

- In an ideal world, how much should each data-point impact the slope and intercept of a model?
  - In this model, which data-point do you think is the *most influential*?



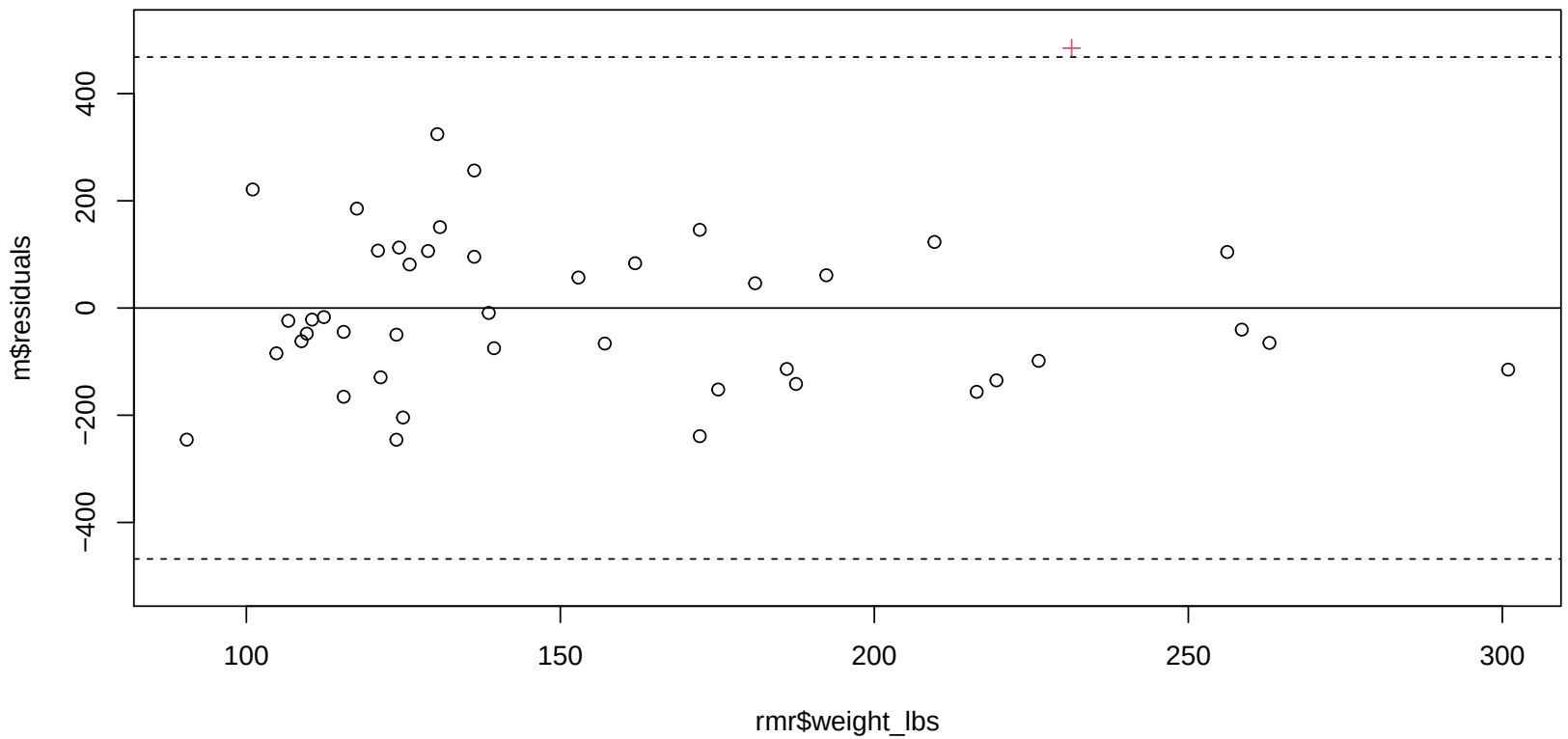
Identifying Unusual Data-points

The data-point marked by a red "+" deviates from the trend seen in the other women in this dataset



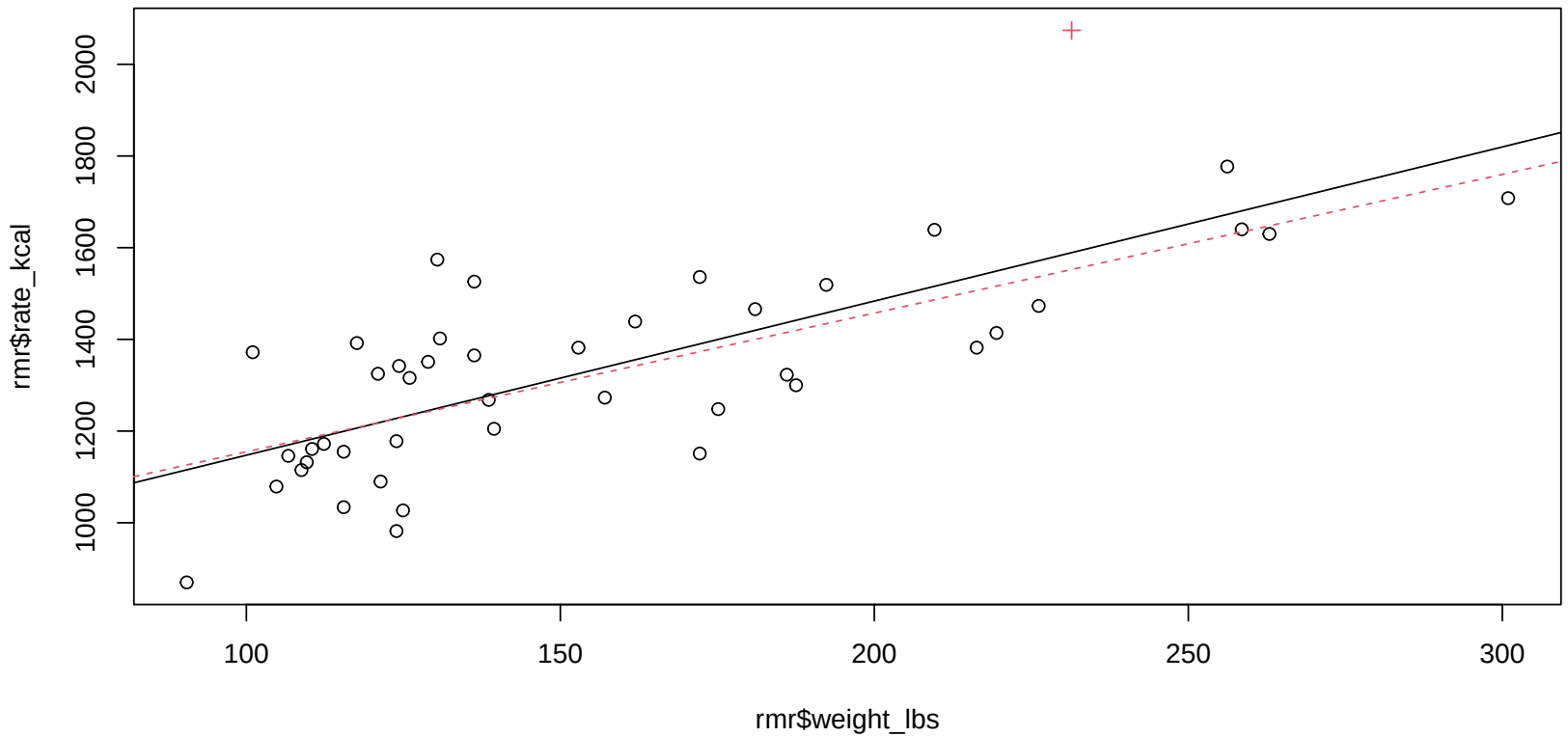
Is it an Outlier?

A common rule of thumb labels a data-point an outlier if it's residual is more than 3 standard deviations away from zero



Influence

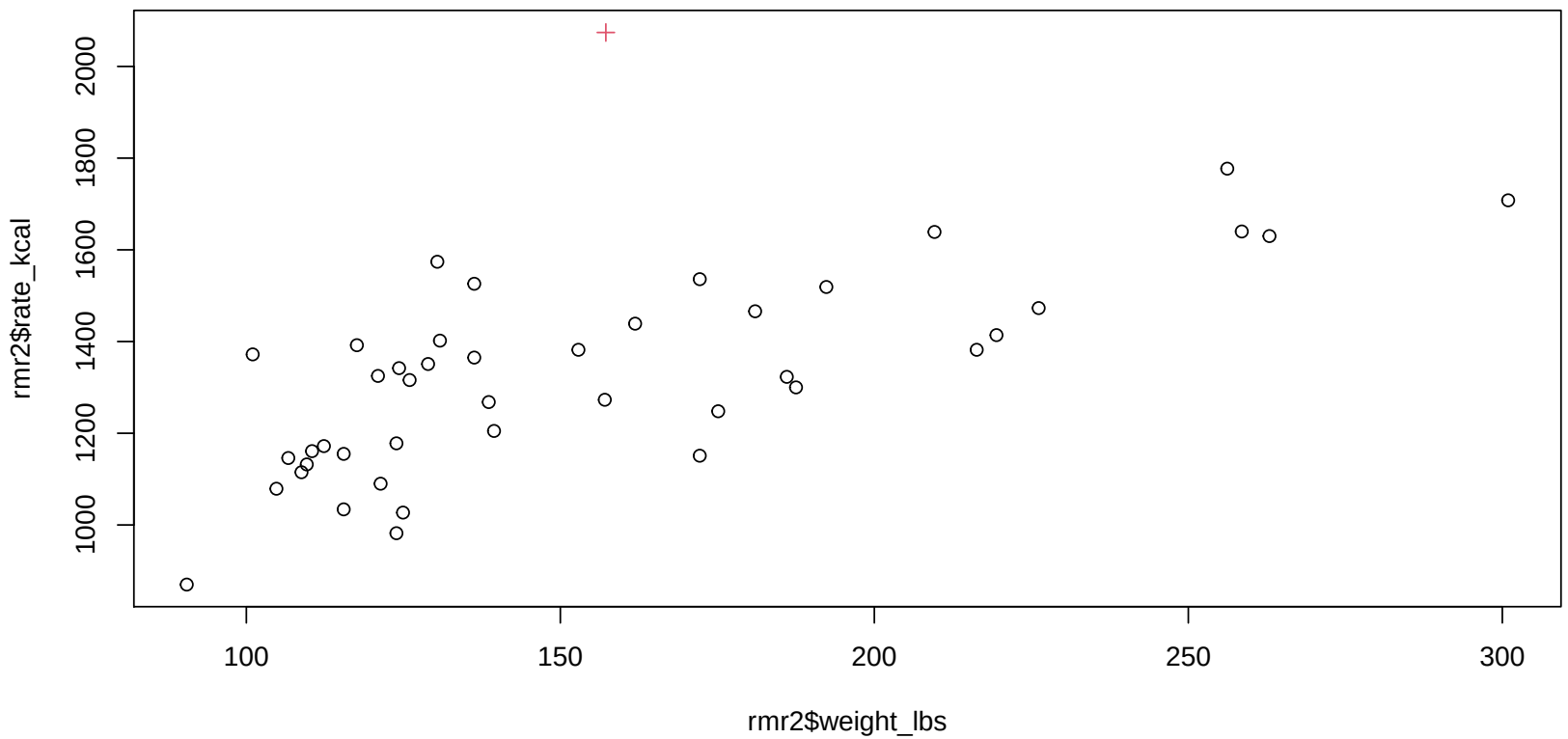
What happens if we remove this data-point?



The slope decreases from 3.36 to 3.03, which might not seem like much but it's an 11% swing attributable to a single data-point

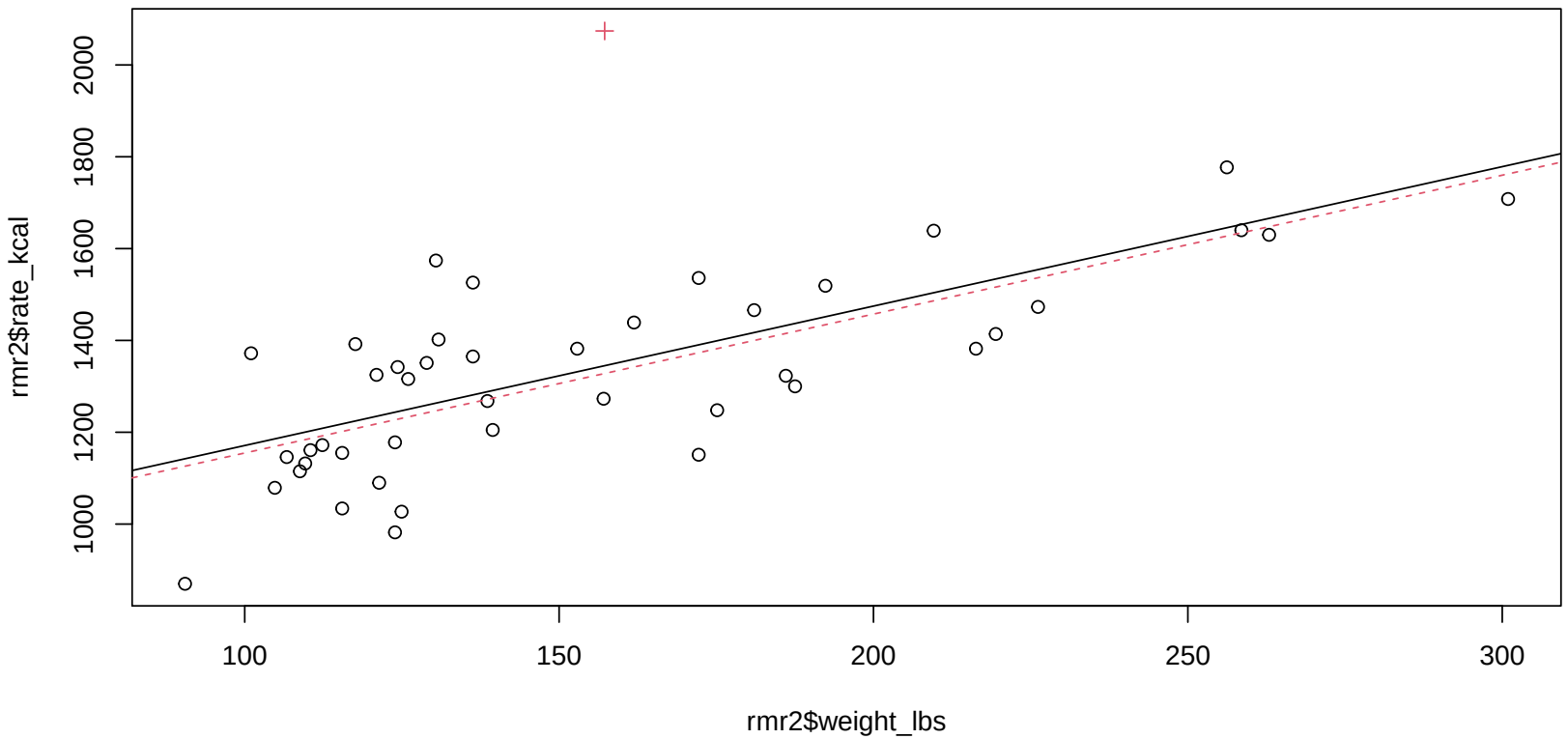
Outliers

- Now suppose we artificially move this data-point's x-value to 155
  - Will this make the point more or less of an outlier?
  - How influential do you think this data-point will be now?



Leverage

The outlying data-point now has *very little influence* on the model's slope (it's 3.026 with all of the data, and 3.025 if the outlier is excluded)



Why isn't such a large outlier influencing the model's slope?

Leverage

- In order to be **influential**, a data-point must be both an **outlier** (unusual y-value) and **high leverage** (away from the mean of x)
  - In the original RMR dataset, the unusual data-point was located at ($x = 231$, $y = 2074$)
  - This is about 1.5 standard deviations beyond the average weight in the dataset (157 lbs)

Leverage

- In order to be **influential**, a data-point must be both an **outlier** (unusual y-value) and **high leverage** (away from the mean of x)
  - In the original RMR dataset, the unusual data-point was located at ($x = 231$, $y = 2074$)
  - This is about 1.5 standard deviations beyond the average weight in the dataset (157 lbs)
- The exact mathematical definition of **leverage** is more technical than we'll get into
  - Conceptually, it can be thought of as the impact of a data-point on its own fitted value
  - Practically speaking, the further away a data-point is from the average predictor, the higher it's leverage tends to be

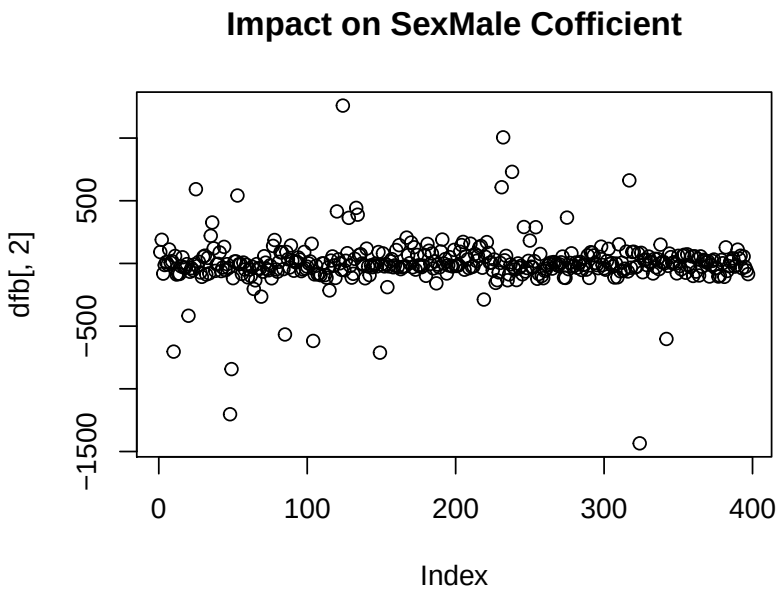
Leverage vs. Outlier

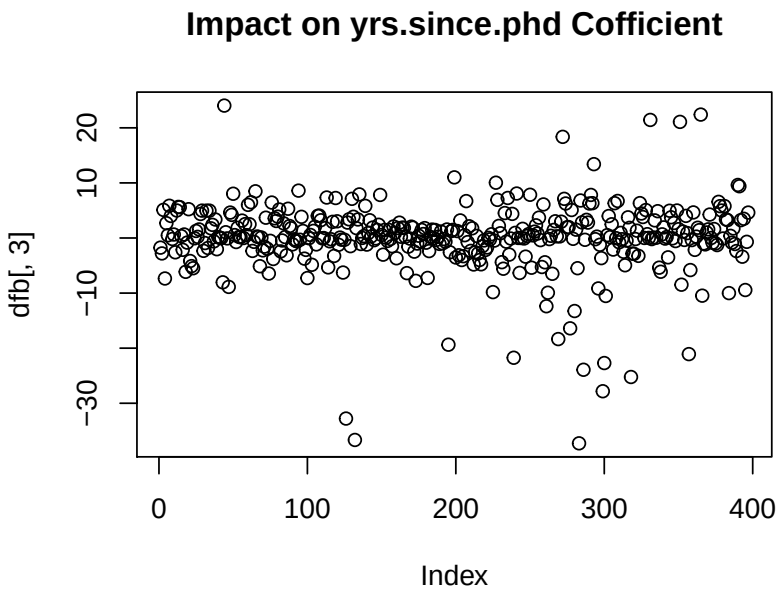
In summary:

- Data-points with small residuals tend to have little impact on model fit, regardless of their leverage
- Data-points with large residuals (outliers) but low leverage also tend to have little impact on model fit
- It's only data-points with large residuals *and* high leverage that substantially impact the model

Influence and Multiple Coefficients

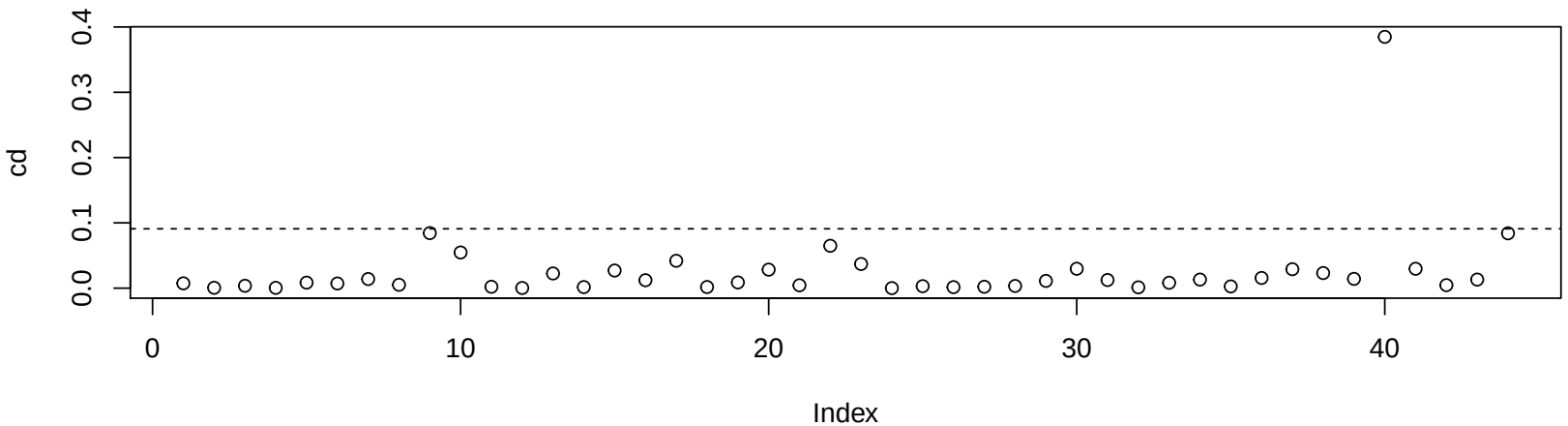
- When a model includes many predictors, data-points with large residuals might be influential on one component of the model, but not others
  - The `dfbeta()` function will calculate how much the removal of an individual data-point will change each coefficient in the model





Cook's Distance

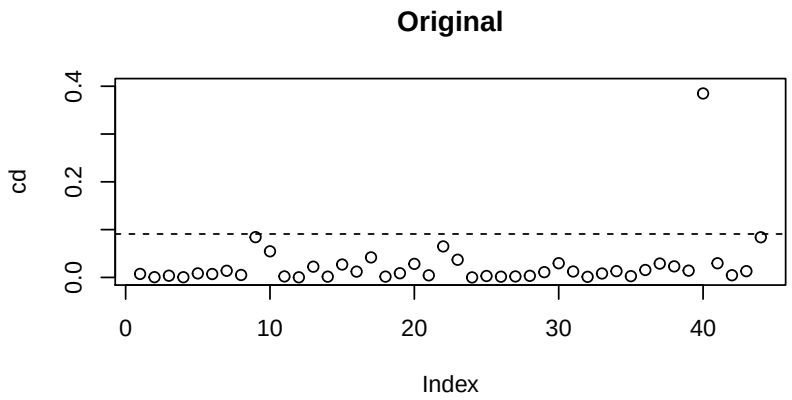
- Cook's distance is an attempt to provide a single, standardized measurement of the influence of an individual data-point on an *entire model*
- It can be viewed as the *sum of all changes* in model's fitted values when a data-point is deleted
  - A Cook's distance larger than $4/n$ is deemed influential

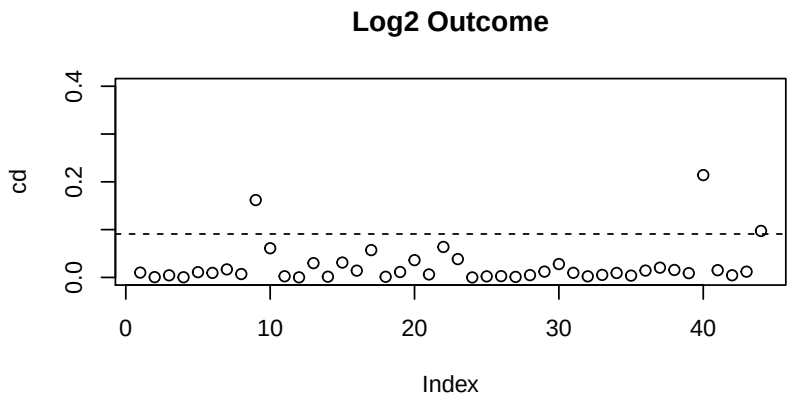


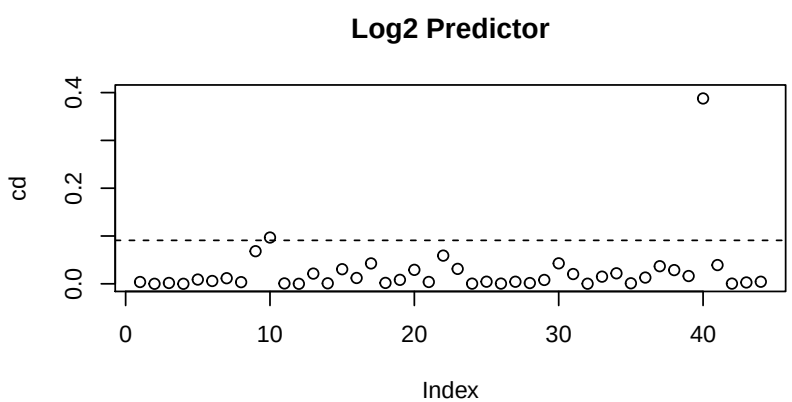
Remediation

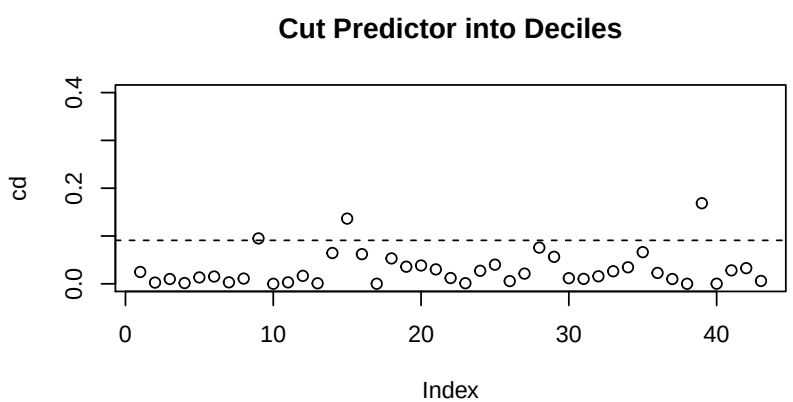
- Upon identifying an influential observation, the action you should take depends upon the goals of your model
  - If you deem the observation to not belong to the population you're studying, you might exclude it entirely
- Otherwise, you might consider transforming your data to reduce the data-point's influence
  - Transforming the outcome variable
  - Transforming the predictor with the largest DFBETA

Remediation









Closing Remarks

- Influence diagnostics are an advanced step in a regression analysis that can help you find a model that is broadly generalizable and not disproportionately based upon a small set of data-points
- There is a high degree of subjectivity when deciding how to handle influential observations
  - At minimum, you should check for anything that really stands out and report on it as a limitation of your model (assuming you make no changes)
  - At maximum, you might completely redesign the predictors you are using to ensure the observations in your data are contributing more equally to the model