SOME INTUITION ON WHY COLLINEARITY CAN BE "BAD", AND WHEN IT SHOULD BE A CONCERN

Dan Putler, Chief Data Scientist, Alteryx

alteryx | The Thrill of Solving

- Why this talk?
- Some intuition on how to think about the effects of (multi)collinearity
- A set of Monte Carlo experiments that examine the effects of differing levels of predictor collinearity and the number of available observations
- Some “rules of thumb” for determining how concerned you should be about collinear predictors for a particular project

WHY THIS TALK?

- Based on personal observation, I find that a fair number of people who develop predictive models know that having predictors that are collinear is “bad”, but they do not have an intuitive understanding why it is potentially bad, or how it works
- Why does this matter?
  - It can lead to unnecessary efforts associated with initial data preparation
  - It can lead to poor predictor filtering practices
  - It often leads to a misunderstanding about the potential bias associated with a model’s predictions

HOW TO THINK ABOUT COLLINEARITY

- A concrete example: Two continuous predictors and a continuous target
  - In this case the Pearson correlation coefficient between the two predictors is an appropriate measure of predictor collinearity
- The value of the Pearson correlation coefficient between the two predictors is an indication of the relative overlap of the information that the two variables provide with respect to predicting the target
  - As the value of the correlation coefficient increases the overlap in the information provided by the two predictors increases, while the amount of information specific to each predictor decreases
  - This overlap in the information between the two predictors means that the amount of information specific to each of these two predictors that each row of the data provides is reduced

THE EFFECTS OF COLLINEARITY

- The reduction in the effective information content of a row of data means that the precision with which we are able to determine the effect of a predictor variable on the target is reduced
  - In the context of a traditional linear regression model, what this means is that the uncertainty around the value of a regression coefficient increases with the level of collinearity
- In a linear regression model, the uncertainty about a coefficient estimate is captured by its standard error
  - The systematic increase in value of the standard error as the level of collinearity increases results in a *bias* whereby traditional tests of the significance of a coefficient are overly conservative

THE EFFECTS OF COLLINEARITY

- Does this bias matter? It depends on whether our objective of an analysis is inference or prediction
  - In the case of inference, it is potentially huge
  - For prediction, it can matter, but ***there is no systematic bias***
- Since the underlying impact of collinearity is to reduce the information content of each observation of data, we can make up for it by having more observations

THE MONTE CARLO SETUP

The equation used to generate the experimental data

$$y = 0.5 - x1 + x2 + 1.5(x3) + e,$$

where e is a random deviate that follows the standard normal distribution, and $x1$, $x2$, and $x3$ are taken from a multivariate normal distribution.

The estimated linear regression models are of the form

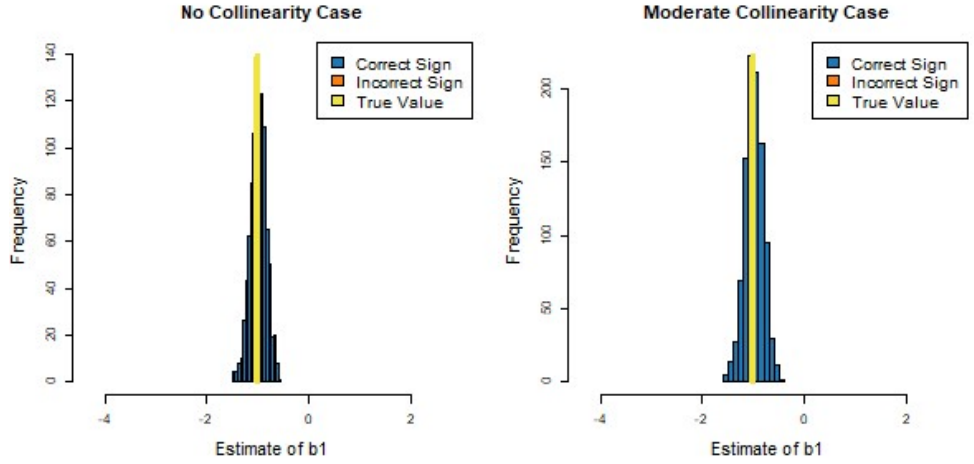
$$y = b0 - b1(x1) + b2(x2) + b3(x3) + error$$

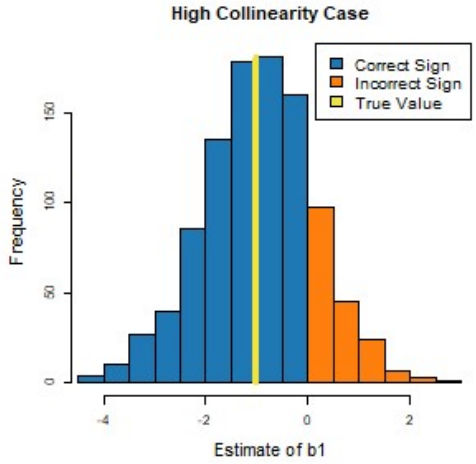
The analysis focuses on the estimated values of $b1$

THE MONTE CARLO SETUP

- Collinearity treatments
  - **None**: All three variables have a correlation of zero
  - **Moderate**: The correlation between *x1* and *x2* is 0.5, and both *x1* and *x2* have a correlation of 0.15 with *x3*
  - **Highly**: The correlation between *x1* and *x2* is 0.99, and both *x1* and *x2* have a correlation of 0.3 with *x3*
- The number of records within each data set is varied from 50 to 2500
- Each specific experiment makes use of 1000 replicates
- This setup allows the findings to be presented as a series of histograms concerning the estimated values of *b1* across the replicates

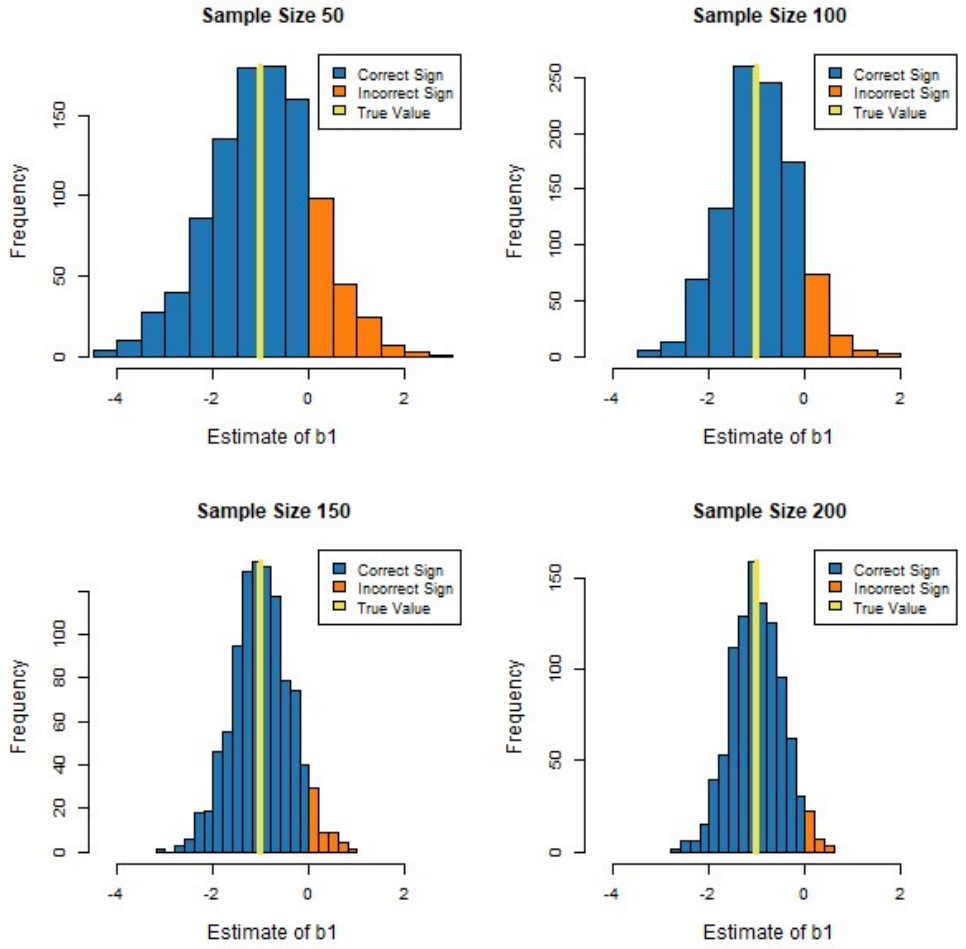
COLLINEARITY WITH 50 RECORDS





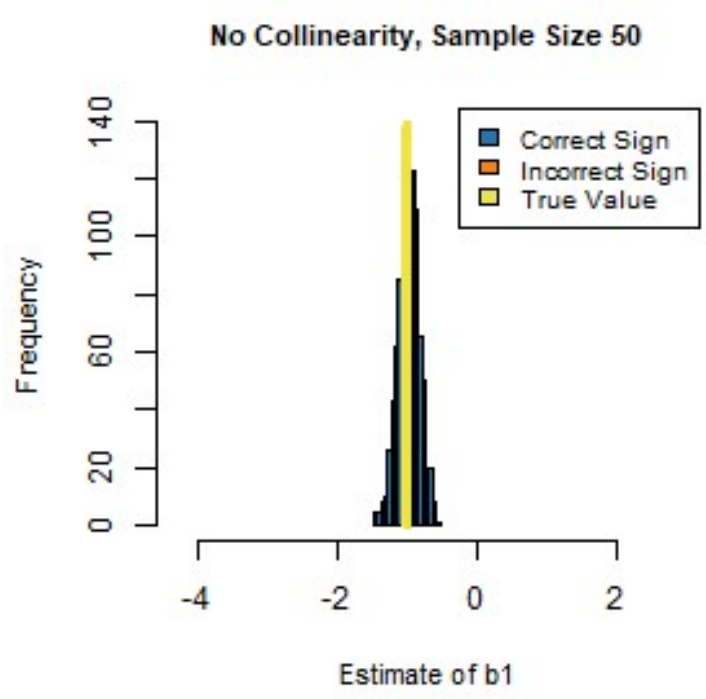
alteryx | The Thrill of Solving

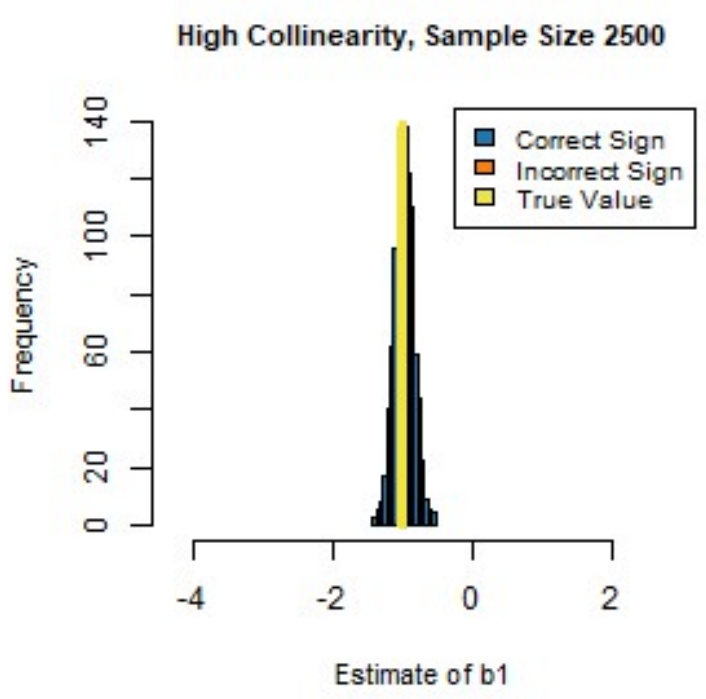
HIGH CASE / RECORDS RELATIONSHIP



alteryx | The Thrill of Solving

COMPENSATION VIA RECORDS





alteryx | The Thrill of Solving

SOME "RULES OF THUMB"

- The effect of moderate levels of collinearity on the precision of estimates is fairly minimal, but the level of precision degrades substantially as the level of collinearity becomes very high
  - There is a need to be concerned about high levels of collinearity, but not about moderate or low levels of collinearity, when the number of observations available is fairly small
- Even in the presence of high collinearity, acceptable levels of precision in determining the effect of predictors on the target can be achieved if there is a sufficient number of rows of data available
  - If there is a lot of data (several thousand records of data or more), even high levels of collinearity do not pose a major problem

Thank You!

https://github.com/dputler/presentations

