# Advanced Regression – Project Proposal


## Jared Hansen

## Monday, March 23, 2020
## TITLE: "From-scratch implementation of the Least Angle Regression algorithm"

1. **The goal. What problem am I trying to solve?**

   The goal of this project is to develop a deep understanding of the Least Angle Regression (LAR hereafter) algorithm. To do this, it is my plan to develop code that implements the LAR from scratch in Python (and hopefully C++), and demonstrate its use in variable selection for LASSO regression. I will ideally get the same (or similar) results as a standard implementation of LAR-selected LASSO in software like SAS, R, or Python.

2. **The data. What data will I use to address this problem?**

   I will use two different data sets. The first data set is collected from the dating app Tinder. The goal is to quantify the genuineness of an individual's self-presentation in their profile, as measured on a continuous scale from -4.0 to 4.0 (higher score meaning more genuine). There are 498 observations and 20 predictor variables, all of which are numeric. Predictors quantify things like self-esteem, importance of fitness, and education level on a scale similar to that used for genuineness.

   The other data set I'm going to use is the residential building data set from the UCI repository. This data set contains 373 observations, 81 numeric economic predictors, and two response variables (construction cost and sales price). Since there are two response variables, I'll be able to use this as though it's two separate data sets if needed. https://archive.ics.uci.edu/ml/datasets/Residential+Building+Data+Set

3. **The methods. What approaches do I plan to use to solve the problem? What will I compare to for a baseline?**

   I will first code this up in Python. After arriving at a final Python implementation, I will compare my from-scratch results to ready-to-use versions of the LAR-selected LASSO in SAS and potentially Python and R. I'll visually compare my coefficient progression plot to that of ready-made methods and use MSPR on a held-out test data set to quantify how well my implementation does relative to existing packages. I'm sure my implementation will be slower, so I'll time the ready-to-use versions and see how much faster they are.

   Depending on how long all of this takes me, it's also my goal to come up with a from-scratch implementation of LAR-selected LASSO in C++. If I'm able to do this, I'll similarly compare my results (e.g. compare coefficient progression plot and test set MSPR) to the ready-made implementations, as well as seeing how fast I can compute the results.