**Udacity GitHub Augmentation**

This 2-pager summarily describes what was done, why it was done, and what I found. What was found takes the form of 5+ talking points.

**I. What Was Done**

Udacity profiles were scraped for GitHub profile data. GitHub profiles, in turn, were scraped for GitHub profile data. GitHub profile data collected included the number of repositories created, the number of code commits within the last year, and the number of followers the individual has on GitHub. These factors were interacted and tested for linear and nonlinear relations to employment. These factors were then treated as left hand variables to proxy productivity and they were checked to see whether alternative education correlates positively.

**II. Why it Was Done**

GitHub acts as a portfolio for professional programmers. Portfolios can be interpreted both as a credential and as an output of labor. Portfolios-as-credentials are expected to explain labor outcomes including salary and employment rate. Portfolios-as-product should be produced more frequently and with higher quality when produced by an individual of comparatively high skill.

In the regression of employment on alternative education, including GitHub data on the right hand reduces omitted variable bias in alternative education. In the regression of GitHub data on alternative education, observation is obtained on whether alternative education significantly changes productivity or merely acts as a signal.

**III. What I Found**

1. Summary and exploratory [TODO: review.]
   1. 48.6% of Udacity users claimed a GitHub profile, but 44.1% of Udacity users claimed a GitHub profile which was observed. About 4.5% of Udacity users appear to have deactivated their GitHub account after declaring it on Udacity. From informal exploration, one explanation seems to be that these individuals are not professional programmers. They may be business analysts, managers, or others who tried programming for a bit, then deleted their GitHub later.
   2. In an uncorrected regression of GitHub data on employment, r2 was .13, but repos were the only factor with a significant, positive, linear effect (p=.01). Other factors had important complex effects, including a negative linear effect on followers and commits, both reversing on the quadratic, and a positive cubic on followers with a negative cubic on commits.
   3. There appears to be a sweet spot number of commits per year, and plausibly per repository. Repositories, on the other hand, have an unbounded positive effect on employment.
   4. Repository and follower effects tend to attenuate one another, but the cubic follower effect is much stronger. This may indicate that huge followership is most likely to occur by having one or a few breakout repositories, rather than simply having many repositories. In other words, for a developer to have a few high-quality repositories is more important than having many repos or commits. This is exactly what we would expect if the portfolio is to be an effective indicator of skill quality.
   5. Employment was invariant to the interaction of commits and repos. Interpreted another way, the number of repositories was good for employment without regard to the number of commits.
2. Non-logistic
   1. Nnano has a negative linear effect
   2. Githubfollowers are more explanatory of voi\_employed than age! And yes, nedu and nexp too.
3. Logistic
   1. Emphasized state effects

**IV. Next Steps**

1. Obtain code quality metrics
   1. Repo star count
   2. Repo issue count
   3. README exists
   4. Last updated date
   5. Is forked from other
2. Obtain non-Udacity GitHub data.
   1. LinkedIn employment data may be more accurate and less noisy compared to Udacity.
   2. Observe GitHub users without a Udacity profile to see whether having a Udacity profile is associated positively with productivity or employment.
3. Obtain non-Udacity non-GitHub data
   1. Administer a survey to identify salary and other details.
   2. Perhaps administer also the Baseline Attitudinal Survey to this population.