**Udacity Scrape**

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**

A JavaScript program[[1]](#footnote-1) was developed which runs in the context of a Udacity user profile web page and downloads data related to the user. Not all data was available for all users. Where available, collected data included the user’s real name, email address, profile image, linkedIn URL, GitHub URL, resume, professional experience information, and education information which includes both university education and online courses. Users were specifically checked for completing one or more Nanodegrees, which is a peculiar flagship alternative credential established by Udacity.

Whether or not the individual is presently employed, a rough estimate of their age, and which languages they claim to speak were also noted. Linear, nonlinear, and logistic regression analysis was carried out with employment as the left-hand variable and factors of alternative education as the key right-hand variables. Analysis followed the standard reduction pattern established in the Attitudinal Baseline Study[[2]](#footnote-2). It’s worth noting that between samples the Udacity platform apparently changed their website, limiting data available in the last sample.

Samples are divided into three samples groups, which can in turn be placed in two categories. Profile URLs were guessed by inputting human names. The defensible sample is based on taking the top 20 names for each gender over the last century according to the Social Security Administration[[3]](#footnote-3). The broad sample was obtained by inputting every name which crossed my mind. As such it could be a priori doubted for selection bias, and that was checked using the objective, defensible name selection standard. The third sample is another broad sample of random names I thought of, but it was collected on a different day.

**II. Why it Was Done**

The economics of education has not caught up to the modern EdTech industry. Alternative credentials are known in the literature, but robust analysis is absent and scholars have cited a dearth of relevant data. This Udacity scrape study involves a major contribution of raw data and analysis of that data. Because the study harvests personally identifiable information, it provides for integrated cross-platform data sets with balanced panel data in the future.

Perhaps the height of the economics of education literature today is reflected in Caplan’s recent *The Case Against Education*. Caplan’s work focuses on traditional education and the university degree in particular. A more holistic analysis of education would include alternative forms of education, including online learning credentials. It may be that alternative education does a better job of enhancing human capital, or that it too is only a matter of signals. Even if alternative education is only a matter of signals, it may be a more efficient signaling process which uses less time and expense to produce a better-fit signal of employment.

This topic has significance over and above the literature. Understanding how to utilize alternative education may immediately lead to financial or utilitarian gain by employers and individuals in job markets of several industries.

**III. What I Found**

1. Non-logistic analysis
   1. The weak model explains more than 50% of the variable of interest
   2. Sample effects did not survive the weak model
   3. Name truncation did not survive the weak model
      1. TODO: what about first-name-only analysis? Will be done during augmentation variance analysis.
   4. Plain education effects fell out before Nanodegree effects fell out. Nedu2 fell out of the weak model, but all nnano and interacted factors survived.
   5. Structural effects were lost in the strong factor model, so the adjusted r2-maximizing model is generally preferred.
   6. nnano/interacted survived longer than nlang, the number of languages spoken
   7. Country effects weren’t significant, possibly due to sampling, although sampling well-represents the actual platform userbase. State effects were significant, but only two states survived until the strong factor model: CA and MI. These states had vertically-robust negative effects.
   8. Speaking languages other than English had positive effects and more significance than speaking English.
   9. Age has the most robust relation to employment. It’s overwhelmingly robust and also non-linearly complex. There is a positive linear effect, a negative quadratic effect, and a positive cubic effect. Age also has a robust interaction with nnano which attenuates the negative cubic effect on nnano.
   10. Surprisingly, ndet and lastupdate had strong effects. I interpret ndet as reflecting transparency and portfolio diversity. Factual transparency might be due to an individual’s desire to signal that they have much to offer, but transparency might also be attitudinal. Attitudinal transparency may be detected by an individual’s willingness to expose their portfolio even when the portfolio is unimpressive. Attitudinal transparency may be more a matter of personality traits than technical skill.
2. Logistic analysis

1. The program, called udacity-console-scraper.js, is openly available at <https://github.com/Vandivier/data-science-practice/tree/master/js/udacity-study/manually-scraped> [↑](#footnote-ref-1)
2. Described in the file called 2-pager-survey-monkey-1-off.docx which is openly available at <https://github.com/Vandivier/data-science-practice/tree/master/stata/udacity-exploratory-analysis/manually-scraped> [↑](#footnote-ref-2)
3. https://www.ssa.gov/oact/babynames/decades/century.html [↑](#footnote-ref-3)