**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-regression analysis
   1. The typical profile was last updated during 2017. Plain comparison of the employment rate of Udacity profile holders puts the employment rate at .53, which is lower than the average 2017 US employment rate, or employment-to-population ratio of about 60%[[4]](#footnote-4).
   2. However, the US employment rate statistic only counts individuals aged 16 or over, while individuals of any age can obtain a Udacity profile. An estimate of age was crudely constructed, but it was the most robust factor in regression analysis, so it’s pretty good. Still, better measures of age can be developed.
   3. I generated working = 1 if age1 >= 16, and it was positive and highly significant in a simple linear regression with p ~= 0.000 and a coefficient of .46.
   4. `tab voi\_employed workingage` showed an employment rate of 84.9 among Udacity profiles with an estimated age in the working age range.
   5. After correcting for age and country, `tab voi workingandusperson` has an employment rate of 69.8. Employment rates for Udacity profile holders seem to be generally much higher than the general public of their native countries, and the benefit may go disproportionately to non-US countries.
2. 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. Exploratory2-4 show structural importance of states, interacted3, and nnano3.
   6. nnano/interacted survived longer than nlang, the number of languages spoken
   7. Simple regression on country had an r2 of about .6, but country effects were entirely omitted from non-simple standard reduction due to collinearity. 78% of sampled users were from the US.
   8. Country effects were omitted for collinearity. This may be disconcerting, 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.
   9. Speaking languages other than English had positive effects and more significance than speaking English.
   10. 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.
   11. 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.
   12. Q-complexity of the strong model is no different than the maxar model
3. Logistic analysis
   1. The long logit has a pseudo-r2 of .45, noticeably less than the weak linear model’s r2.
   2. Weak and strong models had a q-complexity of 7. It appears a good analysis has irreducibly complex structure at q-complexity of 7.
   3. The weak model obtained a pseudo-r2 of .42, exceeding the standard for excellent model fit[[5]](#footnote-5).
   4. Unimportance of sample group and name truncation reproduced.
   5. Since adjr2 doesn’t exist on logit, a medium model with p-value .3 threshold was used. nnano survived the weak model, interacted survived the medium model, but neither survived the strong model.
   6. resinserting nnano1 and interacted3 on the strong model results in strong intertacted3 and marginally superweak nnano1 (p = .517) but, both coefficients are directionally expected. This multi-directional robustness indicates structural correctness with insignificance attributable to sample size
   7. nedu ‘noob effect’ preserved, nnano non-noob effect preserved. nnano is similar to nexp in this non-noob regard, so maybe larger composition of human capital? Theoretically market forces may shape alternative education to have greater human capital share of composition, like vocational education.

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)
4. <https://data.bls.gov/timeseries/LNS12300000> [↑](#footnote-ref-4)
5. According to McFadden, the creator of pseudo-r2, “values of 0.2 to 0.4 for rho-squared represent excellent fit.” See: <https://stats.stackexchange.com/questions/82105/mcfaddens-pseudo-r2-interpretation> [↑](#footnote-ref-5)