**Udacity Kairos 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 scrape data was augmented using Kairos face recognition technology[[1]](#footnote-1). Kairos was used to identify age, gender, and ethnicity. Kairos is free for limited use, making its use compatible with easy replication.

**II. Why it Was Done**

Generally, sociological factors including gender, ethnicity, and age are standard controls which correlate and explain swaths of economic observations. Specifically, the alternative credential research to date has shown gender to be a significant factor[[2]](#footnote-2), but Udacity scrape data does not specify gender. Moreover, Udacity scrape data uses a rough estimation of age which is expected to be inaccurate compared to the machine learning-driven age estimate provided by Kairos.

This reanalysis of the Udacity scrape study should be considered a significant robustness check, and it obtain superior effect coefficients and model explanation through the reduction of omitted variable bias and the introduction of new data.

**III. What I Found**

1. Most Udacity users provided a profile picture. About 7% of pictures provided were unable to be processed by Kairos due to low quality or other reasons. While this study is limited to use with data provided directly on the Udacity platform, profiles often linked to other websites from which even more images could be obtained. Such websites include LinkedIn, Twitter, and GitHub accounts, as well as personal websites.
2. Linear analysis
   1. The long linear regression obtained an r2 of .73, and the long logit regression obtained a pseudo-r2 of 1, both significantly improving over the models lacking Kairos data.
   2. Rough age is far more accurate than Kairos age. See exploratory7-10 and d1longkairos.
   3. Kairosmaleconfidence had the directionally expected result, but it was superweak, with p = .987 in the long linear model. I converted this continuous variable into the dummy kairosmale, and I was surprised to see that when both variables were included they had p < .3 effects in opposite directions.
   4. Speaking other languages did not survive to the weak model. Apparently, the underlying factor is better explained by the Kairos ethnicity metrics.
   5. nnano1, the linear effect, was the only pure nnano to survive the weak model, and it had a positive effect. Interacted1, 2, and 3 all survived. Samplegroup3 survived in this case, with a weak negative effect. Age and state effects were strong, with moderate ethnic effects, a strong negative male effect, but a strong positive maleconfidence effect.
   6. Kairos ethnic data had p ~.25. When these were dropped, model r2 dropped significantly and many states became unstable. It’s clear that ethnicity and state have significant cross-correlation, and they jointly much more stable and informative than either is independently.
   7. The strong linear model only had age and a single state, New Jersey with a negative coefficient.
3. Logistic analysis
   1. Age was the only factor robust to all models. The noob effect from nedu1 and nedu3 survived to the medium model, as did interacted3 with the expected negative sign.
   2. The odd male/maleconfidence effect was robust to the medium model.
   3. A few states and \_samplegroup3 were robust to the medium model.

1. See <http://kairos.com/about> [↑](#footnote-ref-1)
2. See 2-pager-survey-monkey-1-off.docx at <https://github.com/Vandivier/data-science-practice/tree/master/stata/udacity-exploratory-analysis/manually-scraped> [↑](#footnote-ref-2)