Dating Profile Predictors

Eitan Darwish, Bertan Gulsen, Henry Phipps, Santosh Sivakumar, Shun Yamaya

Background

- OkCupid is a US-based international dating site
    - Registration based, uses many questions to match up members
- Provides a fun real-world application of linear modeling
- OkCupid is sustained through paying users/ user data sold to the public
- Data collected through publicly available github repository

Data

- Data collected from 59,946 San Francisco OkCupid users
- Metrics:
  - <25 miles from San Francisco
  - Had active profiles June 2012 (online in past year)
  - Had at least one picture on profile
- Info was originally scraped using Python script

Model

- Goal: to predict length of descriptive bio based on various categories
  - Interesting variable to explore, might provide insight onto which populations tend to/feel the need to be more verbose
- Used R to clean + process data

Other categories

- Ethnicity
  - 5 categories (white, black, asian/middle eastern/pacific islander, hispanic/latinx, native american)
- Height
  - Left as is, restricted to values between 55 and 85 inches
- Income
  - Left as is
- Sexual orientation
  - 3 categories (bisexual, gay, straight)
- Sex
  - 3 categories (male/female/other)
- Age
  - Left as is

Binned categories

- Education
    - Some college (1) / No college (0)
- Employment
    - Employed (1) / unemployed (0)
- Religiousness
    - Religious (1) / Not religious (0)
- English speaking ability
    - Fluent english speaker (1) / not english speaker (0)
- Relationship status
    - In a relationship (1) / not in a relationship (0)

Principle Equation

Bio Length = B1*Race + B2*Education + B3*Height + B4*Income + B5*Employment + B6*Gender + B7*Sexual 0rientation + B8*Religiosity + B9*English ability + B10* age + B11* Relationship status + Error

Regression Output

COEFFICIENT	POINT ESTIMATE	STANDARD ERROR	COEFFICIENT	POINT ESTIMATE	STANDARD ERROR
Intercept	35.8	34.82	***Income***	-4.68*10^-6	7.83*107
Age	1.58 ***	0.146	***Employment***	7.86	4.65
College Education	22.5 ***	4.493	***Orientation (Gay)***	-26.8 **	8.42
Race (Asian)	3.26	9.97	***Orientation (straight)***	-30.3 ***	6.89
Race (Mixed race)	5.94	6.99	***Sex (Male)***	-6.4	4.11
Race (White)	22.5 ***	6.73	***Single/Availability***	-11.3	6.31
Height	0.23	0.495	***Religiosity***	-3.95	2.78
			English Ability	14.1 ***	2.82

Discussion

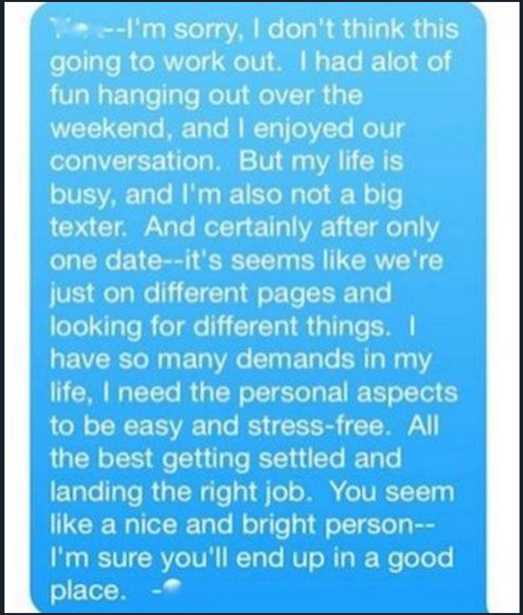
As maybe expected, being:

- Educated
- Employed
- Straight
- White
- Good at english

contribute to longer bios.

This may indicate a positive relationship between US majority-group status and wordiness. Might be interesting to explore the relationship between bio length and dating-match success!

“Casual”



[redacted]--I'm sorry, I don't think this going to work out. I had alot of fun hanging out over the weekend, and I enjoyed our conversation. But my life is busy, and I'm also not a big texter. And certainly after only one date--it's seems like we're just on different pages and looking for different things. I have so many demands in my life, I need the personal aspects to be easy and stress-free. All the best getting settled and landing the right job. You seem like a nice and bright person-- I'm sure you'll end up in a good place. -[redacted]

Every instance of the word “casual” in an essay adds **135** words to a person’s bio!

Woah, there, sheep!