CS 484 Presentation (Title TBD?)

Nathaniel Bechhofer ⋆	nbechhof@gmu.edu
Omnia Elemary	oelemary@gmu.edu
Iman Khalil	ikhalil2@gmu.edu
Jaclyn Lasky	jlasky2@gmu.edu
Yuran (Helena) Niu	yniu3@gmu.edu

September 18, 2016

How does education predict income?

We'd like to find out how much we can infer about someone's income from their education, both in the United States and Europe.

Data Sources

We use two datasets, the United States Current Population Survey (CPS) and the European Social Survey (ESS).

- Both datasets are from the years 2010, 2012, and 2014.
- Both datasets contain rich information about respondents.
- We can use the US data from 2011, 2013, and 2015 to test our predictions.

The US data has 610,756 observations; the ESS data has 157,261 observations.

Architecture

- We obtained data files from the ESS website and the IPUMS (Integrated Public Use Microdata Series) in `dta` and `dat` formats.
- We used the software Stata to apply given data definitions from IPUMS to the `dat` file to obtain an informative `dta` file.
- With `pandas`, a Python package, we read the `dta` files as dataframe objects to use in a Python 3.5 ecosystem.
- Within this ecosystem, we used the `seaborn` package to visualize the data.

Feature Selection

For both data sets, we have variables telling us the following about respondents:

- Educational Attainment
- Age
- Gender
- Location (US state or European country)

In addition, the European respondents consistently report parental education. Since we are using the Annual Social and Economic Supplement (ASEC) data from the CPS, we have fairly rich data about the occupation and employment status of respondents, which can help us clarify the particular channels that education affects income through.

Performance Evaluation

. . .

Model Selection

. . .

Platforms

- Python 3.5
- `pandas` `dataframe`
- `sci-kit learn` for ML algorithms