Investment Value of Education in the U.S. and Europe

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

Team 4

November 8, 2016

How does education predict income?

Problem

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.

Why it matters

If income matters for variables of interest from health to happiness, we can make indirect inferences about individuals and groups using individual or average education levels *combined with other information*.

Some motivating facts

- "College degree holders enjoy an 84 percent increase in earnings over their high-school-educated counterparts [in the United States]."
- In the United States, income earners at the 90th percentile make roughly 16 times as much as those at the 10th percentile.
- Americans born in the bottom income quintile have a 10% chance of entering the top quintile. That probability **doubles** with a college degree.

Data Sources

We use two datasets, the European Social Survey (ESS) and the United States Current Population Survey (CPS); we obtained data files from the ESS website and the IPUMS (Integrated Public Use Microdata Series) in `dta` and `dat` formats.

- 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 (some of) our predictions.

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

Features

Our dependent variable is measured differently in these datasets: the ESS gives us within country deciles, while the CPS gives a numerical answer in dollars. For the US, we only analyze those between the ages of 30 and 34 making between $10K and $500K.

Architecture

- For the ESS, we obtained a `dta` file with missing values and value labels already applied.
- We used the software Stata to apply given data definitions from IPUMS to the CPS `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 `matplotlib` and `seaborn` packages for exploratory visualization.

Preprocessing

Normalization

Why use the natural log of income?

- The fact that $\ln(1+x) \approx x$ enables us to think of coefficients as percentage changes.
- Economists typically model earnings as a function of education and work experience using the *Mincer equation*, which assumes complementarity.

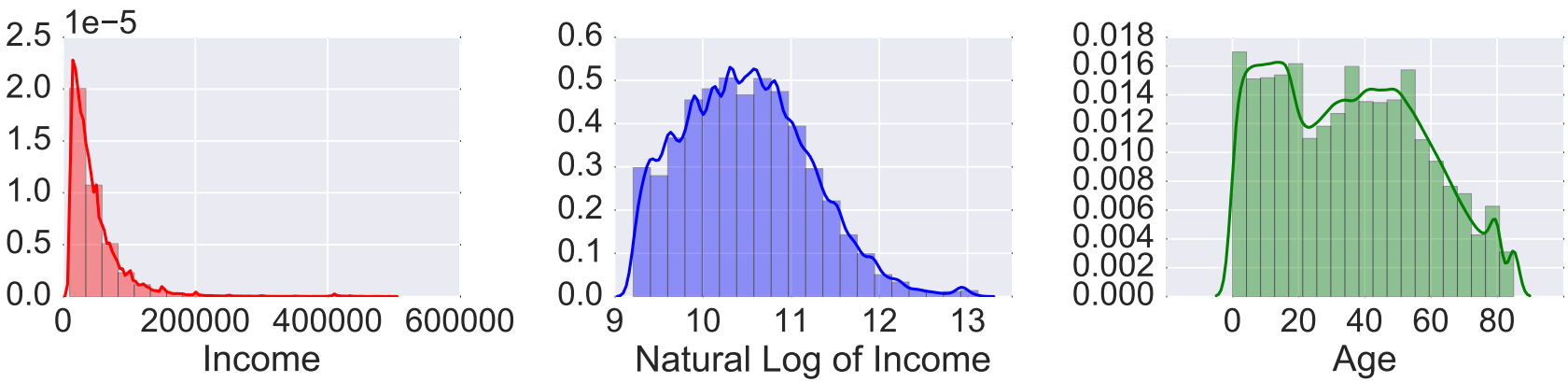

Figure: Sample histograms from `seaborn`

Feature Selection

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

- Educational Attainment
- Age
- Gender

In addition, the European respondents consistently report parental education.
The Annual Social and Economic Supplement (ASEC) data from the CPS (obtained every March) that we are using has rich data about the occupation and employment status of respondents, which can help us clarify how education affects income.

Dimensionality Reduction

Model Selection

Results from Linear Regression

Table: Coefficients for the natural log of income

College	0.511	0.510	0.542	0.541	0.528
Age		0.025		0.025	0.024
Female			-0.282	-0.282	-0.274
Worker					0.362
N	31640	31640	31640	31640	31640
R^2	0.153	0.156	0.201	0.204	0.222

For all of these regression results, we have a sample size of 31640 people, all between the ages of 30 and 34 and with income between \$10000 and \$500000.

Evaluation

Table: Coefficients for the natural log of income

College	0.511	0.510	0.542	0.541	0.528
Age		0.025		0.025	0.024
Female			-0.282	-0.282	-0.274
Worker					0.362
N	31640	31640	31640	31640	31640
R^2	0.153	0.156	0.201	0.204	0.222

For all of these regression results, we have a sample size of 31640 people, all between the ages of 30 and 34 and with income between \$10000 and \$500000.

Ranked Performance

ML Algorithms

Details

Prediction

Plots

Platforms

- Python 3.5
- `pandas` dataframe (built on `NumPy`)
- `sci-kit learn` for ML algorithms
- All brought together with the Jupyter Notebook
- Version control using Github

Discussion

Conclusion