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:

- European Social Survey (ESS)
- 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 both data sets, we have variables telling us the following about respondents:

- Educational Attainment
- Age
- Gender

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.

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

For the US, we only analyze those between the ages of 30 and 34 making between \$10K and \$500K.
We recoded our categorial predictors as dummy variables:

- Whether the person has a college degree (for education).
- If the person is female (for gender).
- If the person is looking for work or has found work (for labor force status).

Moreover, we use the natural log of income as our outcome variable, rather than income itself. Consequently, the fact that $\ln(1+x) \approx x$ means that differences in outcomes can be interpreted as percentage differences; i.e. predicting an increase of 0.1 in our outcome variable corresponds to an approximately 10% increase in income.

Normalization

Why use the natural log of income?

- The distribution of the natural log of income is closer to a Gaussian.
- Economists typically model earnings as a function of education and work experience using the *Mincer equation*, which assumes complementarity.

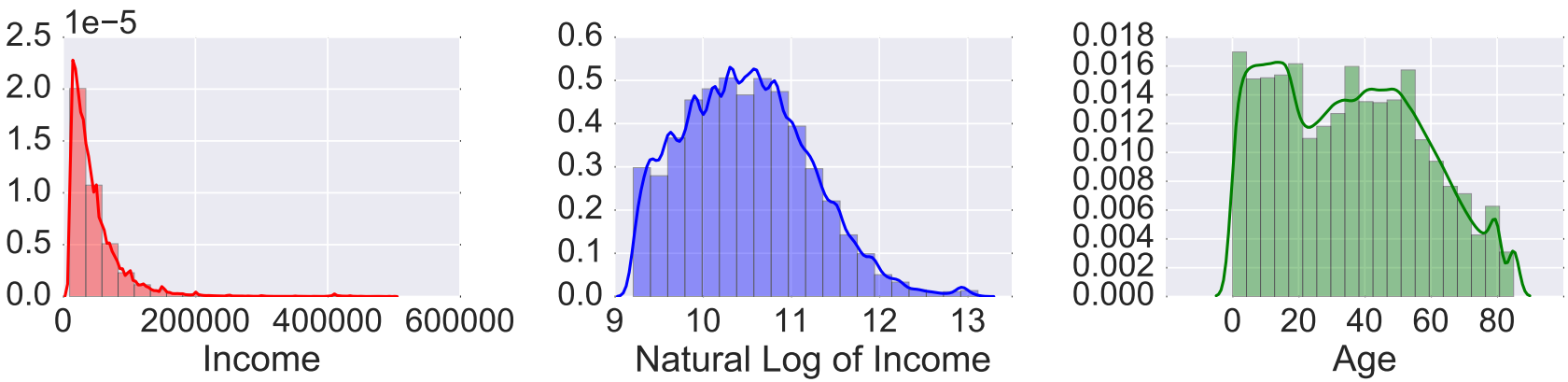

Figure: Sample histograms from `seaborn`

Feature Selection

To select features for prediction, we refer back to the basic problem: how can we best predict income using independent variables that have effects that are not operating via other independent variables include?
We can expect that the effects of age or gender are not primarily operating via, e.g. education.

Dimensionality Reduction

Some variables we don't use:

- Whether one is currently in college
- Birthplace
- Poverty
- Marital Status
- Veteran Status

We are currently working on reducing the dimensionality of occupation, to test whether the returns to education vary a lot by occupational choice.

Model Selection

Our primary tool for model selection will be the R^2 values for linear regression. This calculates how much of the variation in measured income can be accounted for by the model. The values we obtain differ very little from those obtained when using an R^2 that penalizes for overfitting, as our sample size is large with every predictor both statistically and practically significant.

When working with the ESS data, which classifies income by deciles, we compute the total number of false positives, false negatives, and true positives over all classes, and then we compute precision, recall, and f-score using these counts. We then average these scores.

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, our sample is all between the ages of 30 and 34 and with income between \$10000 and \$500000.

Evaluation

For all of our linear regression results, models with more predictors have increased R^2 values.
Moreover, we get consistent results for many of our coefficients, indicating relatively result estimation.

ML Algorithms

For the Regression Tree of maximum depth 2, we obtain an average cross validation score of 0.200 when using 4-fold cross validation. This increases to 0.222 when allowing the maximum depth to be 5, indicating that higher depth is unambiguously better for our purposes.
To come (with the ESS categorical income categorization, where our problem turns into classification):

- Classification Decision Trees
- Back-propogation
- Random Forests
- AdaBoost

Prediction

We are currently in the process of importing the sequestered data for the CPS to test the predictions we made, along with testing for the effects within our dataset for people aged 35-39.

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

All of the variables included appear to be both statistically and practically significant; we plan to add in additional features that are available to us from both datasets.

Conclusion

Education matters a lot for income. We get a remarkably good fit when using our log-linear model, indicating that education primarily affects how income changes, rather than income levels themselves. Further work might explore the interaction of education and age, along with hierarchical modeling that can incorporate variables than we excluded.