Returns to 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

September 22, 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.

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 `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.

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.

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.

Performance Evaluation

Starting with R^2

Our main tool will be the *coefficient of determination* R^2, defined as the proportion of variance explained by any particular model. For our first models, we can distinguish our choices using the *adjusted* R^2 criterion, defined as $\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$, where n is the sample size and p is the number of independent variables (constructed to penalize overfitting).

Other criteria

- Bayesian information criterion (BIC)
- Cross validation on random subsets and years not included
- Least Absolute Selection and Shrinkage Operator (LASSO) with scaled coefficients

Some sense of the size of coefficients for the natural log of income (with only those of ages 30-34!)

College	1.204	1.201	1.327	1.324	0.887
	(0.0337)	(0.0337)	(0.0328)	(0.0328)	(0.0277)
Age		0.0383		0.0402	0.0248
		(0.0114)		(0.0110)	(0.00924)
Female			-1.609	-1.609	-0.969
			(0.0312)	(0.0312)	(0.0267)
Worker					4.170
					(0.0318)
Constant	8.685	7.461	11.10	9.817	1.958
	(0.0198)	(0.364)	(0.0506)	(0.355)	(0.304)
N	41254	41254	41254	41254	41254
adj. R^2	0.030	0.030	0.089	0.089	0.357

Standard errors in parentheses

Model Selection

Our initial pass through the data consisted exclusively of Ordinary Least Squares (OLS) regression models that include gender, age, and educational attainment as independent variables for the natural log of income. We then introduced interaction terms that allow the effects of age and educational attainment to vary by gender. (We plan to use hierarchical modeling to estimate how these effects vary by location and year).

For now, our regressions do not use cases with missing entries for the variables in question.

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.

Platforms

- Python 3.5
- pandas dataframe (built on NumPy)
- sci-kit learn for ML algorithms
- PyStan for Bayesian hierarchical modeling using the Stan probabilistic programming language (itself implemented using C++)
- All brought together with the Jupyter Notebook
- Version control using Github