Returns to Language Skills in the US Labor Market*

Natalie Cook†

April 27, 2021

Abstract

Figuring out which language gives you the biggest wage increase in the US holding as much other stuff constant as possible. Using linear regression plus lots of fancy data science techniques. This is really for fun and I don't think there are any serious implications of the findings. Maybe on a personal level people could choose which language to learn based on what will make them the most money.

*Acknowledgements here, if any.

†Department of Economics, University of Oklahoma. E-mail address: natalie.c.cook-1@ou.edu

1 Introduction

This is a silly topic. I'm doing it to entertain myself and learn. There are no cancer-curing implications of it.

2 Literature Review

Previous work by Ingo Isphording shows that educational decisions are an important determinant of later-life earnings. This point is driven further in follow-up work by **?** and **?**.

There is one paper that is almost exactly my topic, and then everything else is hyper specific to some other circumstance. I don't think that people actually research this seriously so it's kinda hard to do a lit review.

3 Data

The primary data source for this research is the 2019 American Community Survey from the US Census Bureau. I will put a fancy table reference here to the summary statistics 1. Now I will talk about the survey in general a little bit and the go into depth on the variables I will use and how they are measured. This is important for interpretation later on.

4 Empirical Methods

The primary empirical model for the analysis can be depicted in the following equation:

$$Y_{it} = \alpha_0 + \alpha_1 Z_{it} + \alpha_2 X_{it} + \varepsilon, \tag{1}$$

where Y_{it} is a continuous outcome variable for unit i in year t, and Z_{it} are characteristics about the firm at which i is working, while X_{it} are characteristics about i. The parameter of interest is α_1.

log wage is a function of language, occupation, industry, tenure/age, sex, race, location, educational attainment I'll tell you about why I'm using logwage when I figure it out myself - does that count as data cleaning and mutation?

I think I have to make like 100,000 dummy variables because almost everything is categorical here. Y which is log wage is continuous. I think it's gonna be a big long mess.

5 Research Findings

The main results are reported in Table 2. This is where I try to interpret my results. I'll comment on whether the model overall is valid/significant. I'm hoping that it is. I'll talk about heteroskedasticity and robust standard error and other things that I definitely remember that make a model legit. Then I'll interpret the coefficients that get calculated for each variable. Maybe not each variable since some of them are supposed to be controls.

6 Conclusion

If you want to pick a language to learn that will increase your salary - please do an actually nuanced assessment of your situation. If you're gonna go into national security or intelligence of course you should learn Russian or Korean or Arabic or Mandarin. If you need something that is broadly

applicable you should learn Spanish. This is dumb and not actually useful and you shouldn't use it as advice because I can almost guarantee that whatever it tells you to do will actually be wrong. It's a fun exercise though so there's that. Don't actually take this conclusion seriously, this is for my learning and entertainment purposes.

References

Ingo Isphording, Mathias Sinning. ???? "The Returns to Language Skills in the US Labor Market."
.

Figures and Tables

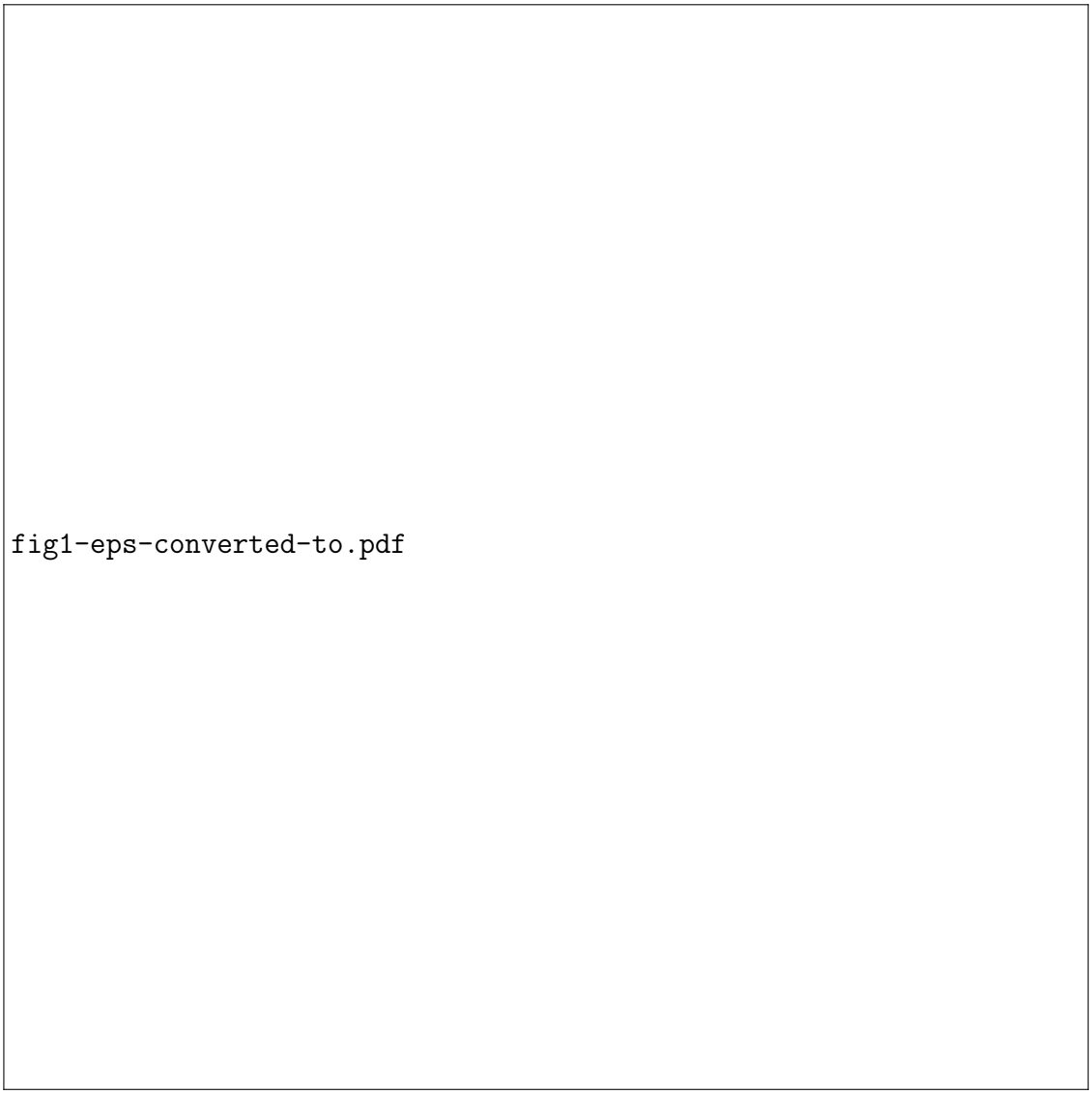

Figure 1: Figure caption goes here

Table 1: Summary Statistics of Variables of Interest

Panel A: Summary Statistics for Variables of Interest

	Mean	Std. Dev.	Min	Max
Outcome variable 1	4.127	1.709	0.000	8.516
Outcome variable 2	1.293	0.648	0.000	0.216
Policy variable	0.685	0.464	0.000	1.000
Control variable 1	0.451	0.497	0.000	1.000
Control variable 2	0.322	0.467	0.000	1.000

Panel B: Sample Means of Outcome Variables for Subgroups

	Group 1	Group 2	Group 3	Group 4
Outcome variable 1	1.782	2.181	3.749	4.127
Outcome variable 2	0.824	0.971	1.215	1.693
N	25,796	75,879	37,157	33,839

Notes: Put any notes about the table here. Sample size for all variables in Panel A is $N = 172,671$.

Table 2: Empirical estimates of parameter of interest

	Few Controls	Many Controls
Variable of interest	-1.977***	-0.536**
	(0.219)	(0.214)
Individual characteristics	✓	✓
Firm characteristics		✓
Location dummies		✓
N	172,671	172,671

Notes: Table notes here. Standard errors in parentheses. ***Significantly different from zero at the 1% level; **Significantly different from zero at the 5% level.