Baseline Regression Results

Luke DiMartino

April 30, 2022

1 Summary

I am investigating gendered wage gaps in developing India. I am interested in two major aspects of the gaps. First, I am conducting a Machado-Mata decomposition of the wage gap. The MM decomposition builds on the traditional Blinder-Oaxaca decomposition, which splits the difference in means between an advantaged group and a disadvantaged group into a sum of two parts: the difference in endowed characteristics and the difference in returns to those characteristics. For example, men in the United States are on average more educated than women. Knowing only this fact about the two groups, a difference in mean wages is not economically unreasonable — men are better endowed with characteristics that make them more productive (whether this endowment is systematically unjust or unfair is outside of the scope of this paper). But, men also earn greater returns to these characteristics. For instance, a man experiences greater return for the same graduate degree than a woman. This is economically unreasonable, the so-called "unexplained" part of the gap.

The MM decomposition expands on the Blinder-Oaxaca decomposition by decomposing this gap in a similar manner but at different quantiles of the wage distribution.

I am also interested in the different returns women in particular experience at different quantiles of the wage distribution. As a policymaking tool, this has more obvious implications. Certain covariates may be correlated with higher wages, and causal inference might suggest that in general they raise wages significantly. But, quantile regression allows us to zoom in on certain parts of the distribution to determine whether that effect varies by wage level.

2 Baseline OLS Results

This table is a general ordinary least squares regression on the entire dataset. This table demonstrates a few key results.

1. The naive estimation of the mean wage gap is somewhere between 41% and 50%, using exact calculations from the coefficient on female.

2. There is not a dramatic difference between White and clustered standard errors. Comparing columns 1 and 2, the standard errors in the second are slightly larger, but not dramatically and not enough to affect results. This is important because although clustered standard errors are theoretically correct, computing them is infeasible for more complex models presented below.

3. Occupation is a massive source of wage inequality. Controlling for it between columns 2 and 3 absorbs a significant amount of the wage gap.

4. The coefficient on literacy is negative and statistically significant, suggesting that controlling for all other factors, being literate reduces wages by nearly 8%.

Appendix Table A1: **Pooled OLS Regression Results**

	White SE's	**Clustered SE's**	**Occupation FE's**	**District FE's**
female	-0.402***	-0.402***	-0.338***	-0.348***
	(0.00487)	(0.00519)	(0.00482)	(0.00462)
Literacy	-0.115***	-0.115***	-0.0508***	-0.0765***
	(0.00763)	(0.00789)	(0.00724)	(0.00681)
Years of Education	0.0498***	0.0498***	0.0238***	0.0206***
	(0.00101)	(0.00105)	(0.000980)	(0.000937)
Marital Status	0.0409***	0.0409***	0.0612***	0.0909***
	(0.00577)	(0.00598)	(0.00547)	(0.00516)
Age	0.0345***	0.0345***	0.0258***	0.0193***
	(0.000938)	(0.000963)	(0.000885)	(0.000836)
Squared Age	-0.000307***	-0.000307***	-0.000232***	-0.000176***
	(0.0000114)	(0.0000117)	(0.0000107)	(0.0000101)
Graduate Degree	0.353***	0.353***	0.226***	0.277***
	(0.0194)	(0.0199)	(0.0191)	(0.0188)
Program Income	-0.165***	-0.165***	-0.119***	-0.112***
	(0.0111)	(0.0113)	(0.0108)	(0.0100)
Little English Ability	0.183***	0.183***	0.0910***	0.0362***
	(0.00789)	(0.00798)	(0.00744)	(0.00713)
English Fluency	0.594***	0.594***	0.353***	0.215***
	(0.0145)	(0.0147)	(0.0142)	(0.0140)
Constant	2.265***	2.265***	3.487***	3.920***
	(0.0181)	(0.0188)	(0.122)	(0.125)
Occupation FE's	No	No	Yes	Yes
District FE's	No	No	No	Yes
Observations	95747	95747	95747	95747
Adjusted R^2	0.358	0.358	0.436	0.512
AIC	185057.7	185057.7	172710.3	159337.3
BIC	185209.2	185209.2	173685.7	163939.4
F	2987.2	2637.2	665.2	216.8

Standard errors in parentheses

All models include caste fixed effects.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

This table is a OLS regression on two subsamples, men and women. This is equivalent to using interaction effects with female for every variable, but makes the results a bit more clear. The difference in returns between men and women is considered economically unreasonable, and there are a few variables where this difference is significant. Marital status, in particular, affects men's wages dramatically more than women's, while a graduate degree provides far greater returns to women than to men.

Appendix Table A2: **Pooled OLS Subsample Regression Results**

	Men	Women
Literacy	-0.0585***	-0.105***
	(0.00790)	(0.0143)
Years of Education	0.0201***	0.0187***
	(0.00104)	(0.00230)
Marital Status	0.114***	0.0130
	(0.00660)	(0.00930)
Age	0.0235***	0.0105***
	(0.00105)	(0.00145)
Squared Age	-0.000220***	-0.0000970***
	(0.0000125)	(0.0000179)
Graduate Degree	0.227***	0.425***
	(0.0214)	(0.0431)
Program Income	-0.164***	-0.0798***
	(0.0153)	(0.0136)
Little English Ability	0.0268***	0.0661**
	(0.00752)	(0.0207)
English Fluency	0.178***	0.343***
	(0.0150)	(0.0367)
Constant	3.874***	3.379***
	(0.133)	(0.474)
Occupation FE's	Yes	Yes
District FE's	Yes	Yes
Observations	68798	26949
Adjusted R^2	0.472	0.457
AIC	114995.0	42140.0
BIC	119418.3	46076.8
F	130.1	.

Standard errors in parentheses

All models include caste fixed effects.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3 Moving Forward

Moving forward, I would like to potentially add more control variables and fine-tune the more complex models.