Evaluating and Improving Proxy Selection Frameworks for Welfare Schemes

Development Economics

The Challenge of Accurate Welfare Targeting

We start with the broader problem at hand- **how do you effectively target households for social programs?** There are three reasons why targeting is often ineffective especially in low- and middle-income countries

Issues with targeting

Poverty Lines	Outdated poverty line definitions
Income	Income is misreported, unobservable or unverifiable
Mobility	High degree of mobility closer to the poverty line

A possible solution is the use of Proxy Means Tests(PMTs). The idea is to find household level characteristics that are **able to predict income well** and are also **easily obervable**

1. PMTs use static proxies that aren't updated with socio-economic changes
2. Not specific to diverse population characteristics.

What we aim to do with our paper

1. Find a framework that maximises the predictive capability by choosing the right number and combination of proxies

2. Run said model in a dataset with consumption expenditure (such as consumption survey 2011) and compare with currently used PMTs

3. Take note of the coefficients and see how well they predict consumption in another time period (here, 2022)

4. Understand policy implications by looking and leakage and under coverage for different poverty lines

For reviewing our model, we compare it with the Grosh and Baker model that has been widely used in developing countries. They use different proxy baskets such as education, durables, household characteristics.

Understanding How PMTs work and how we can improve

Proposition: A Model with the Goal of Maximizing Adjusted R square

Data from Household Consumption Survey (2011)

We have 141 variables including consumption on various items, ownership of durables, and household characteristics (age, gender, etc)

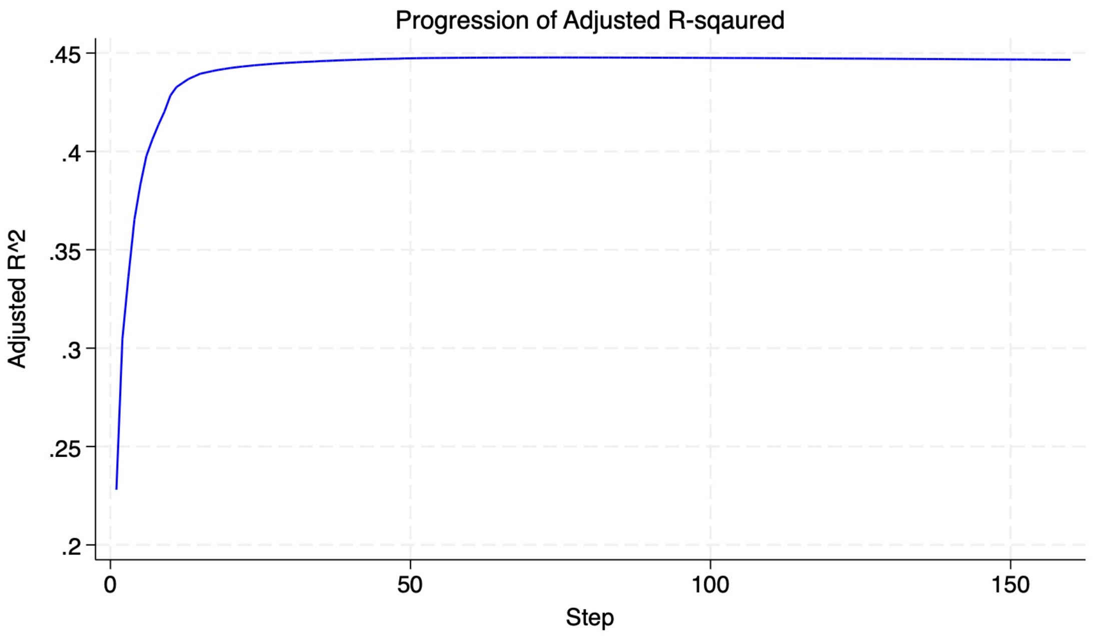
$$\boldsymbol{Consumption} = \boxed{\boldsymbol{\beta_{\{1\}}} * \boldsymbol{Variable\ 1}} + \boxed{\boldsymbol{\beta_{\{2\}}} * \boldsymbol{Variable\ 2}} + \boldsymbol{\beta_{\{3\}}} * \boldsymbol{Variable\ 3} + \ldots$$

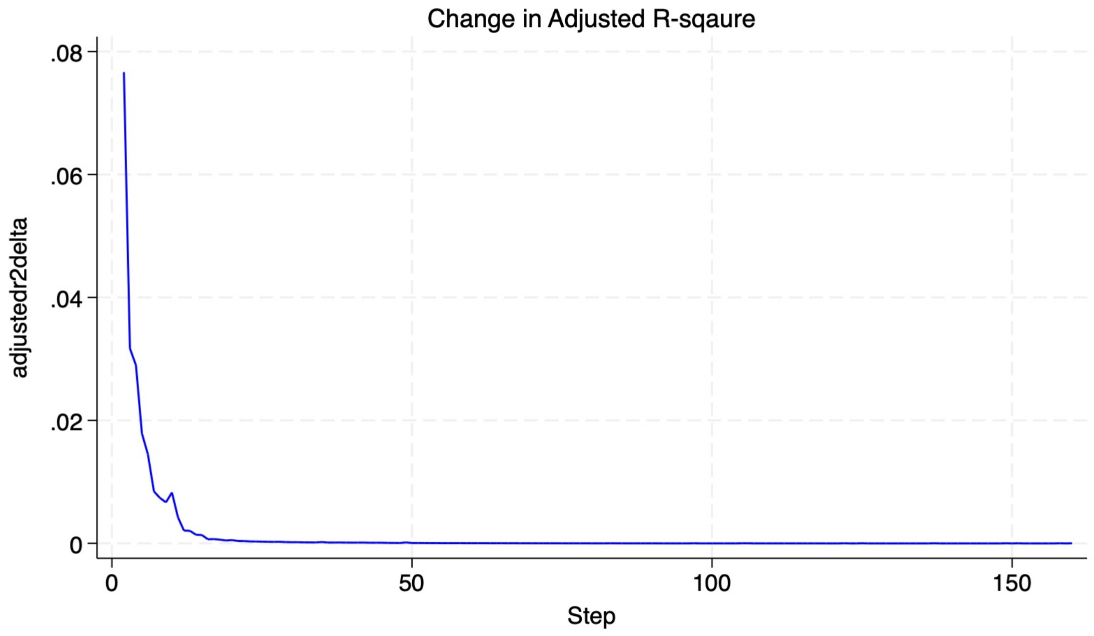
Start with a blank slate. We have 140 variables, regress all on consumption and select variable with highest $\overline{R}^2$. Attach it to the model

We're left with 139 variables, and regress all of them on the first model including variable 1. Again, pick the one with the highest $\overline{R}^2$, and append it to the model

We continue doing this adding the variable that maximize $\overline{R}^2$ until we see that $\overline{R}^2$ starts reducing

Results of the regression: Adjusted R square





Adjusted R-Squared maximizes at the 73rd addition of variables, however the marginal increase (as shown in the right figure) is insignificant after the 30th iteration

Regression results: Comparison

Source	SS	df	MS			
Model	2.6247e+12	29	9.0507e+10	Number of obs	=	99,522
Residual	1.3554e+13	99,492	136230118	F(29, 99492)	=	664.37
				Prob > F	=	0.0000
				R-squared	=	0.1622
				Adj R-squared	=	0.1620
Total	1.6179e+13	99,521	162563793	Root MSE	=	11672

Cons_exp	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
rural	464.34	91.45751	5.08	0.000	285.0844	643.5956
Dwelling_unit_Code_num	-650.3634	40.79347	-15.94	0.000	-730.3181	-570.4087
cons_electricity	5.018252	.1562174	32.12	0.000	4.712067	5.324436
cons_water_bill	4.294743	.7962306	5.39	0.000	2.73414	5.855345
HH_Size_num	792.8436	18.1106	43.78	0.000	757.347	828.3402
Education_num						
02	-220.1121	730.9088	-0.30	0.763	-1652.684	1212.46
03	-998.1222	1521.679	-0.66	0.512	-3980.594	1984.349
04	553.4163	748.2905	0.74	0.460	-913.2239	2020.056

Grosh-Baker Regression

Source	SS	df	MS			
Model	6.9980e+12	47	1.4889e+11	Number of obs	=	100,351
Residual	9.5644e+12	100,303	95355130.6	F(47, 100303)	=	1561.47
				Prob > F	=	0.0000
				R-squared	=	0.4225
				Adj R-squared	=	0.4223
Total	1.6562e+13	100,350	165046510	Root MSE	=	9765

Cons_exp	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
cons_clothes_tot	2.867983	.03585	80.00	0.000	2.797718	2.938249
cons_entertainment_tot	6.917222	.1386698	49.88	0.000	6.645431	7.189013
cons_lemon	111.8916	2.054903	54.45	0.000	107.864	115.9192
cons_med_insti_tot	1.024448	.0144747	70.78	0.000	.9960779	1.052818
cons_misc_HH_consumables_tot	7.2481	.235505	30.78	0.000	6.786513	7.709687
cons_educ_exp_tot	1.011077	.0139518	72.47	0.000	.9837313	1.038422
cons_non_insti_med_tot	1.265053	.040189	31.48	0.000	1.186283	1.343823
cons_egg_meat	1.608287	.0759773	21.17	0.000	1.459373	1.757202
WH_car	2378.127	134.9584	17.62	0.000	2113.611	2642.644
cons_milk	1.086139	.0517	21.01	0.000	.9848077	1.18747
cons_servant	1.781702	.1110405	16.05	0.000	1.564064	1.999341
cons_refined_liquor	1.99706	.1772259	11.27	0.000	1.649699	2.34442

$\overline{R}^2$ Maximising Regression

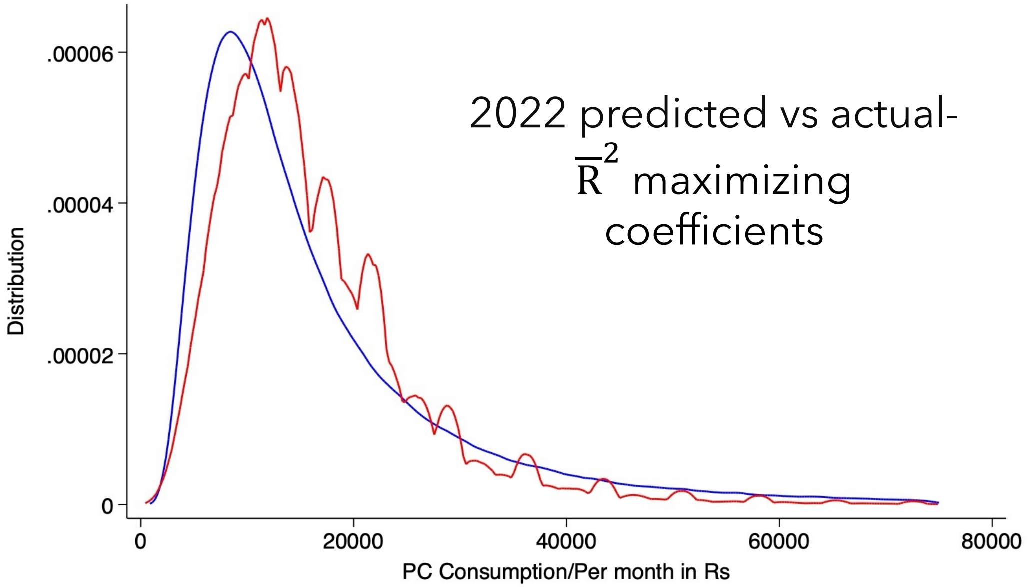
Comparing our regression with the widely used Grosh-Baker model: we see that the adjusted r square is much higher. Also, the selection of covariates in Grosh-Baker is not statistically based and relies on educated guesses. The regression was run on 2011 consumption data

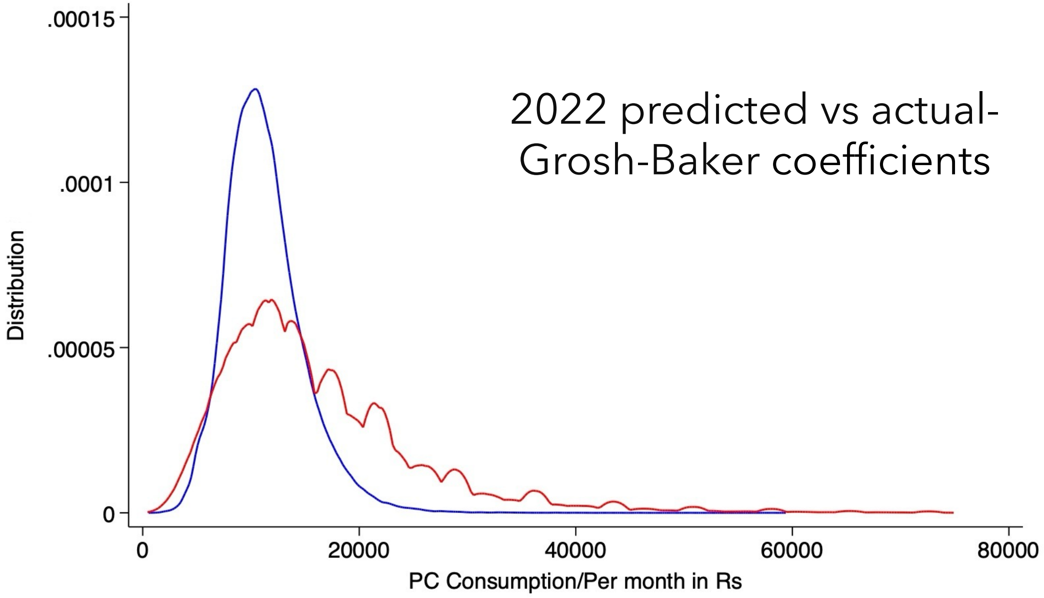
Note: not all covariates are attached in the figures. Please refer to the term paper for the entire list

But do they predict consumption better?

We take note of the coefficients which was shown in the previous slide, and using those coefficients and covariates predict consumption expenditure for 2022. The question remains: **which model predicts consumption better?**

If we overlap actual expenditure on predicted values, here's what they look like:





Policy Applications for this Framework

We aim to study how improved predictors enhance the effective delivery of social programs by targeting benefits to households below specific thresholds, comparing errors across various poverty lines used as eligibility criteria.

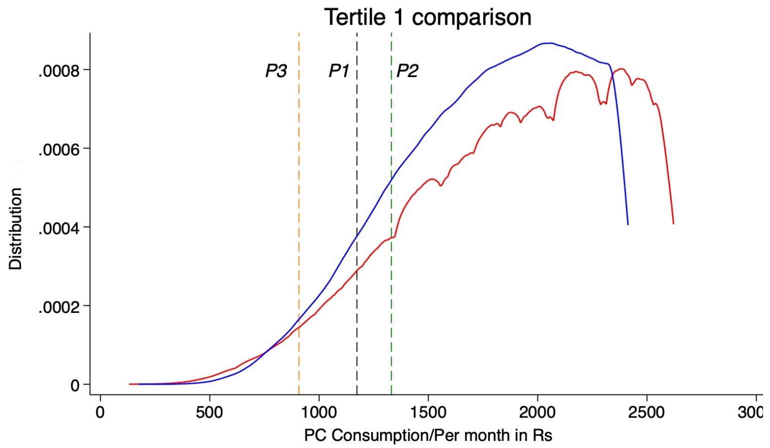
Poverty Lines	Exclusion Error	Accuracy	Inclusion Error
Line 1 @ Rs. 1172.5 per person/mo (*2011 poverty line*)	2.2%	95.23%	2.54%
Line 2 @ Rs. 1331.925 per person/mo (*World Bank poverty line*)	3.35%	92.46%	4.18%
Line 3 @ Rs. 908 per person/mo (*Tendulkar Committee poverty line*)	0.89%	98.35%	0.75%

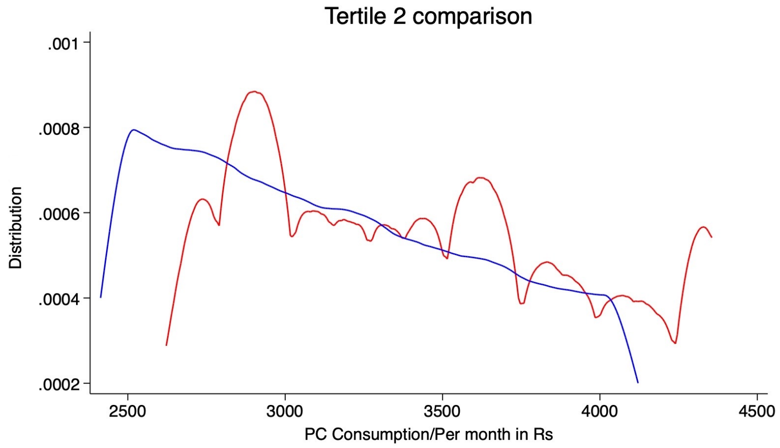
Table 3: Impact of Change in Poverty Lines

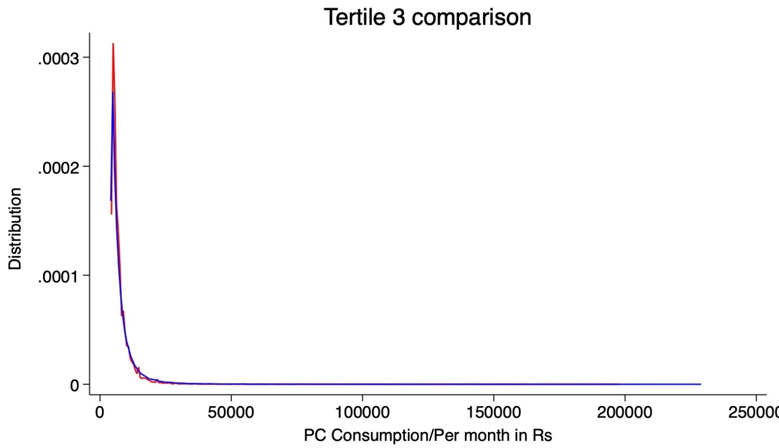
We find that errors, both inclusion and exclusion are minimised when the poverty line is reduced.

Dividing Household by Consumption

We divide households by consumption expenditure into three categories, or tertiles (quartiles but for 3 subdivisions). Then overlap predicted 2022 consumption expenditure from our model and actual expenditure







An interesting observation we find is that our model **is able to better predict the first tertile, or the bottom section of consumption expenditure**

Conclusion and Scope

Takeaways

1. We have provided a framework to better predict income, but don't suggest the mentioned covariates as being the "optimal" ones

2. Covariates and consumption proxies change with income, countries and cultures. A similar framework can be used to determine proxies across regions

3. It's evident PMTs need constant revision to update proxies that utilise adaptive targeting mechanisms

Scope for Future Research

Machine Learning Algorithms to adapt proxies in real time

State specific model that account for regional differences in Indian culture