Problem Set 12

Samuel Hunt

May 1, 2025

1 Imputation Methods

First, I created a summary statistics table and attached it below into this document. From this we can see that around 90% of the women in our sample did not go to college, 63% are married, and 23.7% are holding union jobs. We also see that the mean logwage is 1.7, so it must be in terms of something like thousands of dollars. Overall, this data is fairly reasonable, with a plausable concern that we are underrepresenting women who went to college at just 10%. Additionally, logwages are missing at a rate of 31% which is fairly high. We can assume that this wages dataset is some form of survey data, and so we can assume the worst that the data is MNAR. It is possible that there is an attenuation bias in log wages, such as those with no wages reporting their value as NA, or people being embarrassed to report their log wages. This means that there could be an omitted variable explaining systemic differences in the true value of the omitted log wage variables, meaning the data would be MNAR.

We also created a linear regression to estimate logwages given several different imputation methods: listwise deletion, mean imputation, and Heckman selection. The table of regression outputs for each model is also attached in this document. We know that the true value of $\hat{\beta}_1 = 0.091$. This is the coefficient for returns to schooling. This is fairly different from the coefficients that we actually got from our models, as listwise deletion returned a coefficient of 0.059, mean imputation returned 0.036, and Heckman selection returned a coefficient of .091 This means that listwise deletion was fairly close but still an inaccurate prediction to the true value of the variable. Listwise deletion is consistent with MCAR, and so this indicates that MCAR is an inaccurate description of our data. Mean imputation was even less accurate at 0.036, and because the mean imputation model is consistent with MAR, it indicates that our data is also likely not MAR. Finally, the Heckman selection model was extremely accurate and returned a coefficient of 0.091, identical to the true value of $\hat{\beta}_1$. This model is consistent with the idea that the data is MNAR, which we originally predicted, and so because this model was the best fit we can safely assume the data to be MNAR.

2 Probit Model

We also made a probit model to estimate the proportion of our observations that should choose to work in a union job by estimating their utility of working in a union job. We then altered this model to look at a counterfactual policy in which wives and mothers were unable to work at union jobs. We found our predicted probability of union job employment to be 23.7%, and under the counterfactual in which we did not consider wives and mothers, we found the predicted probability of union job employment to be 22.7%. This indicates that wives and mothers were more attracted to union jobs, as when we do not consider this group the interest in union jobs decreases. The difference between these two estimates is about 1%, so we can say that our counterfactual policy would lead to about a 1% decrease in employment in union jobs. Overall I do think that this model is fairly realistic, however it does have some flaws in the data. I think it would likely be quite important to include the region and industry as well as each are plausible correlated to the prevalence of union employment.

	Unique	Missing Pct.	Mean	SD	Min	Median	Max
logwage	1546	31	1.7	0.7	-1.0	1.7	4.2
hgc	14	0	12.5	2.4	5.0	12.0	18.0
exper	1932	0	6.4	4.9	0.0	6.0	25.0
kids	2	0	0.4	0.5	0.0	0.0	1.0
		N	%				
college	0	1996	89.5				
	1	233	10.5				
married	0	814	36.5				
	1	1415	63.5				
union	0	1700	76.3				
	1	529	23.7				

Table 1: Regression Outputs

	Listwise Deletion	Mean Imputation	Heckman Selection
(Intercept)	0.834	1.149	0.446
	(0.113)	(0.078)	(1.111)
hgc	0.059	0.036	0.091
	(0.009)	(0.006)	(0.066)
union1	0.222	0.068	0.186
	(0.087)	(0.047)	(0.213)
college1	-0.065	-0.126	0.092
	(0.106)	(0.048)	(0.227)
exper	0.050	0.021	0.054
	(0.013)	(0.007)	(0.030)
I(exper^2)	-0.004	-0.001	-0.002
	(0.001)	(0.000)	(0.001)
Num.Obs.	1545	2229	2229
R2	0.038	0.020	0.092
R2 Adj.	0.035	0.018	0.088
AIC	3182.4	3808.4	
BIC	3219.8	3848.4	
Log.Lik.	-1584.189	-1897.193	
F	12.106	9.207	
RMSE	0.67	0.57	0.66