# PS7

Rafael Zago

March 2019

## 1 Question 6

Table 1:

| Statistic | N | Mean | St. Dev. | Min | Pctl(25) | Pctl(75) | Max |
|---|---|---|---|---|---|---|---|
| logwage | 1,669 | 1.625 | 0.386 | 0.005 | 1.362 | 1.936 | 2.261 |
| hgc | 2,229 | 13.101 | 2.524 | 0 | 12 | 15 | 18 |
| tenure | 2,229 | 5.971 | 5.507 | 0.000 | 1.583 | 9.333 | 25.917 |
| age | 2,229 | 39.152 | 3.062 | 34 | 36 | 42 | 46 |

The are 560 missing values for logwage, so the rate is $560/1,669$, with gives us 33.5% of the observations with missing logwage. Thus, this varibale is probably MNAR.

## 2 Question 7

Columns (1) and (2) are exactly the same, which is expected, since column (2) is estimated replacing the missing values of logwage with the mean value, what should not change $\beta_1$. Column (3), on the other hand, presents a lower coefficient for $\beta_1$, a change we should expect, since now, the missing values of logwage are replaced my predictions using the linear regression in the first item of the question. The last models are more realistic then the first two, given that the missing values are not missing at random (MAR) nor (MCAR). One explanation for this is that we cannot observe wages for people who are not in the labor market, which generates a lot of missing variables in a non random manner. The returns to an additional year of schooling, then, is of 5.4% in the model estimated in the third column(and eith the mice package) and 6.2% for the first two estimations.

## 3 Question 8

I am still thinking about estimation strategies and trying to set up some alternative data sets, talking to some professor about it, etc. The main data set I will use for this class is Brazilian Census data. In order to estimate the model, I am still thinking about a best strategy. I first thought about using a Diff-in-Diff, but, since I am studying a phenomenon that is still happening (Venezuelan migration to Brazil), it may not be the best one. Now, I am thinking of using a simple OLS with some interaction terms, or even a synthetic control method. Since I am treating the study as a natural one, a simple OLS should not be a problem.

Table 2: Results

| | *Dependent variable:* | | |
|---|---|---|---|
| | logwage | | |
| | (1) | (2) | (3) |
| hgc | 0.062*** | 0.062*** | 0.054*** |
| | (0.005) | (0.005) | (0.005) |
| collegenot college grad | 0.145*** | 0.145*** | 0.171*** |
| | (0.034) | (0.034) | (0.027) |
| poly(tenure, 2)1 | 4.855*** | 4.855*** | 4.037*** |
| | (0.346) | (0.346) | (0.323) |
| poly(tenure, 2)2 | −1.836*** | −1.836*** | −1.919*** |
| | (0.345) | (0.345) | (0.322) |
| age | 0.0004 | 0.0004 | 0.0002 |
| | (0.003) | (0.003) | (0.002) |
| marriedsingle | −0.022 | −0.022 | −0.022 |
| | (0.018) | (0.018) | (0.014) |
| Constant | 0.709*** | 0.709*** | 0.808*** |
| | (0.145) | (0.145) | (0.119) |
| Observations | 1,669 | 1,669 | 2,225 |
| $R^2$ | 0.208 | 0.208 | 0.156 |
| Adjusted $R^2$ | 0.206 | 0.206 | 0.153 |
| Residual Std. Error | 0.344 (df = 1662) | 0.344 (df = 1662) | 0.319 (df = 2218) |
| F Statistic | 72.917*** (df = 6; 1662) | 72.917*** (df = 6; 1662) | 68.081*** (df = 6; 2218) |

*Notes:* $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$.