# Problem Set 7

Ana Gallart

March 28, 2023

## 1 Question 6

| | Unique (#) | Missing (%) | Mean | SD | Min | Median | Max |
|---|---|---|---|---|---|---|---|
| logwage | 670 | 25 | 1.6 | 0.4 | 0.0 | 1.7 | 2.3 |
| hgc | 16 | 0 | 13.1 | 2.5 | 0.0 | 12.0 | 18.0 |
| tenure | 259 | 0 | 6.0 | 5.5 | 0.0 | 3.8 | 25.9 |
| age | 13 | 0 | 39.2 | 3.1 | 34.0 | 39.0 | 46.0 |

Log wages are missing at a 25 percent rate, given the data is from women in the '80s in the workforce, I think the the gaps in the data are missing not at random. I believe there is a systematic reason for this 35 percent missing logwage.

## 2 Question 7

I tried veery hard to get the modelsummary table of the 4 regression models to print. I do not know why my RStudio continued to fritz up on that last line. I tried many different ways to do it, I left a few of them on the R script so you could see and potentially tell me where it failed. I was able to print "regmod" on its own to see the results, as well as seeing the individual summary tables.

'Complete Cases Regression'
Call: lm(formula = logwage hgc + college + tenure + I(tenure$^2$) + $age$ + $married, data$ = $wagesMCAR$)
Coefficients: (Intercept) hgc collegenot college grad tenure 0.5335692 0.0623931 0.1451682 0.0495251 I(tenure$^2$)$agemarriedsingle$ − 0.00155970.0004412 − 0.0220462
'Mean Imputation Regression'
Call: lm(formula = logwage hgc + college + tenure + I(tenure$^2$) + $age$ + $married, data$ = $wagesMI$)
Coefficients: (Intercept) hgc collegenot college grad tenure 0.7075956 0.0496877 0.1682280 0.0381679 I(tenure$^2$)$agemarriedsingle$ − 0.00132990.0002001 − 0.0268326
'Fitted Value Imputation Regression'
Call: lm(formula = logwage hgc + college + tenure + I(tenure$^2$) + $age$ + $married, data$ = $wagesMAR$)
Coefficients: (Intercept) hgc collegenot college grad tenure 0.5335692 0.0623931 0.1451682 0.0495251 I(tenure$^2$)$agemarriedsingle$ − 0.00155970.0004412 − 0.0220462
'Multiple Imputation Regression' Class: mipo m = 5 term m estimate ubar b t dfcom df riv 1 (Intercept) 5 6.063326e-01 1.609689e-02 7.624749e-03 2.524658e-02 2222 29.81306 0.56841425 2 hgc 5 6.097254e-02 2.270250e-05 7.806441e-06 3.207023e-05 2222 45.52285 0.41262982 3 collegenot college grad 5 1.253245e-01 7.920086e-04 1.535529e-04 9.762721e-04 2222 105.69561 0.23265343 4 tenure 5 4.078755e-02 1.816404e-05 1.039432e-05 3.063722e-05 2222 23.69809 0.68669656 5 I(tenure$^2$)5 − 1.072165$e$ − 035.259045$e$ − 083.378360$e$ − 089.313078$e$ − 08222220.759650.770868476$age$54.871539$e$ − 055.587311$e$ − 062.159169$e$ − 068.178314$e$ − 06222238.831810.463729827$marriedsingle$5 − 1.769301$e$ − 022.235500$e$−041.101504$e$−052.367680$e$−042222796.021090.05912796$lambdafmi$10.362413340.4012751020.292100460.3212

**2.1 The true value of $\beta_1 = 0.093$. Comment on the differences of $\beta_1$ across the models. What patterns do you see? What can you conclude about the veracity of the various imputation methods? Also discuss what the estimates of $\beta_1$ are for the last two methods.**

Complete Cases Regression:0.0623931 Mean Imputation Regression: 0.0496877 Fitted Value Imputation Regression: 0.0623931 Multiple Imputation Regression: .06097254

None of the predicted values are super close to the true value, but the Complete Cases and Fitted Value approximations were the closest. It makes sense that they have the same (or at least would be similar) prediction since the fitted value uses the predicted value of the logwage using the complete cases regression. Multiple Imputations Regression method was the second closest to the true value, that shows it is significantly better than the mean imputation method.

# 3 Question 8

What data are you using? What kinds of modeling approaches do you think you're going to take?

I'm thinking of using housing data from the US Census Bureau, either about Housing Affordibility or Rental Housing Finance. I would like to implement that data with somthing about income to see how housing prices respond to income and other factors. I would like to find what houses people could afford. I have seen that the FRED also has a Housing Affordability Index, but I'm not sure that will work with the more detailed Census Bureau data. I will likely have to use some data imputation methods if too much of my data gets dropped if I remove incomplete observations.