1 Data and Figures

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max
logwage	670	25	1.6	0.4	0.0	1.7	2.3
hgc	17	0	13.1	2.5	0	12.0	18
tenure	259	0	6.0	5.5	0.0	3.8	25.9
age	13	0	39.2	3.1	34	39.0	46
tenure.squared	259	0	66.1	102.5	0.0	14.7	671.7

Log wages are missing at a rate of 25 percent. It is likely that the missing log wages are Missing Not at Random (MNAR). The wage values are missing due to reasons that are not random. Log wage cannot be fully accounted by other variables for which we have information.

	Model 1	Model 2	Model 3	Model 4
(Intercept)	0.534	0.708	0.534	0.536
	(0.143)	(0.114)	(0.112)	(0.112)
hgc	0.062	0.050	0.062	0.062
	(0.005)	(0.004)	(0.004)	(0.004)
tenure	0.050	0.038	0.050	0.049
	(0.005)	(0.004)	(0.004)	(0.004)
tenure.squared	-0.002	-0.001	-0.002	-0.002
	(0.000)	(0.000)	(0.000)	(0.000)
as.factor(college)not college grad	0.145	0.168	0.145	0.145
	(0.034)	(0.024)	(0.023)	(0.025)
age	0.000	0.000	0.000	0.000
	(0.003)	(0.002)	(0.002)	(0.002)
as.factor(married)single	-0.022	-0.027	-0.022	-0.023
	(0.018)	(0.014)	(0.013)	(0.013)
Num.Obs.	1669	2229	2229	2231
Num.Imp.				5
R2	0.208	0.147	0.277	0.277
R2 Adj.	0.206	0.145	0.275	0.275
se_type	HC2	HC2	HC2	

Compared to the true value of 1 (0.093), we observe lower 1's across all three of the models. Across all models, we observe very low R-squared values. This shows that the independent variables in the sample do not account for a large portion of the variation in logwage. Models 1, 2, and 4 all exhibit the same 1 (0.062) When performing the mean imputation in Model 2, we observe the lowest 1 (0.050) and the lowerst R-squared (0.147). The results show that the mean imputation method performed the worst since it is the farthest from the true value of 1. In Model 1, a one year increase in years of schooling (completed by a woman) yields a 6.2 percent increase in wages. In Model 2, a one year increase in years of schooling (competed by a woman) yileds a 5 percent increase in wages.

For my final research project, I am thinking of looking at the wage gap between African Americans and Caucasian individuals living in the United States. I am also interested at looking at unemployment rates. For the former, I will use data from the Current Population Survey data (CPS) that I retrieved from IPUMS-USA.

I am using a dataset I retrieved from the Kaggle datasets which includes unemployment rates across individuals with different levels of education, gender, and race during the period of 2010 to 2020. I mostly provide summary statistics for the latter. I want to quantify the residual race-based discrimination by examining the wage gap and the differences in unemployment rates.

I believe the discussion about race is important and relevant in 2021. In particular, the topic has gained significant traction in 2020 due to the increase in social awareness which stemmed from the controversial deaths of Breonna Taylor and George Floyd. So far, I have cleaned data, created visualizations, and performed preliminary research on the topic.