

# PS7 Smith

March 26, 2024

## 1 Question 6 Answer and Data Summary

The missing log wage value is 25

	Unique	Missing Pct.	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
logwage_mean	670	0	1.6	0.3	0.0	1.6	2.3
logwage_pred	1221	0	1.7	0.3	0.0	1.7	2.3

## 2 Question 7 Answer and Data Summary

The estimated returns to schooling coefficient varies across the models, but none of them are close to the true value of .093:

- Listwise deletion: 0.062
- Mean imputation: 0.050
- Impute with model 1 predictions: 0.062
- Multiple imputation: 0.062

	Listwise deletion	Mean imputation	Prediction imputation	Multiple imputation
(Intercept)	0.534 (0.146)	0.708 (0.116)	0.534 (0.112)	0.534 (0.146)
hgc	0.062 (0.005)	0.050 (0.004)	0.062 (0.004)	0.062 (0.005)
collegenot college grad	0.145 (0.034)	0.168 (0.026)	0.145 (0.025)	0.145 (0.034)
tenure	0.050 (0.005)	0.038 (0.004)	0.050 (0.004)	0.050 (0.005)
I(tenure^2)	-0.002 (0.000)	-0.001 (0.000)	-0.002 (0.000)	-0.002 (0.000)
age	0.000 (0.003)	0.000 (0.002)	0.000 (0.002)	0.000 (0.003)
marriedsingle	-0.022 (0.018)	-0.027 (0.014)	-0.022 (0.013)	-0.022 (0.018)
Num.Obs.	1669	2229	2229	1669
Num.Imp.				5
R2	0.208	0.147	0.277	0.208
R2 Adj.	0.206	0.145	0.275	0.206
AIC	1179.9	1091.2	925.5	
BIC	1223.2	1136.8	971.1	
Log.Lik.	-581.936	-537.580	-454.737	
F	72.917	63.973	141.686	
RMSE	0.34	0.31	0.30	

All of the models, with exception of the mean imputation were close in their estimates of .062. However, the answers are not close to actual. While they provide an estimate, they're not accurate. This could be because the data that is provided doesn't provide the right information to accurately predict the missing data

### 3 Question 8

For the final project, I've laid out the draft template for each of the graded sections. I've also gotten with the IT department to gain access to the last five to six years of data (instead of just the running last 365 days). It's important that I look further back because the time between initial quotation to actually shipping product can be 6 - 18 months, depending on whether the project is new construction or renovations.

I'm able to easily identify when a quote converts to an order, and the subset of that which is whether a special discount (called and FPA) was applied.

I have figured out that I will have to do some data clean up to remove duplicate jobs, such as pricing up the same job with and without a product option. Counting both of those as separate jobs could skew the data.

In addition to the internal company data mentioned above, I plan to reference the AHRI market share data. This is data provided to all HVAC members monthly and quarterly that shows your company's overall market share details. By being a member, every company is required to report their monthly job bookings and shipments. I also will try to incorporate the HVAC PPV from FRED, but I'm not quite sure how that will tie in with the internal data yet.

After discussing with you further, I'll be using a linear model approach to model the relationship between special discount requests and closing the job. I might focus that to see if there are stronger relationships in certain regions of the country or time of year.