Craig Chosney

STATS II FINAL PROJECT - SOC 71600
May 14, 2023

I. INTRODUCTION

For Jeremy's class last semester I analyzed EEO-1 survey data longitudinally in an attempt to assess the work of Diversity, Equity, and Inclusion (DEI) initiatives at NYC cultural institutions as is compares to the national level. A friend of mine is an HR manager at one such institution, and it is my hope that this research may develop into a Capstone project in partnership with the cultural institution at which they work. EEO-1 data from 2014 to 2020 was utilized to see if an increase in the proportion of minorities—in this case females, people of color, and female people of color—in management roles had occurred over this span of time. In my analysis for that project, the proportion of female POCs in management roles was used as the metric for measuring the success of DEI initiatives.

For this project I am continuing the work of last semester, but I will be using U.S. census microdata and income as a metric for measuring inequality rather than employment numbers. The specific research questions I address in this project are as follows:

1. Which of the factors of age, sex, race, and managerial role contribute to income inequality for individuals employed in the museum and cultural institution job sector of the New York, Newark, and Jersey City Core Based Statistical Area?
2. How does age affect earnings differences for managers and-non managers within our sample?
3. How do the interactions of sex, race, and managerial role affect earnings differences within our sample?
4. How do sex, race, and age affect who becomes a manager within the museum and cultural institution job sector of the New York, Newark, and Jersey City Core Based Statistical Area?
5. Can our measure of goodness of fit be improved for our OLS linear regression models by using a neural net predictive model?

II. DATA

Data was pulled from IPUMS USA with sampling from the 2021 5-year American Community Survey. Variables pulled were as follows: *REGION*, *STATEFIP*, *COUNTYFIP*, *MET2013*, *SEX*, *AGE*, *RACE*, *EMPSTAT*, *OCCSOC*, *INDNAICS*, *WKSWORK2*, *UHRSWORK*, and *INCWAGE*. A five percent density sample was extracted with specific cases selected for Employed New York/Newark/Jersey City CBSA residents using the *MET2013* code (35620) and *EMPSTAT* code (1) for "employed." Data was then further reduced to only workers of "Museums, art galleries, historical sites, and similar institutions" using the INDNAICS code (712) when importing the dataset to STATA:

```
use STATS_FINAL_DATASET_5YR_NY.dta if (indnaics == "712")
```

Binary "dummy" variables were then constructed for our minority categories (Female and POC) using the *SEX* and *RACE* variables. In this instance "POC" refers to any survey respondents that did not designate "1" for "white." A binary variable was also constructed to indicate whether or not a respondent was employed in a management occupational category using the variable *OCCSOC* (Standard Occupational Classification). From 2018 onward, this classification system designates "managers" with a six-digit code beginning with "11." Thus, the following code was implemented for the construction of this management variable:

```
generate is_mng = 0
replace is_mng = 1 if ((substr(occsoc,1,2)) == "11")
label variable is_mng "is manager"
label define is_mng 1 "MANAGEMENT" 0 "NON-MANAGEMENT"
```

Lastly, because I am using *INCWAGE* as a dependent variable to measure inequality, I took steps to normalize its values. I first controlled for values of *INCWAGE* not representing a year's wages with the *UHRSWORK* and *WKSWORK2* variables. Values were dropped if a respondent's "usual hours worked per week" were less than thirty or if "weeks worked last year" were less than forty. The variable *INCWAGE* was then winsorized using a one percent margin for recoding:

```
generate newincwage=incwage
replace newincwage=. if uhrswork<30
replace newincwage=. if wkswork2<4
winsor newincwage, gen(wincwage) p(0.01)
```

Regression models used for the project utilized *WINCWAGE* as a dependent variable to measure income inequality, and *SEX*, *RACE*, *AGE*, and *IS_MNG* as independent variables affecting *WINCWAGE*. After selecting for industry and CBSA and normalizing income, the sample size is 1,119; thus, N = 1,119 for our linear models. Our logistic model uses our sample without dropped wage values and has an N = 1,488.

III. METHODS AND ANALYSIS

1. Which of the factors of age, sex, race, and managerial role contribute to income inequality for individuals employed in the museum and cultural institution job sector of the New York, Newark, and Jersey City Core Based Statistical Area?

```
. regress wincwage age is_female is_poc is_mng
```

Source	SS	df	MS	Number of obs	=	1,119
				F(4, 1114)	=	55.55
Model	4.1154e+11	4	1.0288e+11	Prob > F	=	0.0000
Residual	2.0631e+12	1,114	1.8520e+09	R-squared	=	0.1663
				Adj R-squared	=	0.1633
Total	2.4747e+12	1,118	2.2135e+09	Root MSE	=	43035

wincwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
age	963.4062	92.91013	10.37	0.000	781.1076	1145.705
is_female	888.9358	2588.9	0.34	0.731	-4190.734	5968.606
is_poc	-12978.04	2743.081	-4.73	0.000	-18360.23	-7595.854
is_mng	39949.49	4400.127	9.08	0.000	31316.02	48582.96
_cons	24562.85	4657.603	5.27	0.000	15424.18	33701.51

```
. estat hettest

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variable: Fitted values of wincwage

H0: Constant variance

    chi2(1) = 363.60
Prob > chi2 = 0.0000

. vif
```

Variable	VIF	1/VIF
age	1.01	0.989820
is_female	1.01	0.989960
is_poc	1.00	0.996570
is_mng	1.00	0.996844
Mean VIF	1.01	

For the first regression I have employed an OLS Regression with winsorized wage income as the dependent variable and age, sex, race, and managerial role as dependent variables. This regression model was chosen because the dependent variable is a continuous interval variable and the independent variables are nominal.

An adjusted R-square of 0.1633 indicates that our model explains sixteen percent of variance, and a p value of 0.000 shows that our model is statistically significant. P values for the independent variables interestingly show that the relationship between *wincwage* and *is_female* is not statistically significant. Beta coefficients show positive relationships between the dependent variable and age and managerial role, and a negative relationship between the dependent variable and being a POC. Finally, tests for heteroskedasticity and multicollinearity show none in the model.

To answer our research question, the factors of age, race, and managerial role contribute to income inequality for individuals employed in the museum and cultural institution job sector of the New York, Newark, and Jersey City Core Based Statistical Area. Our OLS regression model indicates that for each unit increase in age, wage income increases by $963.40 on average. The quality of being a manager increases wage income by $39,949.49 on average, and the quality of being POC decreases wage income by $12,978.04.

2. How does age affect earnings differences for managers and-non managers within our sample?

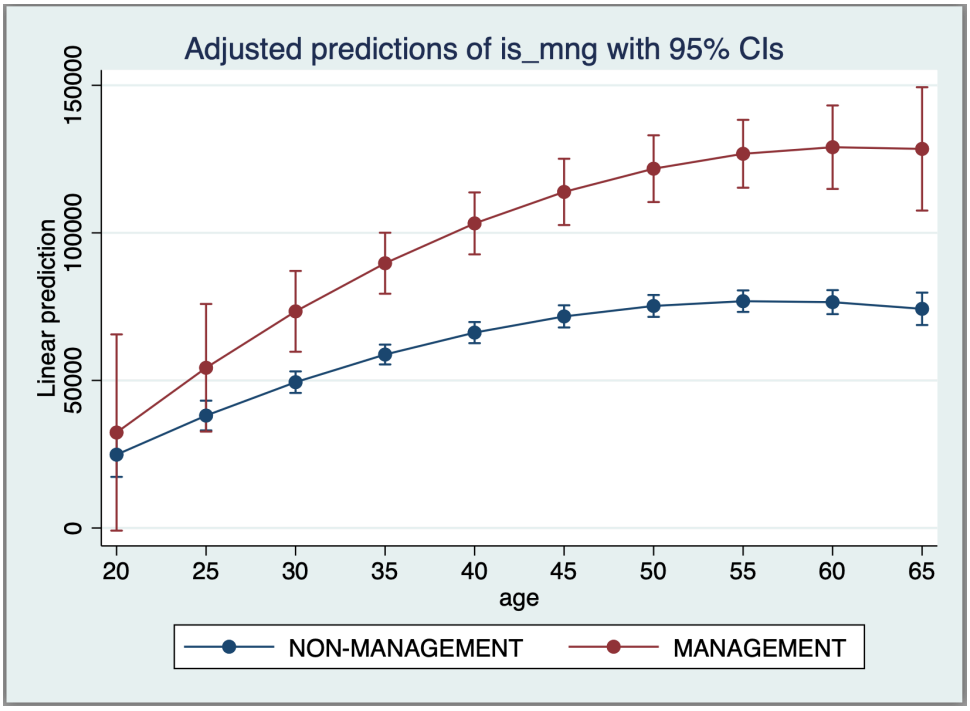
```
. . regress wincwage is_mng##c.age##c.age
```

Source	SS	df	MS	Number of obs	=	1,119
				F(5, 1113)	=	49.25
Model	4.4830e+11	5	8.9659e+10	Prob > F	=	0.0000
Residual	2.0264e+12	1,113	1.8206e+09	R-squared	=	0.1812
				Adj R-squared	=	0.1775
Total	2.4747e+12	1,118	2.2135e+09	Root MSE	=	42669

wincwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
is_mng						
MANAGEMENT	-36021.7	57241.73	-0.63	0.529	-148335.6	76292.17
age	4393.873	612.5542	7.17	0.000	3191.982	5595.764
is_mng#c.age						
MANAGEMENT	2525.819	2614.199	0.97	0.334	-2603.495	7655.133
c.age#c.age	-38.77705	6.737816	-5.76	0.000	-51.9973	-25.55679
is_mng#c.age#c.age						
MANAGEMENT	-17.50766	28.42462	-0.62	0.538	-73.27953	38.26422
_cons	-47506.26	12979.01	-3.66	0.000	-72972.35	-22040.17

```
. quietly margins is_mng, at(age=(20(5)65)) vsquish

. marginsplot
```



In the second model I have again used winsorized wage income as the dependent variable and an interaction term of management status and age as the independent variable. The continuous variable age has been modeled as non-linear in order to allow us to see greater effects in the variance of age with management role.

Although the results of our model look interesting, a p value of 0.538 for our interaction term *is_mng#c.age#c.age* indicates a 53.8 percent chance of seeing these results in the world where the null hypothesis is true, and we must reject our model.

Thus, it is unclear how age affects earnings differences for managers and non-managers in our sample.

3. How do the interactions of sex, race and managerial role affect earnings differences within our sample?

```
. regress wincwage i.sex##is_poc##is_mng
```

Source	SS	df	MS		
				Number of obs	= 1,119
				F(7, 1111)	= 15.79
Model	2.2389e+11	7	3.1984e+10	Prob > F	= 0.0000
Residual	2.2508e+12	1,111	2.0259e+09	R-squared	= 0.0905
				Adj R-squared	= 0.0847
Total	2.4747e+12	1,118	2.2135e+09	Root MSE	= 45010

wincwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
sex						
female	336.7261	3475.62	0.10	0.923	-6482.793	7156.245
is_poc						
POC	-13116.58	4098.616	-3.20	0.001	-21158.48	-5074.674
sex#is_poc						
female#POC	-612.3989	5999.512	-0.10	0.919	-12384.05	11159.25
is_mng						
MANAGEMENT	48341.19	7512.386	6.43	0.000	33601.13	63081.25
sex#is_mng						
female#MANAGEMENT	-14175.42	10708.59	-1.32	0.186	-35186.75	6835.921
is_poc#is_mng						
POC#MANAGEMENT	8727.418	15371.13	0.57	0.570	-21432.31	38887.14
sex#is_poc#is_mng						
female#POC#MANAGEMENT	-20730.36	21035.06	-0.99	0.325	-62003.29	20542.57
_cons	66859.89	2405.889	27.79	0.000	62139.29	71580.48

. **regress wincwage is_poc##is_mng**

Source	SS	df	MS	Number of obs	=	1,119
				F(3, 1115)	=	34.79
Model	2.1183e+11	3	7.0611e+10	Prob > F	=	0.0000
Residual	2.2628e+12	1,115	2.0295e+09	R-squared	=	0.0856
				Adj R-squared	=	0.0831
Total	2.4747e+12	1,118	2.2135e+09	Root MSE	=	45049

wincwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
is_poc						
POC	-13404.85	2995.245	-4.48	0.000	-19281.8	-7527.895
is_mng						
MANAGEMENT	41348.08	5358.105	7.72	0.000	30834.98	51861.19
is_poc#is_mng						
POC#MANAGEMENT	-3697.803	10479.77	-0.35	0.724	-24260.1	16864.5
_cons	67021.23	1737.82	38.57	0.000	63611.47	70431

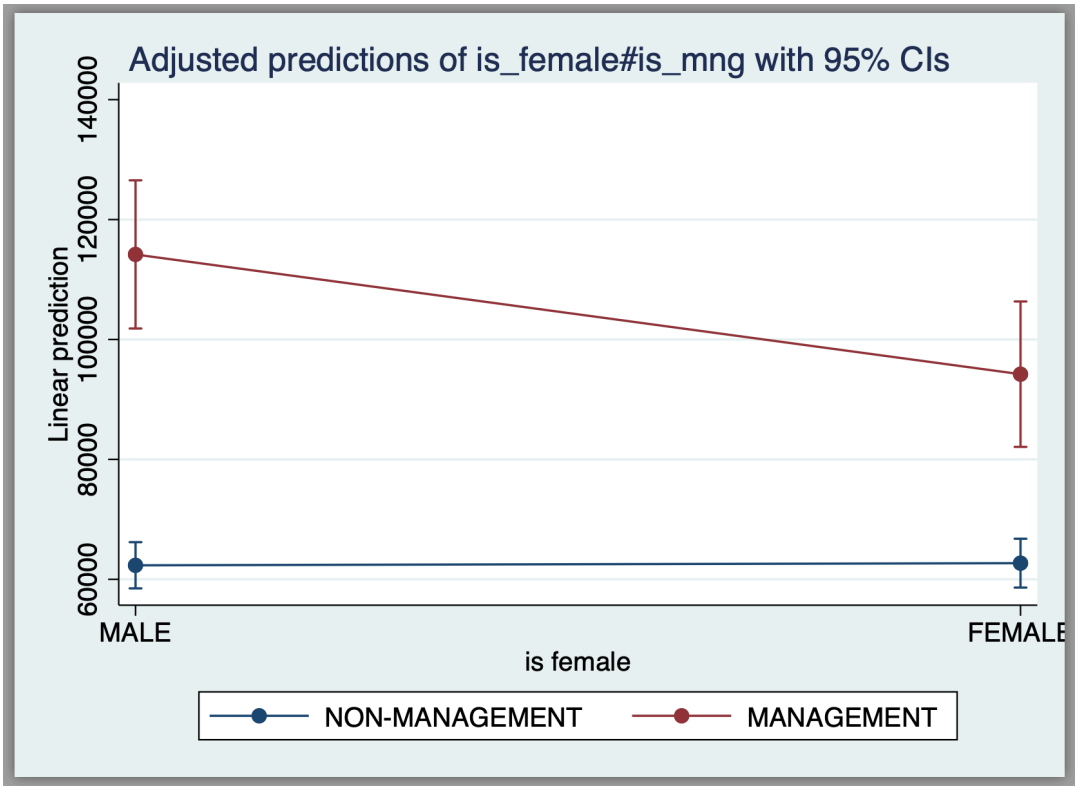
```
. regress wincwage is_female##is_mng
```

Source	SS	df	MS	Number of obs	=	1,119
				F(3, 1115)	=	28.44
Model	1.7590e+11	3	5.8633e+10	Prob > F	=	0.0000
Residual	2.2988e+12	1,115	2.0617e+09	R-squared	=	0.0711
				Adj R-squared	=	0.0686
Total	2.4747e+12	1,118	2.2135e+09	Root MSE	=	45406

wincwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
is_female						
FEMALE	356.4017	2857.439	0.12	0.901	-5250.162	5962.965
is_mng						
MANAGEMENT	51847.88	6596.098	7.86	0.000	38905.71	64790.04
is_female#is_mng						
FEMALE#MANAGEMENT	-20329.96	9273.182	-2.19	0.029	-38524.82	-2135.111
_cons	62340.32	1964.898	31.73	0.000	58485	66195.63

```
.
. quietly margins is_female#is_mng, vsquish

. marginsplot
```



The third model again uses winsorized wage income as the dependent variable and now an interaction term of three nominal variables—race, sex, and management role—as the independent variable.

A p value of 0.325 for our interaction term *sex#is_poc#is_mng* causes us to reject this model due to a lack of statistical significance. The model is run again to see how race and sex interact separately with managerial role. A p value for our interaction term *is_poc#is_mng* of 0.724 causes us to reject this model, but a p value of 0.029 for our interaction term *is_female#is_mng* shows statistical significance.

To answer our research question with this model, we can say that the interaction of the quality of being female and a manager negatively impacts income wages by $20,329.96 on average compared to male managers in the museum and cultural institution job sector of the New York, Newark, and Jersey City Core Based Statistical Area. It is unclear how the interactions of race and managerial role impact earnings.

> *4. How do sex, race, and age affect who becomes a manager within the museum and cultural institution job sector of the New York, Newark, and Jersey City Core Based Statistical Area?*

```
. logit is_mng is_female age is_poc, or

Iteration 0:   log likelihood = -414.71112
Iteration 1:   log likelihood = -412.67812
Iteration 2:   log likelihood = -412.66127
Iteration 3:   log likelihood = -412.66127

Logistic regression                                     Number of obs =  1,488
                                                        LR chi2(3)    =   4.10
                                                        Prob > chi2   = 0.2509
Log likelihood = -412.66127                             Pseudo R2     = 0.0049
```

is_mng	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
is_female	1.137607	.2191159	0.67	0.503	.7799044	1.65937
age	1.005928	.0061101	0.97	0.330	.9940237	1.017975
is_poc	.7111092	.154526	-1.57	0.117	.4644809	1.088691
_cons	.0699138	.0220871	-8.42	0.000	.0376402	.1298595

Note: **_cons** estimates baseline odds.

For our fourth research question we are employing a logistic regression model because our dependent variable is binary with a dichotomous outcome. Our independent variables are sex and race—nominal variables—and age, a continuous interval variable.

Unfortunately, the p values for this model are all above 0.05, and a lack of statistical significance in this model indicates we should reject it. I did, however, run the model again for the national dataset for employees of the museum and cultural institution job sector, with an N of 20,370. The increase in N results in p values of 0.000 across the board, indicating statistical significance. At the national level we see odds ratios greater than one for sex and age, indicating that as the quality of being female and age increase, the log odds of being a manager increase. A odds ratio of less than one for race indicates that as the quality of being POC increases, the log odds of being a manager decrease. Although results for the national sector are compelling, we are unable to answer our research question at the scope of the New York, Newark, and Jersey City Core Based Statistical Area with this model.

NATIONAL SECTOR RESULTS:

```
. logit is_mng is_female age is_poc, or

Iteration 0:   log likelihood =  -5610.868
Iteration 1:   log likelihood = -5566.8896
Iteration 2:   log likelihood = -5566.3656
Iteration 3:   log likelihood = -5566.3655

Logistic regression                                     Number of obs = 20,370
                                                        LR chi2(3)    =  89.00
                                                        Prob > chi2   = 0.0000
Log likelihood = -5566.3655                             Pseudo R2     = 0.0079
```

is_mng	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
is_female	1.214508	.0637647	3.70	0.000	1.095747	1.346141
age	1.01214	.0015663	7.80	0.000	1.009075	1.015215
is_poc	.7706156	.0544666	-3.69	0.000	.6709273	.885116
_cons	.0476912	.0039729	-36.53	0.000	.0405069	.0561497

Note: _cons estimates baseline odds.

5. *Can our measure of goodness of fit be improved for our OLS linear regression models by using a neural net predictive model?*

Using JMP Pro I have performed a neural predictive analysis for our first model, with *wincwage* as our dependent variable and age, sex, race, and management role as our independent variables. This model returns an R-square of 0.2520, indicating that it explains 25.2 percent of variance in our model, as opposed to the sixteen percent of variance that our original, non-neural model explained.

The answer to our research question is that yes, our measure of goodness of fit can be improved for our OLS linear regression models using neural net predictive models. This is a significant improvement in measure of goodness of fit, indicating that deep learning methods are powerful tools for predictive modeling.