Alexander Ryan Hudson
Cottrell – ECN 215

Demographics and Economic Success

Introduction

In this research paper, I endeavor to identify the impact of two major demographic attributes, sex and race, on economic success, namely employment and income. This is obviously a hot button, and admittedly much studied, issue in the United States, and the first step to changing, or coming to terms, with it is to better understand it. In regards to "coming to terms with it," I specifically talk about the impact of sex. There is a well substantiated link between infant health and birth defects and maternal age. Thus, relative to men, women have this fundamental tax/cost on pursuing a career earlier in life, delaying marriage and children. In addition, the logistics of pregnancy and child rearing can place a burden on career pursuits, disproportionately on women. Thus, I believe that, without significant advancements in neonatal and maternal healthcare and virtually ubiquitous access to maternal leave and childcare, there will always be a systemic difference in economic success between the sexes. Obviously econometric analysis could try to account for such differences, but, as long as they are there, biasing will be difficult to completely change. People tend to make decisions in a heuristic based fashion, not algorithmic. Thus, bias based on purely incorrect perceptions can often be, over time, ameliorated through contrary evidence. However, bias based on some fundamental differences will be much more persistent. In essence, even women who devote themselves to their careers, and defer other things, will encounter heuristic based perceptions (bias) as long as there is this fundamental difference. Therefore, systemic changes in employee benefits (such as maternal leave) and reductions in the underlying pressure (medical improvements) would improve the situation, thus reducing the support of the bias and increasing its tendency to change. Such improvements would, obviously, be very complex but valuable. I do not mean to convey that there are no fundamental biases against women, which should be remedied. There is just an additional burden placed on women stemming from the differences in reproductive roles.

Literature Review

As said above, there is a lot of work done. Peterson and Morgan found that compensation differences across occupations were very different, with higher proportion of women indicating a lower wage occupation. However, they found intra occupation differences to be minimal, before and after controlling for individual factors. Gronau looks at the issue I mentioned above. He utilizes a simultaneous equations approach to identify the interactions between wages, planned separations, and skill intensity. However, he does not really delve into the radiating effects, given that he looks primarily at women with planned career interruptions. Ferraro looks at difference in industry and differences in education to conclude a bias, reduced by job evaluations. Raphael and Riker look at race and geographic mobility to explain wage differences, with partial explanation. Overall, some models are above my ability, but they do try to account for differences in ability, effort, and occupation. Some systemic differences in things like proportion of a population with a college degree exist and need to be accounted for in a model.

Data

My data source was the 2016 ACS 1-year Public Use Microdata Samples from the US Census Bureau. Initially, it contained 3,156,487 observations. As data about employment status and income were only gathered for individuals 16 years or older, I restricted the data accordingly. In addition, I wanted to investigate hiring and compensation trends in the civilian sector, so, again, I restricted accordingly. I looked at participation in the labor force in my first model. After, I restricted to remove all non-participants. I looked at employment in my second model. After, I restricted to remove unemployed individuals.

Dependent Variables

The ACS separately gathered 8 different types of income: 1. Wage or Salary Income, 2. Self-Employment Income, 3. Interest, Dividends, Net Rental Income, Royalty Income, or Income from Estates and Trusts, 4. Social Security Income, 5. Supplemental Security Income, 6. Public Assistance Income, 7. Retirement, Survivor, or Disability Income, and 8. All Other Income. Their values were top coded, which reduces the impact of high outliers. Unfortunately, a core attribute of income is its skewed nature. In addition, there is some reason to believe that biasing or other differences would become more pronounced at the higher end of the spectrum. For example, McDonald's is an incredibly blind employer. It does not care about race, sex, or even, to a degree, ability. On the high end however, biasing or other differences could heavily influence compensation and advancement. Thus, the top coded nature of the samples could skew results.

To give a multifaceted insight into the impact of demographics, I modeled with a variety of dependent variables: 1. Participation in the Labor Force, 2. Probability of Employment, 3. Wage, Salary, and Self-Employment Income (1 and 2), 4. Wage, Salary, Self-Employment, and Investment Income (1, 2, and 3), 5. Total Income (All 8 Types). Due to the skewed nature of income, I took the logarithm of those variables. This produced a more normal distribution of values and helped create a more linear relationship, improving the validity of OLS regression.

LaborForce	Boolean variable indicating participation in labor force, civilians only	0 = In labor force 1 = Not in labor force
UnEmploy	Boolean variable indicating employment status, employed is defined as civilians who did any work during the reference week or have a job but did no work due to temporary factors	0 = Employed 1 = Unemployed
TotalEarn	Integer variable indicating sum of wage or salary income and net income from self-employment	Dollars
LogTotalEarn	Log(TotalEarn)	Log(Dollars)
TotalEarnInv	Integer variable indicating sum of wage or salary income, net income from self-employment, and sum of interest, dividends, net rental income, royalty income, or income from estates or trusts	Dollars
LogTotalEarnInv	Log(TotalEarnInv)	Log(Dollars)
TotalInc	Integer variable indicating sum of all eight types of income	Dollars
LogTotalInc	Log(TotalInc)	Log(Dollars)

Independent Variables

As said, I endeavored to look at the impact of demographics factors on economic success, namely sex and race. Given the categorical nature of both, I naturally created dummy variables. Initially, I wanted to test the impact of age as well, but age correlates with experience, which tends to lead to pay increases, more investment income, etc. As the ACS did not provide a quality metric for experience, I could not control for that to allow for possible identification of age discrimination. Thus, age became the control variable for experience.

Sex	Boolean variable indicating gender	0 = Male 1 = Female
Race – A 0 in all subsequent variables indicates solely White		
RBlack	Boolean variable indicating race	0 = Not solely Black or African American 1 = Solely Black or African American
RNative	Boolean variable indicating race	0 = Not solely Alaska Native or American Indian 1 = Solely Alaska Native or American Indian
RAsian	Boolean variable indicating race	0 = Not solely Asian 1 = Solely Asian
RIsland	Boolean variable indicating race	0 = Not solely Native Hawaiian and other Pacific Islander Alone 1 = Solely Native Hawaiian and other Pacific Islander Alone
ROther	Boolean variable indicating race	0 = Not solely another Race, excluding White 1 = Solely another Race, excluding White
RMulti	Boolean variable indicating race	0 = Not Multi Racial 1 = Multi Racial

Control Variables

Given that my goal was to identify possible bias or other factors of demographic attributes, I needed to control for other correlating contributors of success. To control for experience, I used age. To control for ability to mesh into culture, language barrier, etc, I used nativity to the US. To control for work ethic, I used usual hours worked per week and weeks worked in the last 12 months. To control for ability, I used highest educational attainment. To control for differences across industry and employer types, I used occupational category. For occupational category, I utilized OCC codes and categories described by the ACS. For the categorical variables, I created a set of dummies, which includes weeks worked in the last 12 months as the ACS collected as such.

Age	Integer variable measuring age	Years
Nativity	Integer variable indicating nativity to US	0 = Native 1 = Foreign Born
WorkHours	Integer variable indicating usual hours worked per week in last 12 months	Hours
Approximate Weeks worked during the past 12 months – A 0 in all subsequent variables indicates 50 to 52 weeks worked in the past 12 months		
WorkWeek48	Boolean variable indicating weeks worked in last 12 months	0 = Not 48 to 49 weeks worked in the past 12 months 1 = 48 to 49 weeks worked in the past 12 months
WorkWeek40	Boolean variable indicating weeks worked in last 12 months	0 = Not 40 to 47 weeks worked in the past 12 months 1 = 40 to 47 weeks worked in the past 12 months
WorkWeek27	Boolean variable indicating weeks worked in last 12 months	0 = Not 27 to 39 weeks worked in the past 12 months 1 = 27 to 39 weeks worked in the past 12 months
WorkWeek14	Boolean variable indicating weeks worked in last 12 months	0 = Not 14 to 26 weeks worked in the past 12 months 1 = 12 to 26 weeks worked in the past 12 months
WorkWeek1	Boolean variable indicating weeks	0 = Not 1 to 13 weeks worked in the past 12 months

	worked in last 12 months	1 = 1 to 13 weeks worked in the past 12 months
Educational Attainment – A 0 in all subsequent variables indicates highest level of educational attainment is less than high school diploma, GED, or alternative credential		
EHigh	Boolean variable indicating highest educational attainment	0 = Highest education level is not a high school diploma, GED, or alternative credential 1 = Highest education level is a high school diploma, GED, or alternative credential
EAssoc	Boolean variable indicating highest educational attainment	0 = Highest education level is not an Associate's Degree 1 = Highest education level is an Associate's Degree
EBach	Boolean variable indicating highest educational attainment	0 = Highest education level is not a Bachelor's Degree 1 = Highest education level is a Bachelor's Degree
EMast	Boolean variable indicating highest educational attainment	0 = Highest education level is not a Master's Degree 1 = Highest education level is a Master's Degree
EProf	Boolean variable indicating highest educational attainment	0 = Highest education level is not a Professional degree beyond a Bachelor's Degree 1 = Highest education level is a Professional degree beyond a Bachelor's Degree
EDoc	Boolean variable indicating highest educational attainment	0 = Highest education level is not a Doctorate Degree 1 = Highest education level is a Doctorate Degree
Occupational Category – A 0 in all subsequent variables indicates work falls into the managerial category		
OccBus	Boolean variable indicating occupational category	0 = Occupational category is not business 1 = Occupational category is business
OccFin	Boolean variable indicating occupational category	0 = Occupational category is not finance 1 = Occupational category is finance
OccCmm	Boolean variable indicating occupational category	0 = Occupational category is not computer or mathematical 1 = Occupational category is computer or mathematical
OccEng	Boolean variable indicating occupational category	0 = Occupational category is not architecture or engineering 1 = Occupational category is architecture or engineering
OccSci	Boolean variable indicating occupational category	0 = Occupational category is not life, physical, or social science 1 = Occupational category is life, physical, or social science
OccCms	Boolean variable indicating occupational category	0 = Occupational category is not community and social service 1 = Occupational category is community and social service
OccLgl	Boolean variable indicating occupational category	0 = Occupational category is not legal 1 = Occupational category is legal
OccEdu	Boolean variable indicating occupational category	0 = Occupational category is not education, training, and library 1 = Occupational category is education, training, and library
OccEnt	Boolean variable indicating occupational category	0 = Occupational category is not arts, design, entertainment, sports, and media 1 = Occupational category is arts, design, entertainment, sports, and media
OccMed	Boolean variable indicating occupational category	0 = Occupational category is not healthcare 1 = Occupational category is healthcare
OccHls	Boolean variable indicating occupational category	0 = Occupational category is not healthcare support 1 = Occupational category is healthcare support
OccPrt	Boolean variable indicating occupational category	0 = Occupational category is not protective services 1 = Occupational category is protective services
OccEat	Boolean variable indicating occupational category	0 = Occupational category is not food preparation and serving 1 = Occupational category is food preparation and serving
OccCln	Boolean variable indicating occupational category	0 = Occupational category is not building and grounds cleaning and maintenance 1 = Occupational category is building and grounds cleaning and maintenance
OccPrs	Boolean variable indicating occupational category	0 = Occupational category is not personal care and service 1 = Occupational category is personal care and service
OccSal	Boolean variable indicating occupational category	0 = Occupational category is not sales 1 = Occupational category is sales
OccOff	Boolean variable indicating occupational category	0 = Occupational category is not office and administrative support

		1 = Occupational category is office and administrative support
OccFff	Boolean variable indicating occupational category	0 = Occupational category is not farming, fishing, and forestry 1 = Occupational category is farming, fishing, and forestry
OccCon	Boolean variable indicating occupational category	0 = Occupational category is not construction 1 = Occupational category is construction
OccExt	Boolean variable indicating occupational category	0 = Occupational category is not extraction 1 = Occupational category is extraction
OccRpr	Boolean variable indicating occupational category	0 = Occupational category is not installation, maintenance, and repair 1 = Occupational category is installation, maintenance, and repair
OccPrd	Boolean variable indicating occupational category	0 = Occupational category is not production 1 = Occupational category is production
OccTrn	Boolean variable indicating occupational category	0 = Occupational category is not transportation and material moving 1 = Occupational category is transportation and material moving

For a control variable to necessitate inclusion, it must be determinant of the dependent variable and correlated with one or more of the independent variables. Above, I tried to give the deterministic factor that the variables I chose would represent. Below, I have included a correlation matrix. The 5% critical value (two-tailed) = 0.0012, for n = 2584687. I highlighted in red all correlation coefficients that met those requirements. Obviously, it is quite a lot.

	Sex	RBlack	RNative	RAsian	RIsland	ROther	RMulti
Age	0.0523	-0.0403	-0.0205	-0.0466	-0.0131	-0.0846	-0.0729
Nativity	-0.0258	-0.014	-0.0219	0.4214	0.0106	0.2257	0.0017
WorkHours	-0.1931	-0.0227	-0.0022	0.0011	0.0001	-0.0027	-0.0191
WorkWeek48	0.0091	-0.0074	-0.0022	0.0178	-0.0016	-0.0012	0.0032
WorkWeek40	0.0334	-0.0068	-0.001	0.002	-0.0009	-0.0042	0.0053
WorkWeek27	0.0246	0.0081	0.0035	-0.0009	0.0006	0.0044	0.01
WorkWeek14	0.0164	0.0159	0.0083	-0.0017	0.0009	0.0046	0.014
WorkWeek1	0.0172	0.0293	0.0155	0.0033	0.0023	0.0075	0.0198
EHigh	-0.0334	0.0641	0.0282	-0.0965	0.0129	0.0147	0.0073
EAssoc	0.0455	-0.0019	0.0005	-0.0209	-0.0012	-0.023	-0.0002
EBach	0.025	-0.0509	-0.0284	0.0601	-0.0087	-0.0622	-0.0074
EMast	0.0457	-0.0244	-0.019	0.0621	-0.0069	-0.0474	-0.0084
EProf	-0.019	-0.0267	-0.0109	0.0344	-0.0042	-0.0243	-0.0034
EDoc	-0.0148	-0.0194	-0.0078	0.0522	-0.0036	-0.0205	-0.0033
OccBus	0.0232	-0.0076	-0.0073	0.006	-0.0014	-0.0161	0.0001
OccFin	0.0218	-0.0144	-0.008	0.0229	-0.0032	-0.0193	-0.0069
OccCmm	-0.0757	-0.0206	-0.0101	0.0933	-0.0021	-0.0235	0.0017
OccEng	-0.09	-0.0267	-0.0074	0.036	-0.0021	-0.0171	-0.0036
OccSci	-0.0018	-0.0164	-0.0028	0.031	-0.0018	-0.0135	-0.0003
OccCms	0.0429	0.0215	0.0035	-0.0107	0.0005	-0.0095	-0.0003
OccLgl	0.0071	-0.0159	-0.0062	-0.0059	-0.0032	-0.0144	-0.0026
OccEdu	0.129	-0.0154	-0.0048	-0.011	-0.0031	-0.0264	-0.0038
OccEnt	0.0029	-0.0213	-0.0055	-0.0002	-0.0015	-0.0145	0.0041
OccMed	0.135	-0.0093	-0.0104	0.0395	-0.0049	-0.0312	-0.0078
OccHls	0.1129	0.0542	0.0024	-0.0003	0.0003	0.0057	0.0029
OccPrt	-0.0751	0.0274	0.0048	-0.0201	0.0027	-0.0088	0.0021
OccEat	0.0362	0.0091	0.0049	0.0058	0.0039	0.0312	0.018
OccCln	-0.0296	0.0186	0.012	-0.022	0.0026	0.0603	-0.0024
OccPrs	0.1129	0.018	0.0044	0.0265	0.0026	0.0061	0.0055
OccSal	0.0117	-0.0167	-0.009	-0.0077	-0.0009	-0.0058	0.0011
OccOff	0.1874	0.0174	-0.0017	-0.0194	0.0036	-0.0097	0.0022
OccFff	-0.0444	-0.0189	0.0047	-0.0157	-0.0015	0.0475	-0.0018
OccCon	-0.1907	-0.0325	0.0082	-0.0393	0.0019	0.0493	-0.0054
OccExt	-0.0302	-0.0066	0.0026	-0.0077	0	0.002	-0.0015
OccRpr	-0.1551	-0.0206	-0.0002	-0.0221	-0.0012	0.0032	-0.0048
OccPrd	-0.0942	0.0025	0.0017	-0.0024	0.0007	0.0252	-0.0084

OccTrn	-0.1552	0.0372	0.0013	-0.0269	0.0048	0.0251	0

Modeling

Labor Force Participation

To start, I tested the piece of “common knowledge” that women participate in the labor force less than men, due to being a stay-at-home mom or other reasons. Given that women live longer than men and age is such an important factor in labor force participation, I performed a Probit regression with dependent variable LaborForce, independent variable Sex, and control variable Age.

Probit, using observations 3-2584689 (n = 2584687)
Dependent variable: LaborForce
Standard errors based on Hessian

	Coefficient	*Std. Error*	*z*	*p-value*
const	−1.49748	0.00240982	−621.4	<0.0001
Sex	0.203578	0.00164259	123.9	<0.0001
Age	0.0233421	4.31736e-05	540.7	<0.0001

Mean dependent var	0.403748	S.D. dependent var	0.490648
McFadden R-squared	0.095016	Adjusted R-squared	0.095015
Log-likelihood	−1577727	Akaike criterion	3155460
Schwarz criterion	3155499	Hannan-Quinn	3155470
Number of cases 'correctly predicted' = 1895904 (73.4%) f(beta'x) at mean of independent vars = 0.491 Likelihood ratio test: Chi-square(2) = 331299 [0.0000]		Test for normality of residual - Null hypothesis: error is normally distributed Test statistic: Chi-square(2) = 375575 with p-value = 0	

Hypothesis Testing ($p < 0.05$, two-sided)

	Sex	Age
H_0	βSex = 0	βAge = 0
H_a	βSex != 0	βAge != 0
z	123.9	540.7
p	<0.0001	<0.0001
Conclusion	Reject Null	Reject Null

The regression coefficient of Sex is statistically significant at the two-sided $p < 0.05$ level, controlling for age. Thus, we can reject the null and infer a relationship between it and participation in the labor force. $\Phi(-1.49749 + 0.203578) - \Phi(-1.49749) = 0.0307$. Thus, the model predicts that women are 3.07% more likely, than men, to not be in the labor force, not a very large difference. In addition, the regression coefficient of Age is statistically significant at the two-sided $p < 0.05$ level indicating its deterministic value as well, a requirement for its inclusion. After preforming this modeling, I restricted the dataset to remove all of those not participating in the labor force.

Employment

Next, I tested the impact of demographic factors on employment. Given that employment is a categorical variable, I created a dummy. I performed a Probit regression with dependent variable UnEmploy, independent variables of Sex and Race dummies, and control variables of Age, Nativity, WorkHours, and Educational Attainment Dummies.

I tested a model no controls. However, the risk of omitted variable bias necessitated the inclusion of listed controls, with a couple different sets of inclusions tested. In the end, I went with a reasonably inclusive control set, and there was a reasonable shift in the regression coefficients of the independents variables. I did not include occupational coding as that is only gathered for employed individuals, and I did not include weeks worked as it was too predictive of employment and served to drown out the impact of the other variables. Hours worked proved to be a reasonable inclusion because the questionnaire asked for normal hours worked in a week, and, if unemployed, report normal hours worked last time employed in the last twelve months. In addition, it accounts for a tendency to fire part-time individuals first.

Probit, using observations 1-1530499 (n = 1494455)
Missing or incomplete observations dropped: 36044
Dependent variable: UnEmploy
Standard errors based on Hessian

	Coefficient	*Std. Error*	*z*	*p-value*
const	−0.821687	0.00957354	−85.83	<0.0001
Sex	−0.128428	0.00434841	−29.53	<0.0001
RBlack	0.238716	0.00634969	37.59	<0.0001
RNative	0.369894	0.0171068	21.62	<0.0001
RAsian	−0.0541489	0.0115262	−4.698	<0.0001
RIsland	0.0268220	0.0487055	0.5507	0.5818
ROther	0.0210164	0.0107554	1.954	0.0507
RMulti	0.153456	0.0123218	12.45	<0.0001
Age	−0.00920132	0.000143590	−64.08	<0.0001
Nativity	−0.0824832	0.00713099	−11.57	<0.0001
EHigh	−0.121956	0.00671619	−18.16	<0.0001
EAssoc	−0.291272	0.00980728	−29.70	<0.0001
EBach	−0.328444	0.00810639	−40.52	<0.0001
EMast	−0.385411	0.0107098	−35.99	<0.0001
EProf	−0.546148	0.0215966	−25.29	<0.0001
EDoc	−0.491297	0.0249946	−19.66	<0.0001
WorkHours	−0.0122257	0.000168607	−72.51	<0.0001

Mean dependent var	0.030461	S.D. dependent var	0.171851
McFadden R-squared	0.052962	Adjusted R-squared	0.052879
Log-likelihood	−192962.3	Akaike criterion	385958.6
Schwarz criterion	386166.3	Hannan-Quinn	386014.8
*Evaluated at the mean Number of cases 'correctly predicted' = 1448933 (97.0%) f(beta'x) at mean of independent vars = 0.172 Likelihood ratio test: Chi-square(16) = 21582.4 [0.0000]		Test for normality of residual - Null hypothesis: error is normally distributed Test statistic: Chi-square(2) = 735.382 with p-value = 2.05987e-160	

Hypothesis Testing ($p < 0.05$, two-sided)

	Sex	RBlack	RNative	RAsian	RIsland	ROther	RMulti
H_0	βSex = 0	βBlack = 0	βNative = 0	βAsian = 0	βIsland = 0	βOther = 0	βMulti = 0
H_a	βSex != 0	βBlack != 0	βNative != 0	βAsian != 0	βIsland != 0	βOther != 0	βMulti != 0
z	−29.53	37.59	21.62	−4.698	0.5507	1.954	12.45
p	<0.0001	<0.0001	<0.0001	<0.0001	0.5818	0.0507	<0.0001
Conclusion	Reject Null	Reject Null	Reject Null	Reject Null	Fail to Reject	Fail to Reject	Reject Null

Test for omission of variables -
Null hypothesis: parameters are zero for the variables
RBlack
RNative
RAsian
RIsland
ROther
RMulti
Test statistic: F(6, 1.49444e+006) = 325.834
with p-value = P(F(6, 1.49444e+006) > 325.834) = 0
Conclusion: Reject Null

	Age	Nativity	WorkHours
H_0	βAge = 0	βNativity = 0	βWorkHours = 0
H_a	βAge != 0	βNativity != 0	βWorkHours != 0
z	−64.08	−11.57	−72.51
p	<0.0001	<0.0001	<0.0001
Conclusion	Reject Null	Reject Null	Reject Null

Test for omission of variables -
Null hypothesis: parameters are zero for the variables
EHigh
EAssoc
EBach
EMast
EProf
EDoc
Test statistic: F(6, 1.49444e+006) = 529.832
with p-value = P(F(6, 1.49444e+006) > 529.832) = 0
Conclusion: Reject Null

The regression coefficients of Sex, RBlack, RNative, RAsian, and RMulti are statistically significant at the two-sided $p < 0.05$ level. Thus, I can reject the null and infer a relationship between them and probability of being employed, when controlling for experience, ability to mesh into culture, work ethic, ability, and differences across industry and employer types. However, the regression coefficients of RIsland and ROther are not statistically significant at the two-sided $p < 0.05$ level. Thus, I fail to reject the null and cannot infer any relationship. However, hypothesis testing of all race dummies does indicate a statistically significant relationship between race overall, specifically nonwhite, and employment. I included the hypothesis tests of the control variables to confirm their deterministic value as well, a requirement for their inclusion. All have statistically significant regression coefficients. After performing this modeling, I restricted the dataset to remove all of those not employed.

As reference, the dependent variable, UnEmploy, equals 1 if the person is unemployed. Overall, I can infer that women are more likely to be employed than men. I can infer that Asian individuals are more likely to be employed than white individuals. Also, I can infer that Black, Native American, and Multiethnic individuals are less likely to be employed than white individuals. I can infer nothing about the comparative likelihood of being employed for Islandic and Other individuals, which are incidentally very small groups so influencing regression power.

Income

Last, I tested the impact of demographic factors on income. I preformed three different ordinary least squares regression with three different dependent variables measuring income, LogTotalEarn, LogTotalEarnInv, and LogTotalInc. Each dependent variable produced the same basic conclusion. I included the data for LogTotalEarn and LogTotalInc because they are the least inclusive measure of income and most inclusive measure of income. These regressions were performed with independent variables Sex and Race Dummies and with the control variables Age, Nativity, WorkHours, Educational Attainment dummies, Weeks Worked dummies, and Occupational Coding dummies.

I test various models with differing levels of control. In the end, I went with the full inclusion control set, and there was a reasonable shift in the regression coefficients of the independent variables. Also, I performed White's test on my regressions and had heteroscedasticity, so I ran the regressions with robust standard errors.

Model 1: LogTotalEarn as Dependent

OLS, using observations 1-1448933 (n = 1447247)
Missing or incomplete observations dropped: 1686
Dependent variable: LogTotalEarn
Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	*Std. Error*	*t-ratio*	*p-value*
const	8.76518	0.00518961	1689.	<0.0001
Sex	−0.176190	0.00141686	−124.4	<0.0001
RBlack	−0.0760686	0.00207439	−36.67	<0.0001
RNative	−0.0512987	0.00647978	−7.917	<0.0001
RAsian	0.0224375	0.00306480	7.321	<0.0001
RIsland	−0.0133297	0.0143954	−0.9260	0.3545
ROther	−0.0240232	0.00308747	−7.781	<0.0001
RMulti	−0.0317522	0.00419780	−7.564	<0.0001
Age	0.0103696	4.71455e-05	219.9	<0.0001
Nativity	0.0150221	0.00201916	7.440	<0.0001
EHigh	0.201057	0.00246732	81.49	<0.0001
EAssoc	0.325829	0.00305123	106.8	<0.0001
EBach	0.563662	0.00290798	193.8	<0.0001
EMast	0.754569	0.00339032	222.6	<0.0001
EProf	0.932012	0.00581929	160.2	<0.0001
EDoc	0.902796	0.00595037	151.7	<0.0001
WorkHours	0.0328303	8.28777e-05	396.1	<0.0001
WorkWeek48	−0.153982	0.00442873	−34.77	<0.0001
WorkWeek40	−0.307787	0.00279972	−109.9	<0.0001
WorkWeek27	−0.666456	0.00334399	−199.3	<0.0001
WorkWeek14	−1.19031	0.00441109	−269.8	<0.0001
WorkWeek1	−2.11257	0.00571880	−369.4	<0.0001
OccBus	−0.00906603	0.00408698	−2.218	0.0265
OccFin	0.0271042	0.00437181	6.200	<0.0001
OccCmm	0.158194	0.00362897	43.59	<0.0001
OccEng	0.101871	0.00411017	24.79	<0.0001
OccSci	−0.163277	0.00612457	−26.66	<0.0001
OccCms	−0.457534	0.00451441	−101.3	<0.0001
OccLgl	0.0252979	0.00674194	3.752	0.0002
OccEdu	−0.411117	0.00305631	−134.5	<0.0001
OccEnt	−0.329625	0.00591935	−55.69	<0.0001
OccMed	0.0597698	0.00313532	19.06	<0.0001
OccHls	−0.402826	0.00431420	−93.37	<0.0001
OccPrt	−0.219404	0.00445637	−49.23	<0.0001
OccEat	−0.636771	0.00345580	−184.3	<0.0001
OccCln	−0.605061	0.00408213	−148.2	<0.0001
OccPrs	−0.702795	0.00435541	−161.4	<0.0001
OccSal	−0.354235	0.00311581	−113.7	<0.0001
OccOff	−0.335167	0.00266870	−125.6	<0.0001
OccFff	−0.672179	0.00857790	−78.36	<0.0001

OccCon	−0.226466	0.00383735	−59.02	<0.0001
OccExt	−0.143708	0.0177460	−8.098	<0.0001
OccRpr	−0.220526	0.00394490	−55.90	<0.0001
OccPrd	−0.330995	0.00322534	−102.6	<0.0001
OccTrn	−0.439458	0.00341284	−128.8	<0.0001

Mean dependent var	10.33649	S.D. dependent var	1.171426
Sum squared resid	752959.9	S.E. of regression	0.721309
R-squared	0.620860	Adjusted R-squared	0.620848
F(44, 1447202)	37170.34	P-value(F)	0.000000
Log-likelihood	−1580734	Akaike criterion	3161559
Schwarz criterion	3162107	Hannan-Quinn	3161707

Hypothesis Testing (p < 0.05, two-sided)

	Sex	RBlack	RNative	RAsian	RIsland	ROther	RMulti
H_0	βSex = 0	βBlack = 0	βNative = 0	βAsian = 0	βIsland = 0	βOther = 0	βMulti = 0
H_a	βSex != 0	βBlack != 0	βNative != 0	βAsian != 0	βIsland != 0	βOther != 0	βMulti != 0
z	−124.4	−36.67	−7.917	7.321	−0.9260	−7.781	−7.564
p	<0.0001	<0.0001	<0.0001	<0.0001	0.3545	<0.0001	<0.0001
Conclusion	Reject Null	Reject Null	Reject Null	Reject Null	Fail to Reject	Reject Null	Reject Null

Test for omission of variables -
Null hypothesis: parameters are zero for the variables
RBlack
RNative
RAsian
RIsland
ROther
RMulti
Test statistic: F(6, 1.44736e+006) = 341.482
with p-value = P(F(6, 1.44736e+006) > 341.482) = 0
Conclusion: Reject Null

Test for omission of variables -
Null hypothesis: parameters are zero for the variables
WorkWeek48
WorkWeek40
WorkWeek27
WorkWeek14
WorkWeek1
Test statistic: F(5, 1.44736e+006) = 40721.7
with p-value = P(F(5, 1.44736e+006) > 40721.7) = 0
Conclusion: Reject Null

Test for omission of variables -
Null hypothesis: parameters are zero for the variables
EHigh
EAssoc
EBach
EMast
EProf
EDoc
Test statistic: F(6, 1.44736e+006) = 15705
with p-value = P(F(6, 1.44736e+006) > 15705) = 0
Conclusion: Reject Null

Test for omission of variables -
Null hypothesis: parameters are zero for the variables

OccBus OccLgl OccPrt OccFff
OccFin OccEdu OccEat OccCon
OccCmm OccEnt OccCln OccExt
OccEng OccMed OccPrs OccRpr
OccSci OccHls OccSal OccPrd
OccCms OccOff OccTrn
Test statistic: F(23, 1.44736e+006) = 5266.2
with p-value = P(F(23, 1.44736e+006) > 5266.2) = 0
Conclusion: Reject Null

	Age	Nativity	WorkHours
H_0	βAge = 0	βNativity = 0	βWorkHours = 0
H_a	βAge != 0	βNativity != 0	βWorkHours != 0
z	219.9	7.440	151.7
p	<0.0001	<0.0001	<0.0001
Conclusion	Reject Null	Reject Null	Reject Null

The regression coefficients of Sex, RBlack, RNative, RAsian, ROther, and RMulti are statistically significant at the two-sided p < 0.05 level. Thus, I can reject the null and infer a relationship between them and the sum of wage or salary income and net income from self-employment, when controlling for experience, ability to mesh into culture, work ethic, ability, and differences across industry and employer types. However, the regression coefficient of RIsland is not statistically significant at the two-sided p < 0.05 level. Thus, I fail to reject the null and cannot infer any relationship. However, hypothesis testing of all race dummies does indicate a statistically significant relationship between race overall, specifically nonwhite, and the sum of wage or salary income and net income from self-employment. I included the hypothesis tests of the control variables to confirm their deterministic value as well, a requirement for their inclusion. All have statistically significant regression coefficients.

Overall, I can infer that women make less than men, a predicted 17.6% less. I can infer that Asian individuals make more money than white individuals, a predicted 2.2% more. Also, I can infer that all other races, except Islandic, make less than white individuals, predictions ranging from 2.4% less to 7.6% less. I can infer nothing about the comparative earnings of Islandic individuals.

Model 2: LogTotalInc as Dependent

OLS, using observations 1-1448933 (n = 1448020)
Missing or incomplete observations dropped: 913
Dependent variable: LogTotalInc
Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	*Std. Error*	*t-ratio*	*p-value*
const	8.74494	0.00500778	1746.	<0.0001
Sex	−0.213672	0.00138105	−154.7	<0.0001
RBlack	−0.0840602	0.00204253	−41.15	<0.0001
RNative	−0.0559489	0.00640409	−8.736	<0.0001
RAsian	0.00740104	0.00305374	2.424	0.0154
RIsland	−0.0283937	0.0142378	−1.994	0.0461
ROther	−0.0262143	0.00306600	−8.550	<0.0001
RMulti	−0.0233669	0.00415405	−5.625	<0.0001
Age	0.0187944	4.42513e-05	424.7	<0.0001
Nativity	−0.0353920	0.00199776	−17.72	<0.0001
EHigh	0.224182	0.00245246	91.41	<0.0001
EAssoc	0.352821	0.00301323	117.1	<0.0001
EBach	0.608582	0.00287146	211.9	<0.0001
EMast	0.822368	0.00331682	247.9	<0.0001
EProf	1.00528	0.00559405	179.7	<0.0001
EDoc	0.977309	0.00573341	170.5	<0.0001
WorkHours	0.0256363	7.55160e-05	339.5	<0.0001
WorkWeek48	−0.110445	0.00422395	−26.15	<0.0001
WorkWeek40	−0.257155	0.00270396	−95.10	<0.0001
WorkWeek27	−0.564052	0.00328670	−171.6	<0.0001
WorkWeek14	−0.999180	0.00456429	−218.9	<0.0001
WorkWeek1	−1.74953	0.00622172	−281.2	<0.0001
OccBus	−0.0102013	0.00396016	−2.576	0.0100
OccFin	0.0165850	0.00419790	3.951	<0.0001
OccCmm	0.130319	0.00352819	36.94	<0.0001
OccEng	0.0665872	0.00398052	16.73	<0.0001
OccSci	−0.180805	0.00597059	−30.28	<0.0001
OccCms	−0.458272	0.00441723	−103.7	<0.0001
OccLgl	0.00325383	0.00648753	0.5016	0.6160
OccEdu	−0.416031	0.00299610	−138.9	<0.0001
OccEnt	−0.307005	0.00564324	−54.40	<0.0001
OccMed	0.0214953	0.00300189	7.161	<0.0001
OccHls	−0.401843	0.00422720	−95.06	<0.0001
OccPrt	−0.181581	0.00430084	−42.22	<0.0001
OccEat	−0.639651	0.00341543	−187.3	<0.0001
OccCln	−0.586519	0.00396583	−147.9	<0.0001
OccPrs	−0.659791	0.00425119	−155.2	<0.0001
OccSal	−0.341123	0.00297160	−114.8	<0.0001
OccOff	−0.337239	0.00255776	−131.8	<0.0001
OccFff	−0.626745	0.00817360	−76.68	<0.0001
OccCon	−0.248909	0.00367726	−67.69	<0.0001
OccExt	−0.0978516	0.0165888	−5.899	<0.0001

OccRpr	−0.234108	0.00374613	−62.49	<0.0001
OccPrd	−0.335219	0.00306736	−109.3	<0.0001
OccTrn	−0.414555	0.00326387	−127.0	<0.0001

Mean dependent var	10.43441	S.D. dependent var	1.117608
Sum squared resid	720357.5	S.E. of regression	0.705332
R-squared	0.601714	Adjusted R-squared	0.601702
F(44, 1447975)	34663.57	P-value(F)	0.000000
Log-likelihood	−1549144	Akaike criterion	3098378
Schwarz criterion	3098927	Hannan-Quinn	3098527

Hypothesis Testing ($p < 0.05$, two-sided)

	Sex	RBlack	RNative	RAsian	RIsland	ROther	RMulti
H_0	βSex = 0	βBlack = 0	βNative = 0	βAsian = 0	βIsland = 0	βOther = 0	βMulti = 0
H_a	βSex != 0	βBlack != 0	βNative != 0	βAsian != 0	βIsland != 0	βOther != 0	βMulti != 0
z	−154.7	−41.15	−8.736	2.424	−1.994	−8.550	−5.625
p	<0.0001	<0.0001	<0.0001	0.0154	0.0461	<0.0001	<0.0001
Conclusion	Reject Null	Reject Null	Reject Null	Reject Null	Reject Null	Reject Null	Reject Null

Test for omission of variables -
Null hypothesis: parameters are zero for the variables
RBlack
RNative
RAsian
RIsland
ROther
RMulti
Test statistic: F(6, 1.44798e+006) = 304.858
with p-value = P(F(6, 1.44798e+006) > 304.858) = 0
Conclusion: Reject Null

Test for omission of variables -
Null hypothesis: parameters are zero for the variables
WorkWeek48
WorkWeek40
WorkWeek27
WorkWeek14
WorkWeek1
Test statistic: F(5, 1.44798e+006) = 29058.7
with p-value = P(F(5, 1.44798e+006) > 29058.7) = 0
Conclusion: Reject Null

Test for omission of variables -
Null hypothesis: parameters are zero for the variables
EHigh
EAssoc
EBach
EMast
EProf
EDoc
Test statistic: F(6, 1.44798e+006) = 17902.5
with p-value = P(F(6, 1.44798e+006) > 17902.5) = 0
Conclusion: Reject Null

Test for omission of variables -
Null hypothesis: parameters are zero for the variables
OccBus
OccFin
OccCmm
OccEng
OccSci
OccCms
OccLgl
OccEdu
OccEnt
OccMed
OccHls
OccPrt
OccEat
OccCln
OccPrs
OccSal
OccOff
OccFff
OccCon
OccExt
OccRpr
OccPrd
OccTrn
Test statistic: F(23, 1.44798e+006) = 4947.25
with p-value = P(F(23, 1.44798e+006) > 4947.25) = 0
Conclusion: Reject Null

	Age	Nativity	WorkHours
H_0	βAge = 0	βNativity = 0	βWorkHours = 0
H_a	βAge != 0	βNativity != 0	βWorkHours != 0
z	219.9	7.440	170.5
p	<0.0001	<0.0001	<0.0001
Conclusion	Reject Null	Reject Null	Reject Null

The regression coefficients of Sex and all Race Dummies are statistically significant at the two-sided $p < 0.05$ level. Thus, I can reject the null and infer a relationship between them and total income, when controlling for experience, ability to mesh into culture, work ethic, ability, and differences across industry and employer types. Hypothesis testing of all race dummies does indicate a statistically significant relationship between race overall, specifically nonwhite, and total income. I included the hypothesis tests of the control variables to confirm their deterministic value as well, a requirement for their inclusion. All have statistically significant regression coefficients.

Overall, I can infer that women make less than men, a predicted 21.3% less. I can infer that Asian individuals make more money than white individuals, a predicted 0.007% more. Also, I can infer that all other races make less than white individuals, predictions ranging from 2.3% less to 8.4% less.

Conclusion

Overall, my analysis leads to the conclusion that there is bias against race, except Asian, and sex in the workforce and even the free market, as I include those who are self-employed. When controlling for experience, ability to mesh into culture, work ethic, ability, and differences across industry and employer types, these differences continues to exist and are statistically significant. The most biased against group, in compensation, is women, by a large margin. Interestingly, women tend to earn less but are more likely to be employed. This could possibly indicate some sort of value decision in times of layoffs or general firing, where a woman would be kept on over a man because she earns less and does the same thing. As said above, the only group to not experience measurable bias in some way is Asian individuals. In fact, the data points to a pro Asian bias over all other races. Given the degree of separation between the US and most Asian countries, I assumed that Asian immigrants would be disproportionately high skilled. People do not tend to cross half the globe to work at McDonalds. However, this difference would be accounted for in the controls. Thus, the favorable biasing is not something I would not have predicted. Overall, the most biased against race is black individuals they are least likely to be employed and earn the least of the races.

In the future, I would like to improve my model by adding a better metric for work experience (instead of age), more detailed occupational categories, metrics for education types (a person working with an English major in a tech firm might earn less), and include an Hispanic variable. Also, I would want to create a more complex equilibrium model like some that I read about.