Technical Report

Logistic Regression Analysis of US Census Income Data

Douglas Fedorczyk

6/5/2021

Executive Summary

This report examines the US Census income data from 1994. The data was first analyzed to answer the question of whether the interactions between age, sex, and race are important in determining if someone makes more or less than $50K per year and if so, by how much. Next, the data was analyzed to develop a predictive model in determining whether someone's income is greater than $50K per year given the set of variables in the data set. In the first part, there was convincing evidence that a model that included interactions was better than a model that included main effects only. This was based on a drop in deviance test with a resulting p-value much less than 0.0001. While the full model showed that not all interactions had significant p-values, the interaction between age and sex was significant with a p-value much less than 0.0001. The second part of the report found that all variables were important in predicting if an individual's income was greater than $50K per year with an Mean Squared Predictive Error (MSPE) value of 0.1517; using a classification rule of $\hat{p} = 0.49$. However, this is not to say that all categories within each variable were statistically significant and additional discussion is provided in the conclusion as to what changes could be made to improve the MSPE value.

Data

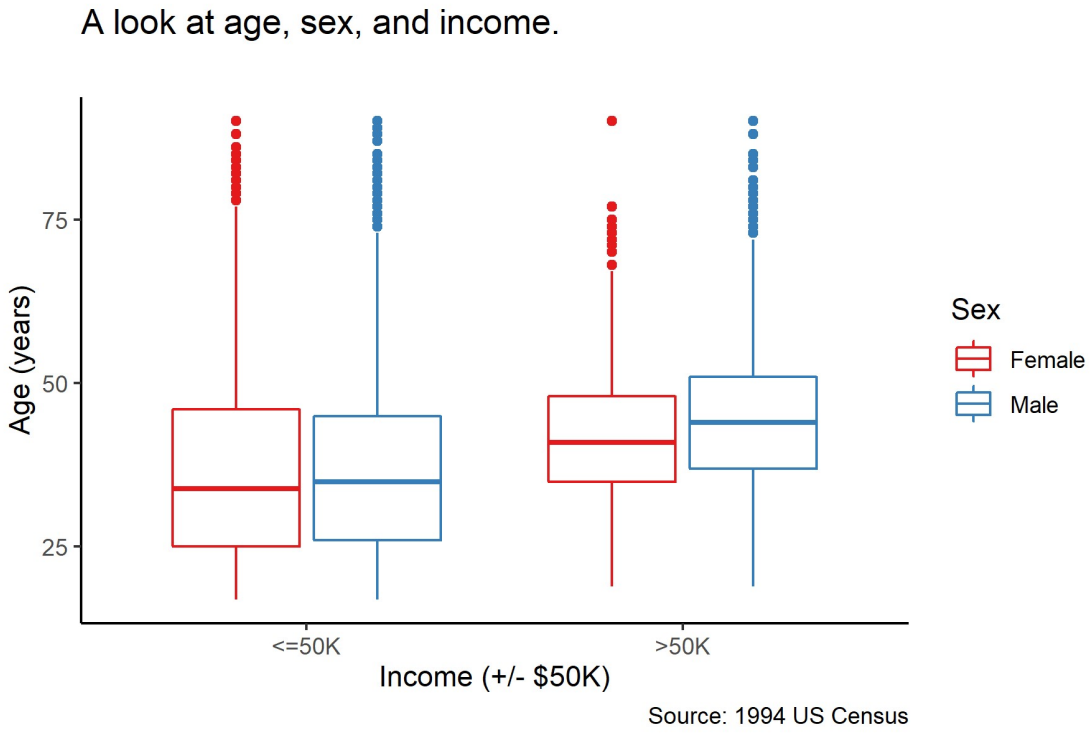
The data set comes from the following University of California Irvine website, , which in turn came from the US Census Bureau website from 1994. The intent of the data set was to determine the probability of earning more than $50K per year using census data. There are 15 variables in total to include the response variable, which indicates if an individual's income is greater than or less than $50K per year. Other variables included age, working class, a weighting factor, education level, number of years of education, marital status, occupation, familial relationship, race, sex, capital gains, capital losses, hours worked per week, and native country. The variables are self explanatory for the most part with the exception of the weighting factor which was a used as a means by the Census Bureau to account for response bias. The main data file contained 32,561 observations. A second data file was created to validate the predictive model generated from the main data file. This second data file contained the same variables with 16,281 observations. There were three variables that had missing values, working class and occupation, both of which had about 6% of their data missing; while the native country variable had roughly 2% missing.

Altogether, the missing values represented less than 10% of the data. As such, it was felt to be low risk to remove the missing values and proceed with the analysis.

Analysis & Results

Explanatory Question

The first question of interest looked at the interactions between age, sex, and race, to determine if these were significant in determining the odds that a person's income is greater than $50K per year and if so, how much of an effect does the interaction have. In the Figure below is a box plot showing income versus age where sex is highlighted to indicate the differences between the two sexes across age.



The figure shows there is a slight difference in the median value of males and females who make more than $50K per year compared to males and females who do not with respect to age. There is a slight difference in the median between males and females who make more than $50K per year with respect to age as well. A similar plot was also generated for age and race which can be found in the Appendix.

The two data files were combined to use all available data to answer the explanatory question. A logistic regression model was fit using a binomial distribution along with a logit link function. Two-way and three-way interactions were considered between the three

variables and a drop in deviance test was used to determine if the interactions were appropriate for the model.

After fitting a model with two-way interactions, it was found that only the interaction between age and sex was statistically significant. For all other interactions, their respective p-values were greater than the nominal cutoff value of 0.05. In fact, as can be seen in Table 1 below, a good portion of the associated p-values were much greater than 0.05.

Table 1: Full Model p-values

Coefficient	Pr(>|z|)
(Intercept)	3.11e-09
age	0.0257
raceAsian-Pac-Islander	0.1386
raceBlack	0.3432
raceOther	0.7129
raceWhite	0.5002
sexMale	0.9805
age:raceAsian-Pac-Islander	0.6174
age:raceBlack	0.7534
age:raceOther	0.6642
age:raceWhite	0.9933
age:sexMale	5.61e-15
raceAsian-Pac-Islander:sexMale	0.2581
raceBlack:sexMale	0.0644
raceOther:sexMale	0.7841
raceWhite:sexMale	0.1156

A second model was fit to the data without interaction terms. In Table 2 below it can be seen that most of the individual variables considered were statistically significant based on their associated p-values being much less than 0.0001.

Table 2: Reduced Model p-values

Coefficient	Pr(>|z|)
(Intercept)	< 2e-16
age	< 2e-16
raceAsian-Pac-Islander	2.21e-09
raceBlack	0.609
raceOther	0.637
raceWhite	1.05e-07
sexMale	< 2e-16

However, after performing a drop in deviance test, there was convincing evidence in favor of the full model with a rather small p-value of 1.3e-11 as shown in Table 3 below. This

suggested that the full model with the two-way interaction terms was the more appropriate model. Also of note was the AIC value for the two models. The full model had an AIC score of 45,679 while the reduced model had a AIC score of 45,732.

Table 3: Main Effects vs Two-way Interactions

Analysis of Deviance Table					
Model 1: income ~ age + race + sex					
Model 2: income ~ age * race + age * sex + race * sex					
	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	45215	45718			
2	45206	45647	9	70.33	1.312e-11

Three-way interactions were also considered. A drop-in-deviance test was performed between the two-way interaction model and the three-way interaction model and found that the two-interactions model was the more appropriate model with a p-value of 0.9031. Output from the test can be found in Table 4 below.

Table 4: Two-way vs Three-way Interactions

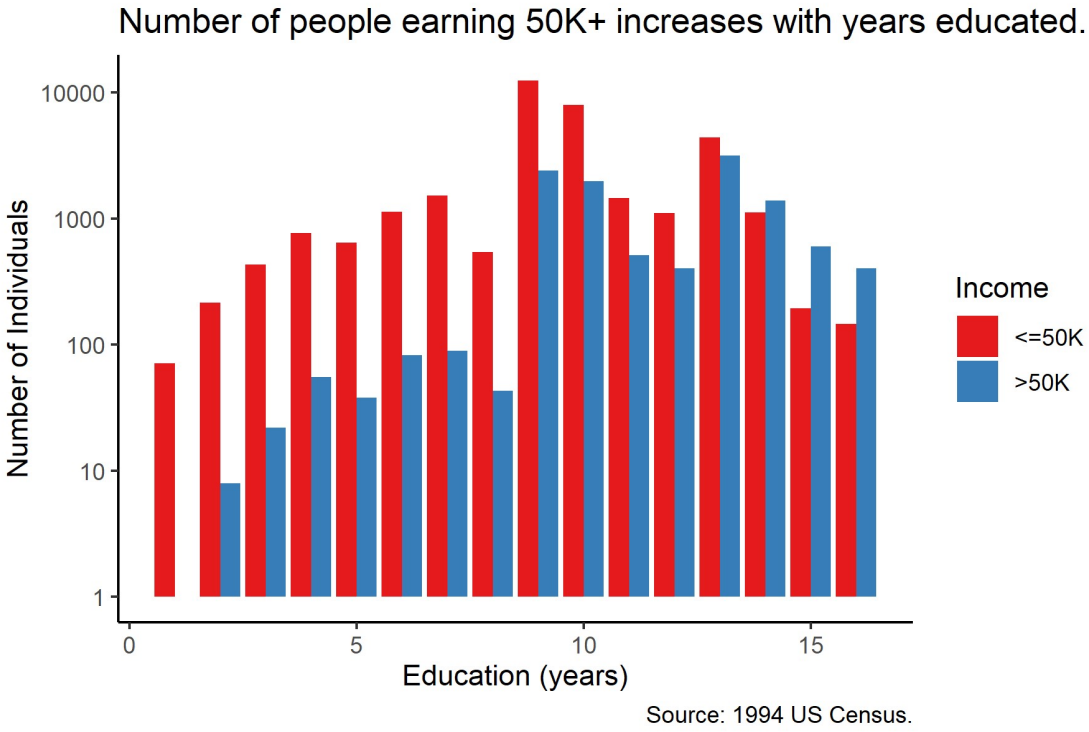
Analysis of Deviance Table					
Model 1: income ~ age * race + age * sex + race * sex					
Model 2: income ~ age * race * sex					
	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	45206	45647			
2	45202	45646	4	1.044	0.9031

A table containing the estimates for each of the terms along with their respective 95% confidence interval can be found in the appendix. From these estimates one can determine the odds ratio for any number of items of interest. For example, if one were to look at the differences between a 30-year-old white male and a 30-year-old white female, based on the coefficient estimates it can be shown that the 30-year-old white male is 2.7 times more likely to make more than $50K per year than the 30-year-old white female. This is illustrated in an example problem located in the appendix. It can be seen how the additional interaction terms for the 30-year-old white male impact the equation both with respect to the intercept as well as the slope.

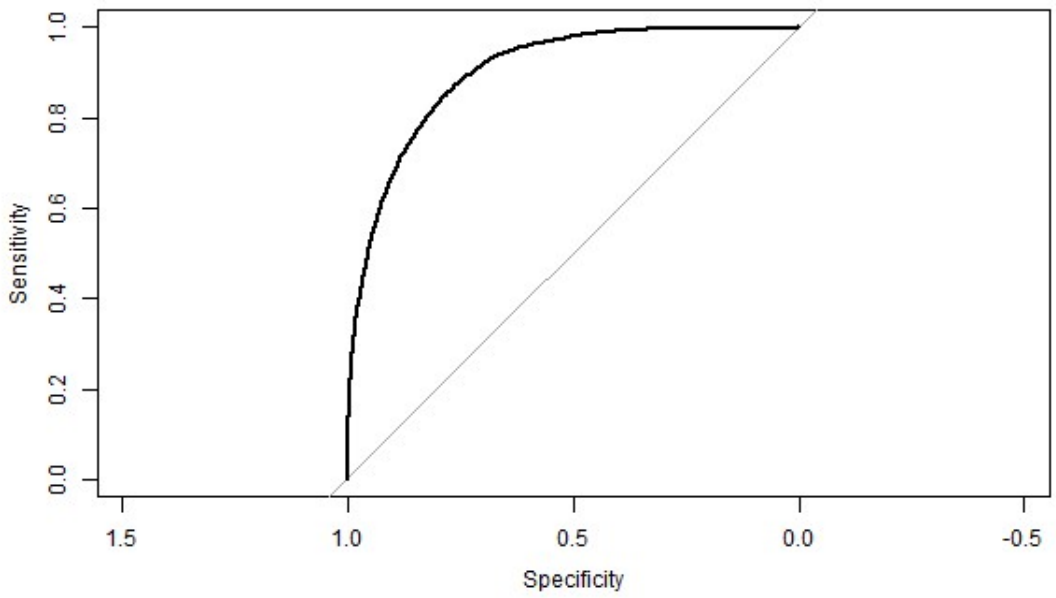
Exploratory Question

The second question of interest considered how accurately a model could predict if someone earned more than $50K per year with the given variables. Several exploratory plots were generated to see if there were any patterns among the variables. The figure below is an

example displaying the plot of the variable "education.num" along with income split across the several categories. One item of note is that as the number of years of education increases so does the number of people who make more than $50K per year. At about 13 years of education the number of those who make more than $50K per year begin to overtake those who do not. The later years of education are most associated with education levels of Masters, Doctorate, and professional schooling. The step increase in number of individuals in the middle of the plot is related to high school graduates.

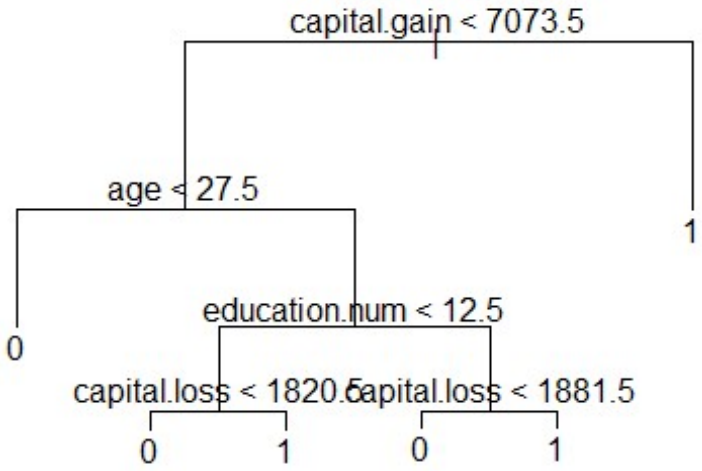


The variable "education.num" appeared to reflect the same information as the "education" variable and so it was not included as part of the model fitting process. A naive approach was used to start by including all variables in the model fitting. Most variables were considered to be statistically significant based on their individual p-values. The next step was then to calculate the predictions with the fitted model using the test data. Then a new predicted response variable, $\tilde{y}$, was calculated using a classification rule of 0.5 as a starting point from which an MSPE value of 0.15226 was calculated. The following ROC curve was generated to check if the $\hat{p}$ value seemed like a good classification rule for this model. As can be seen, 0.5 seemed like a reasonable value; however, other values were checked against the model to see if any improvement could be made. Other $\hat{p}$ values included 0.6 and 0.4, both of which did not improve upon 0.5. The model was tuned a bit more until it was discovered that 0.49 was the best classification rule based on having the lowest MSPE of 0.15172.



ROC Curve

Further investigation into reducing MSPE included using a classification tree. Starting with all the variables, the classification tree determined that the most important variables were "capital.gain," "age," "education.num," and "capital.loss" which is illustrated in the plot below. One item of note with the tree diagram is the split for years of education. The classification found a split at 12.5 years of schooling to be important which corresponds well with the figure detailing the number of individuals and years of education above. However, the misclassification error rate associated with the classification tree was 0.1914, which was higher than the aforementioned model. At this point, it was decided that the MSPE of 0.15172 using a naive approach with the logistic regression model was the lowest achievable.



Classification Tree

Conclusions/Discussion

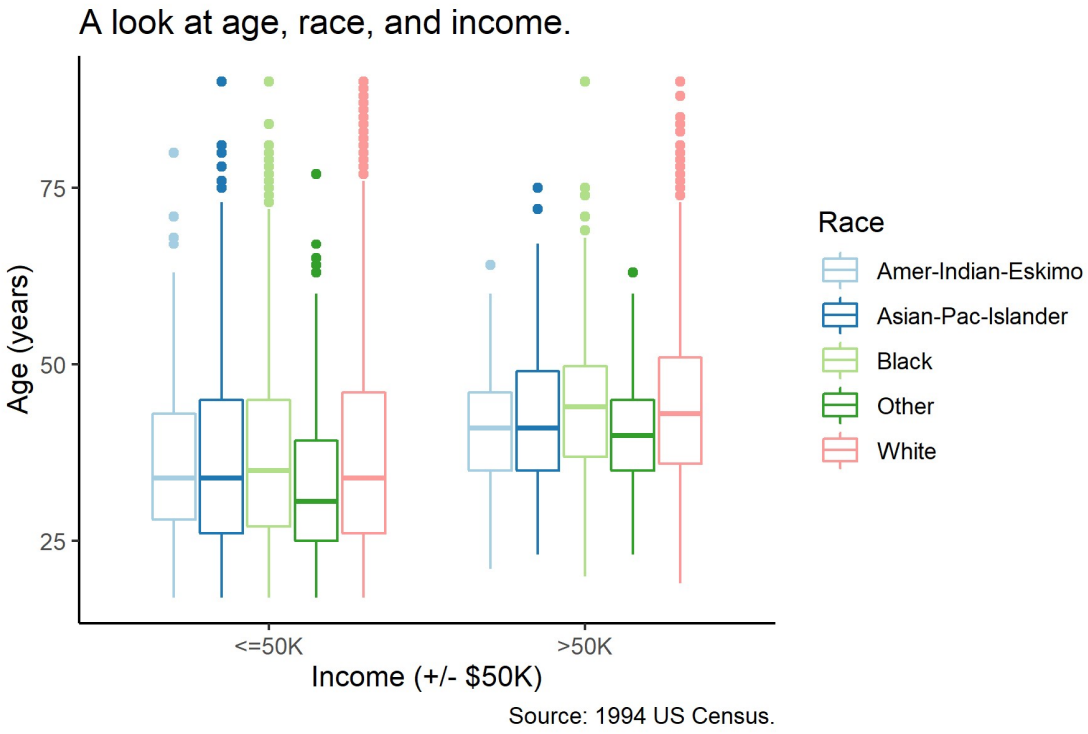
In the explanatory portion of this report, age, race, and sex were considered in terms of how they relate to an individual's income being greater than $50K. It was found that between a model with main effects only and a model with two-way interaction terms, the model with the interaction terms was the more appropriate model even though the individual terms were found to be less significant. Additionally, three-way interactions were also considered and found to be statistically insignificant compared to the two-way interactions model. This was determined by performing a drop in deviance test for all three models. Lastly, the example provided between the 30 year old white female and 30 year white male illustrated how these interaction terms can impact a model by adding or subtracting from the intercept and slope of the model.

The exploratory analysis looked at which variables were statistically significant in predicting if an individual's income was greater than $50K per year. When determining which variables seemed to be statistically significant in predicting a person's income, it was found that to a certain degree all the variables were statistically significant. However, there were certain levels for each of the categorical variables that were not statistically significant. This is believed to play a part in the MSPE value being 0.1517 given a classification value of $\hat{p} = 0.49$. While the MSPE value was not terrible, there may be ways that it can be improved upon such as combining terms, data transformations, or using quadratic terms. Interestingly

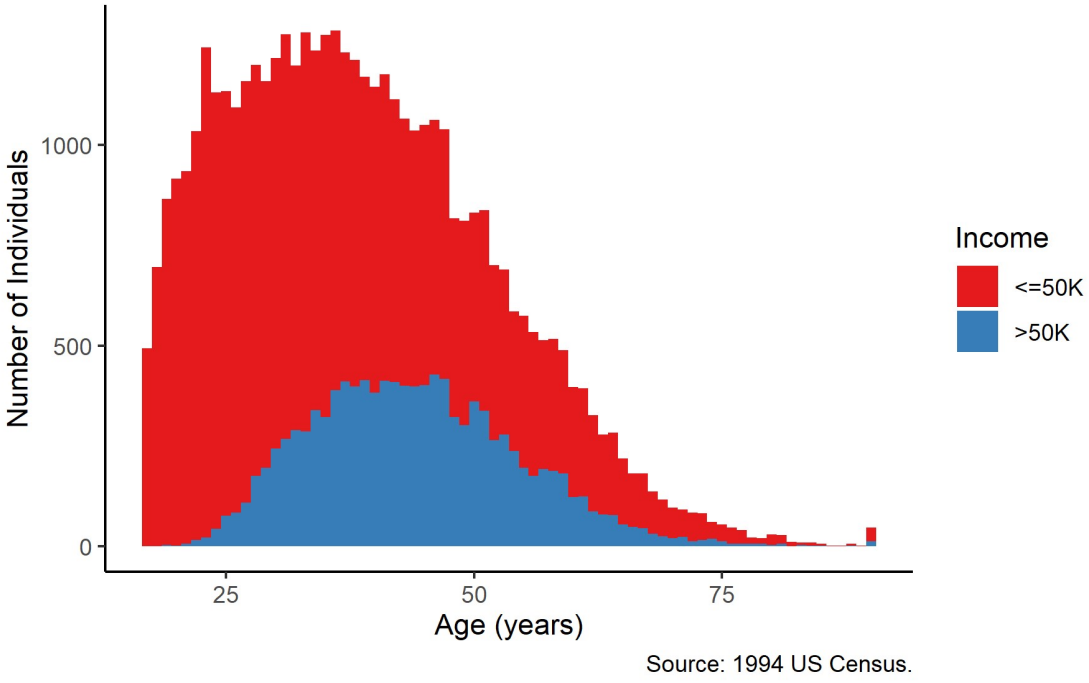
enough, even though the classification tree resulted in a higher error result, it did show that most of the predictive capability could be found with just four variables.

Appendix

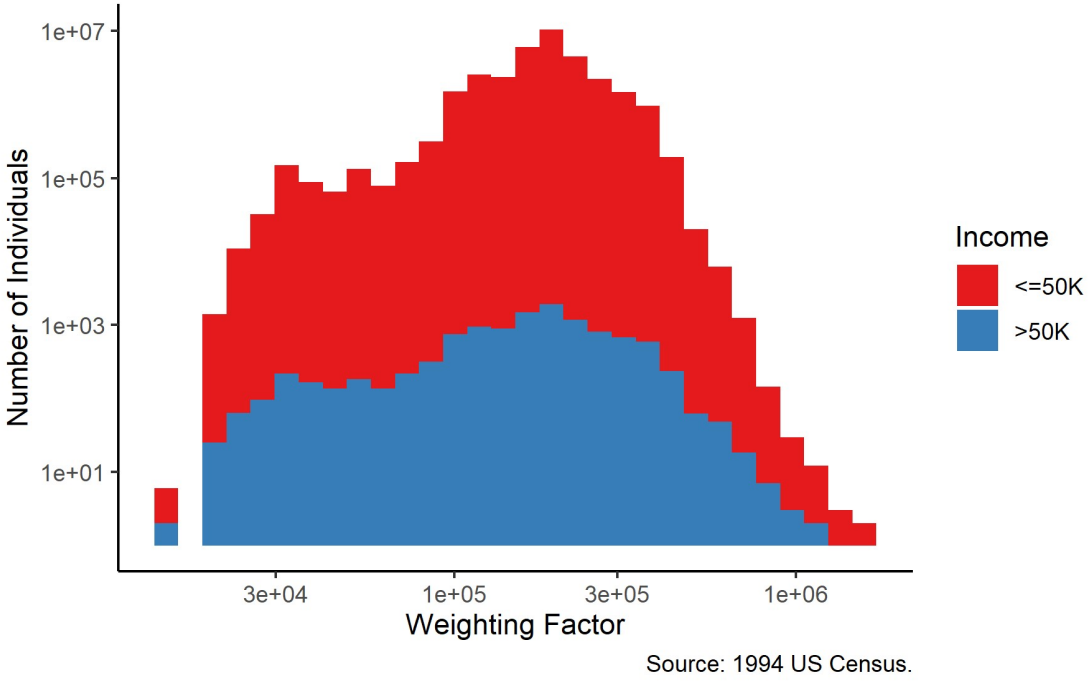
Additional Figures



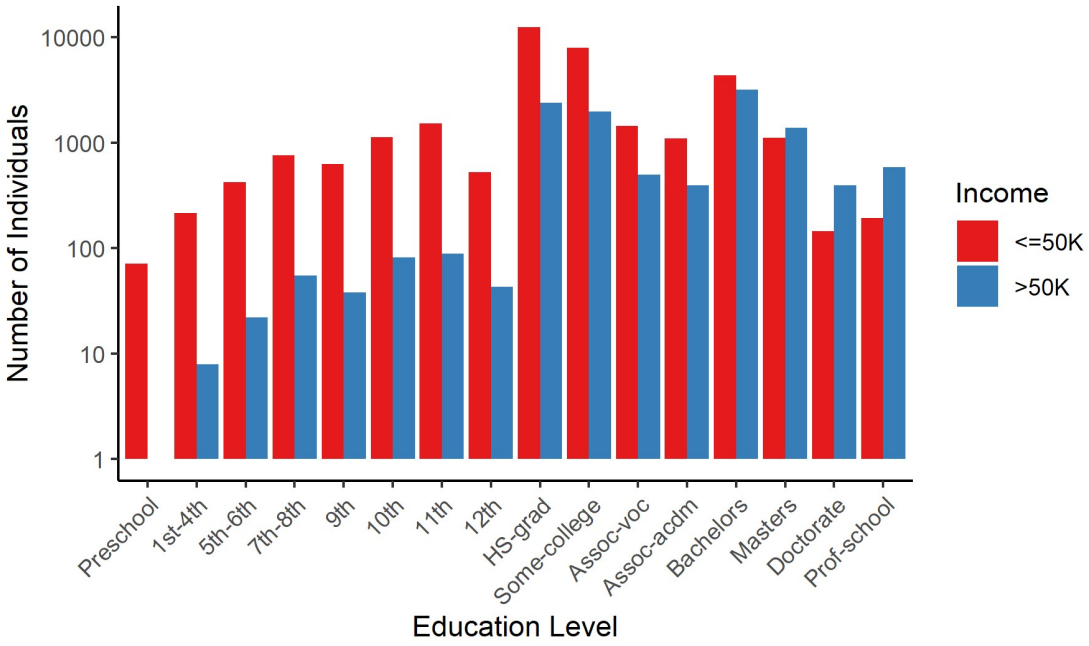
Exploratory Analysis



Source: 1994 US Census.

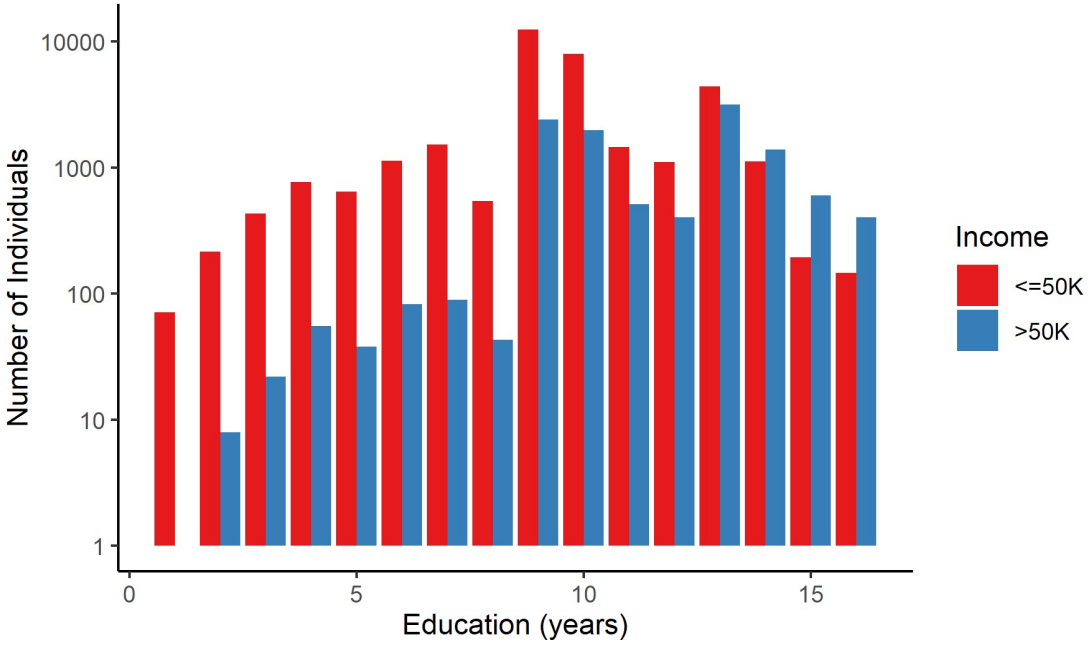


Source: 1994 US Census.

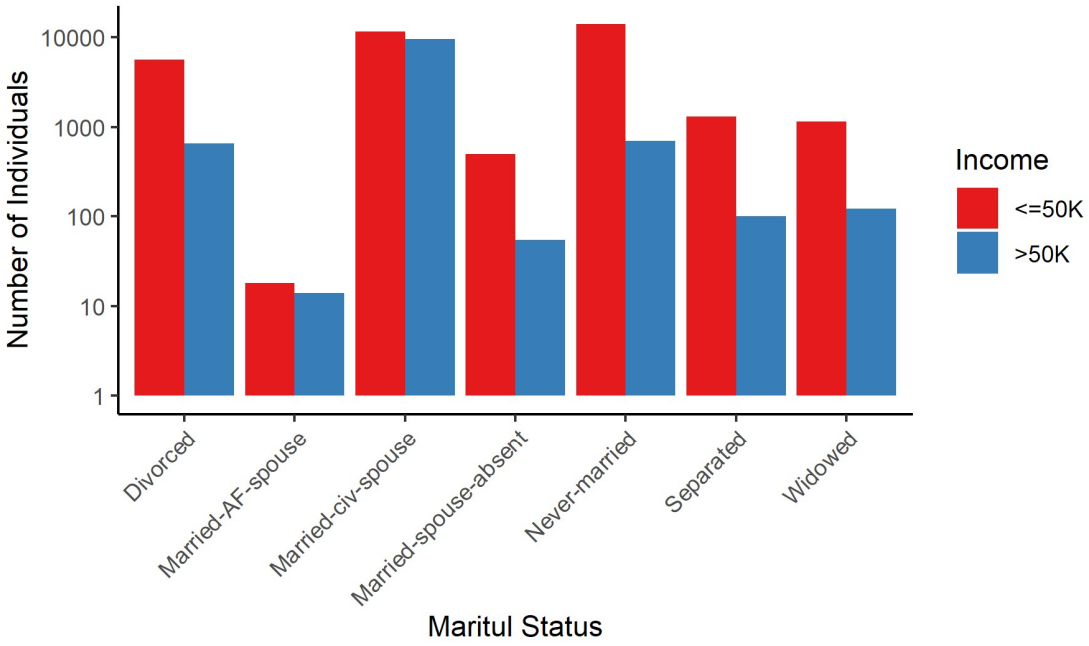


Source: 1994 US Census.

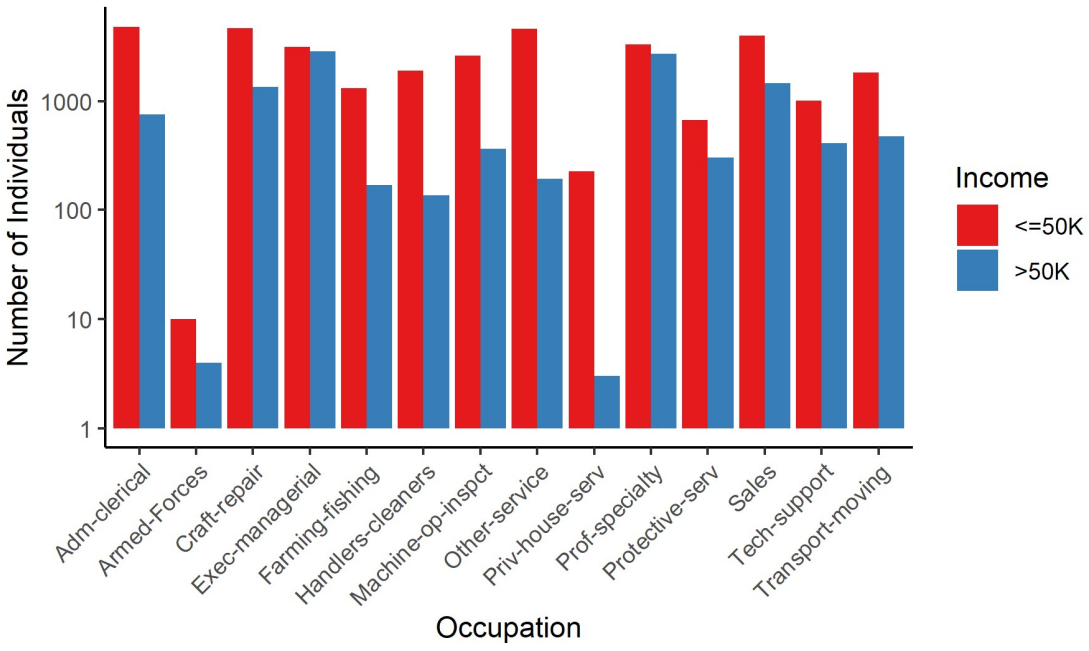
Number of people earning 50K+ increases with years educated.



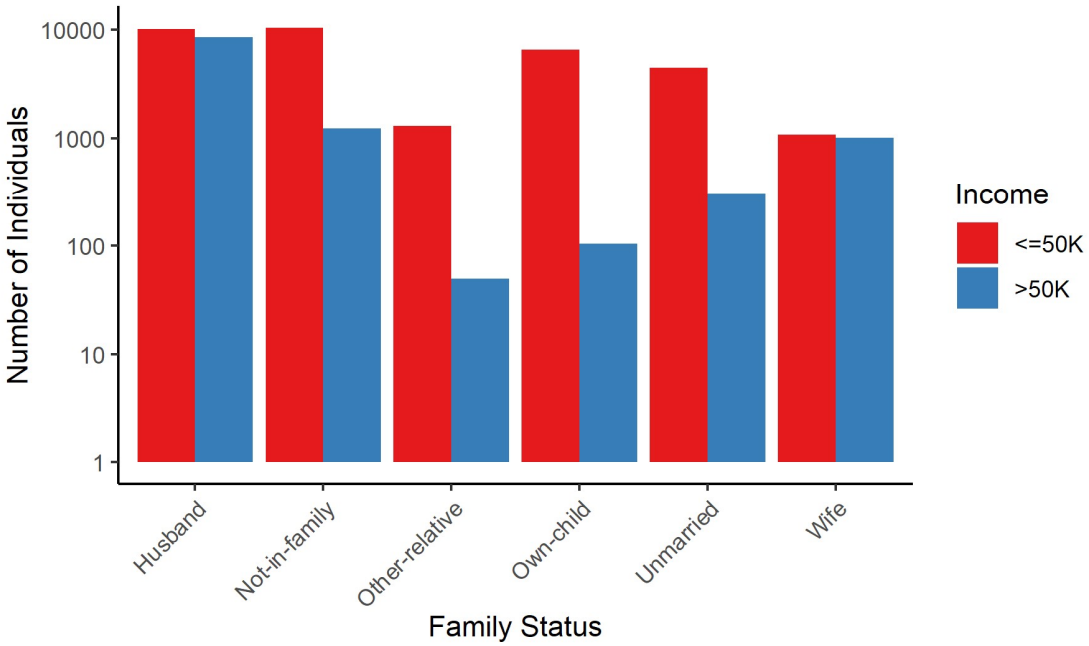
Source: 1994 US Census.



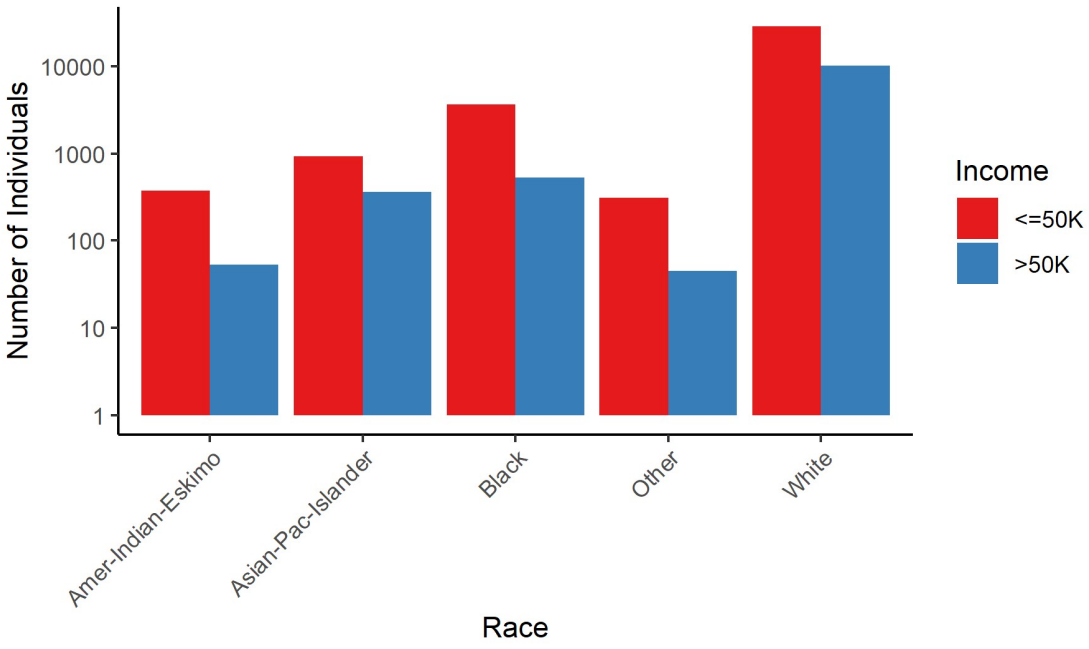
Source: 1994 US Census.



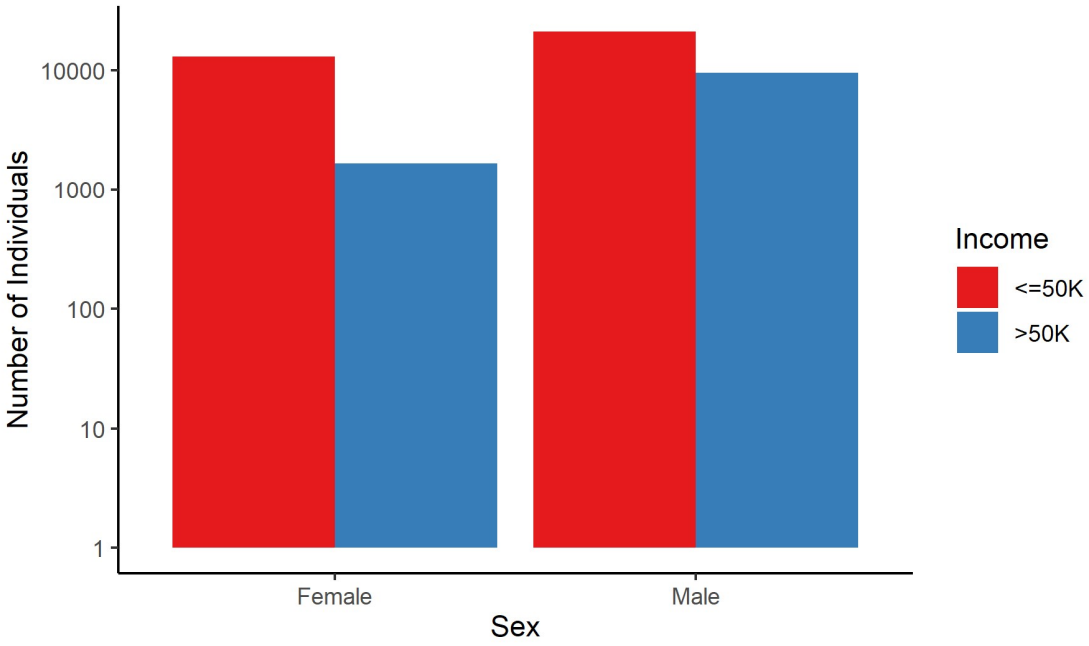
Source: 1994 US Census.



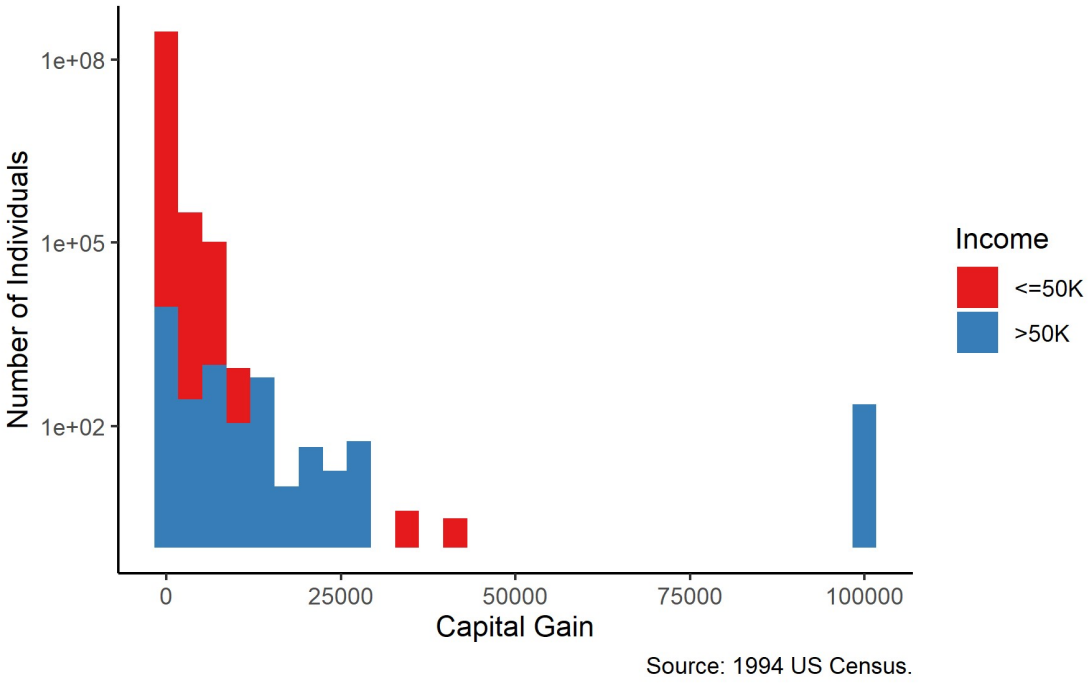
Source: 1994 US Census.



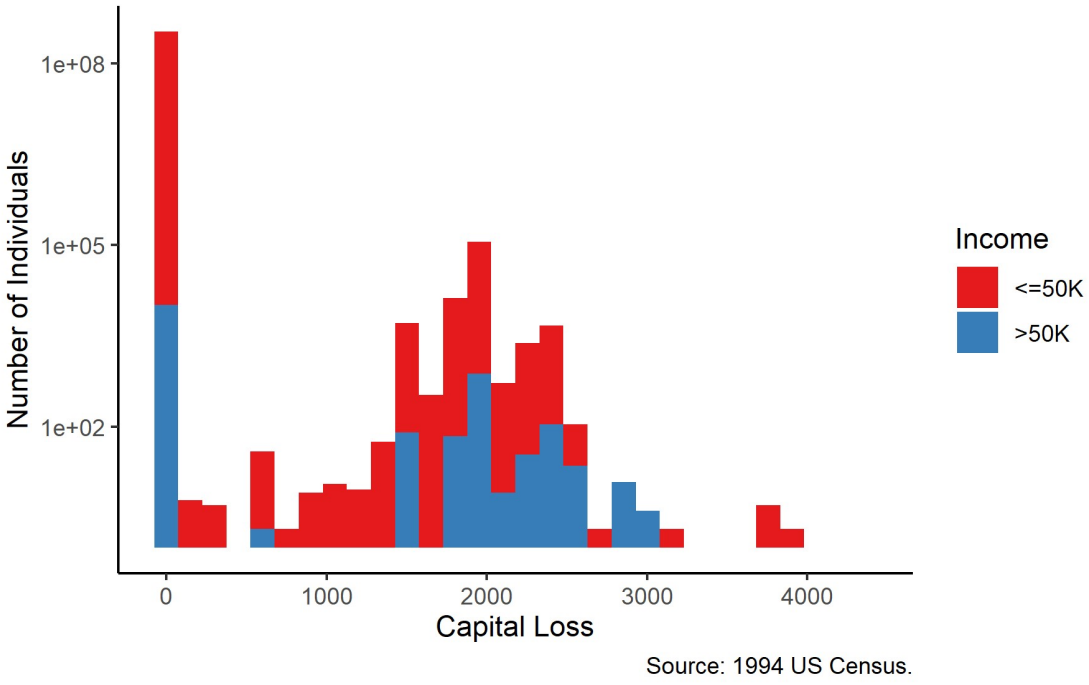
Source: 1994 US Census.



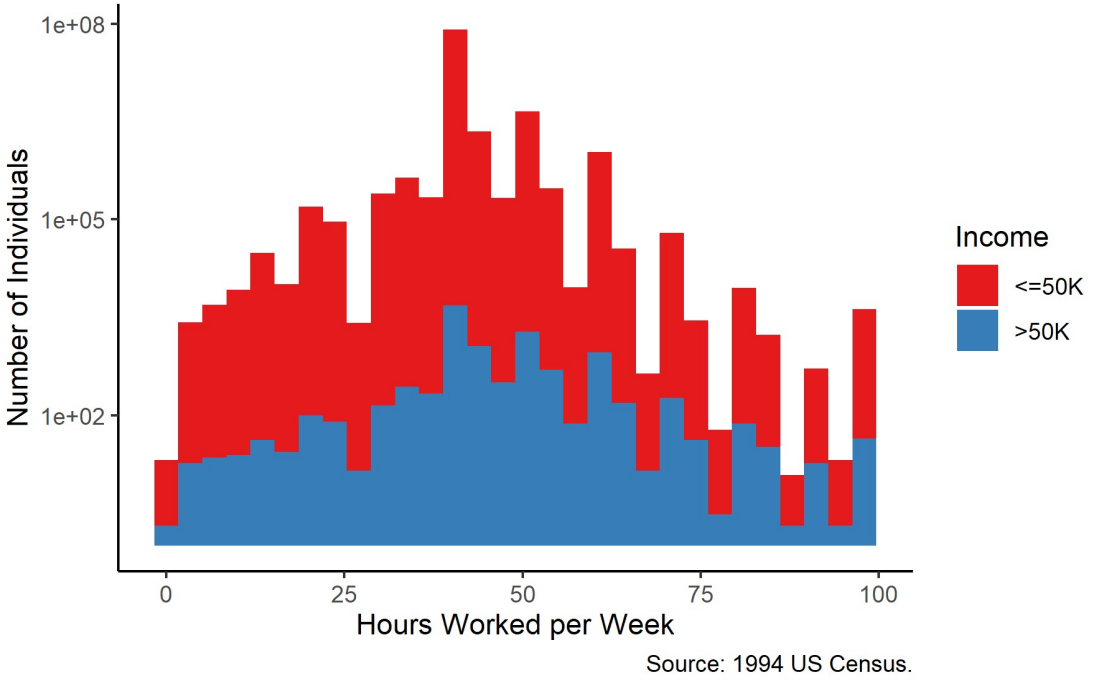
Source: 1994 US Census.



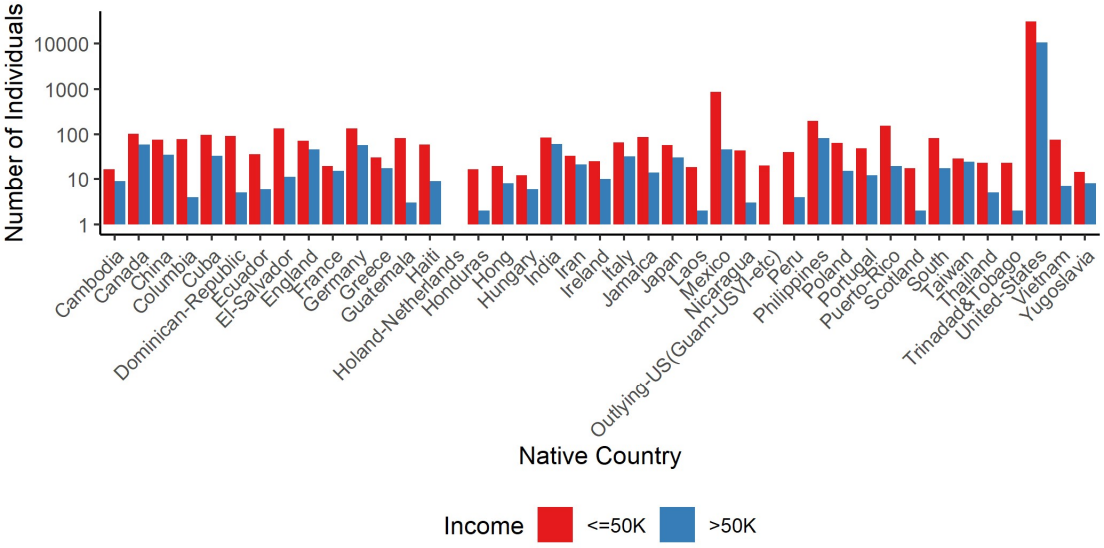
Source: 1994 US Census.



Source: 1994 US Census.



Source: 1994 US Census.



Source: 1994 US Census.

Additional Tables

Log-odds Two-way Interaction Estimates w/ 95% CI

Coefficient	Estimate	95% CI LB	95% CI UB
(Intercept)	-3.4780859026	-4.628535384	-2.32763642
age	0.0284090433	0.003441487	0.05337660
raceAsian-Pac-Islander	0.9346156000	-0.302377885	2.17160908

Coefficient	Estimate	95% CI LB	95% CI UB
raceBlack	-0.5812148535	-1.782949814	0.62052011
raceOther	-0.3063831388	-1.938263275	1.32549700
raceWhite	0.3955078520	-0.754212557	1.54522826
sexMale	0.0083368700	-0.661710395	0.67838414
age:raceAsian-Pac-Islander	-0.0068267014	-0.033608097	0.01995469
age:raceBlack	0.0041513936	-0.021743758	0.03004654
age:raceOther	0.0080829041	-0.028411404	0.04457721
age:raceWhite	0.0001059315	-0.024799900	0.02501176
age:sexMale	0.0161384958	0.012089717	0.02018728
raceAsian-Pac-Islander:sexMale	0.4135373765	-0.303138924	1.13021368
raceBlack:sexMale	0.6453381258	-0.038623534	1.32929979
raceOther:sexMale	0.1422864554	-0.875524424	1.16009734
raceWhite:sexMale	0.5240942799	-0.128778863	1.17696742

Example Problem

30 yr old white female:

$$\boldsymbol{logit(income)} = \boldsymbol{\beta}_0 + \beta_1\boldsymbol{age} + \boldsymbol{\beta}_5\boldsymbol{raceWhite} + \boldsymbol{\beta}_{11}\boldsymbol{age} * \boldsymbol{raceWhite}$$

$$\boldsymbol{logit(income)} = \boldsymbol{\beta}_0 + \beta_5 + (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{11}) * \boldsymbol{age}$$

$$\boldsymbol{exp}(\boldsymbol{-3}.47 + 0.40 + (\boldsymbol{0}.\boldsymbol{0}28 + \boldsymbol{0}.\boldsymbol{00010}) * \boldsymbol{30}) = \boldsymbol{0}.\boldsymbol{11}$$

30 yr old white male:

$$\boldsymbol{logit(income)} = \boldsymbol{\beta}_0 + \beta_1\boldsymbol{age} + \boldsymbol{\beta}_5\boldsymbol{raceWhite} + \boldsymbol{\beta}_6\boldsymbol{sexMale} + \boldsymbol{\beta}_{11}\boldsymbol{age} * \boldsymbol{raceWhite} + \boldsymbol{\beta}_{12}\boldsymbol{age} * \boldsymbol{sexMale} + \boldsymbol{\beta}_{16}\boldsymbol{raceWhite} * \boldsymbol{sexMale}$$

$$logit(income) = \beta_0 + \beta_5 + \beta_6 + \beta_{16} + (\beta_1 + \beta_{11} + \beta_{12}) * age$$

$$exp(-3.47 + 0.40 + 0.0083 + 0.52 + (0.028 + 0.00010 + 0.016) * 30) = 0.30$$

$$0.30/0.11 \approx 2.7$$