Predictors of Family Income

Brian McHatton

University of Oklahoma

Introduction

Indicators for economic prosperity or woe can vary widely across the world. For example, in the Philippines, whether a family has air conditioning or television and how many of each can be a fair estimator of that family's income (Villejo). However, In the United States, due to our higher average standard of living, things like television and air conditioning are commonplace even for families who have little income, so exploring different types of variables is needed. In this paper I will focus more on demographic variables like race, sex, and age; employment status and whether someone works multiple jobs; educational effects like years of education, sheepskin effect of degrees, type of degree attained; and calculating most of these same things with their spouses, if applicable.

Family income also has large effects on several different aspects of life, especially health outcomes educational outcomes. Specific findings include higher levels of education and college choices for students—particularly those of average intelligence (Kinsler and Pavan). Families with more money are more likely to send an average-intelligence child to a high-level university than families with lower income and a comparatively intelligent child. Higher income overall increases mental and physical health outcomes (Ettner), though there are some health drawbacks like increased alcohol consumptions.

From a finance perspective, higher family incomes also lead to the ability to save more money, make investments, and build generational wealth that can perpetuate outcomes in education and health (Source).

Because of its central role in our lives, I decided to collect data on households and predict their family income to better understand the causes of attributions to a family's economic situation.

Description of Model

In this paper, I created a model meant to predict family income based on quantitative factors like number of years of education and additionally a large number of qualitative, demographic data. All data I used was sourced from IPUMS Time Use Survey for the year 2020 (Hofferth). The variables I include in the model include household size, household number of adults, age; whether the respondent is White, Black, Asian Pacific Islander, or Native; whether the respondent has attained a bachelor's, master's, doctorate degree, or a professional license; respondents sex, years of education, and if they have multiple jobs; spousal presence, spousal sex; spouse is White, Black, Native, or Asian Pacific Islander; spousal years of education, spousal employment, spouses' usual hours worked; and if the spouse has attained an associates, bachelor's, master's, doctoral degree or a professional degree or license. I used 8782 observations of households from the survey. My model is a linear regression model in the style of $Y=\beta_0 + \beta_x x +u$, where Y is the dependent variable, β_0 is the intercept, and x represents a change in each individual variable with its magnitude of effect predicted by its accompanying β_x value. This yields the following model:

Equation 1:

$$Log(FAMINCOME) = \beta_0 + \beta_1 HH\_SIZE + \beta_2 HH\_NUMADULTS + \beta_3 AGE + \beta_4 WHITE + \beta_5 BLACK + \beta_6 API + \beta_7 NATIVE + \beta_8 SEX + \beta_9 EDUCYRS + \beta_{10} MULTJOBS + \beta_{11} SPOUSEPRES + \beta_{12} SPSEX + \beta_{13} SPWHITE + \beta_{14} SPBLACK + \beta_{15} SPNATIVE + \beta_{16} SPAPI + \beta_{17} SPEDUC + \beta_{18} SPEMPNOT + \beta_{19} SPUSUALHRS + \beta_{20} BACHELORS + \beta_{21} MASTERS + \beta_{22} PROFESSIONAL + \beta_{23} DOCTORAL + \beta_{24} SPASSOCIATES + \beta_{25} SPBACHELORS + \beta_{26} SPMASTERS + \beta_{27} SPPROFESSIONAL + \beta_{28} SPDOCTORAL + u$$

The following table gives the name and a description of the units of the variables used within the regression model:

Table 1

Variable	Description	Variable	Description
FAMINCOME	Household's income in dollars	SPNATIVE	Dummy for Native spouse
HH_SIZE	Number of people in household	SPAPI	Dummy for Asian Pacific Islander
HH_NUMADULTS	Number of adults in household	SPEDUC	Spousal education in years
AGE	Age in years	SPEMPNOT	Dummy for spousal employment

WHITE	Dummy variable for white	SPUSUALHRS	Spouse's usual hours worked in a week
BLACK	Dummy variable for Black	BACHELORS	Dummy variable for bachelor's degree
NATIVE	Dummy variable for Native	MASTERS	Dummy variable for master's degree
API	Dummy variable for Asian Pacific Islander	PROFESSIONAL	Dummy variable for professional license
SEX	Dummy variable for biological sex	DOCTORAL	Dummy variable for doctorate
EDUCYRS	Years of education	SPASSOCIATES	Dummy for spouse's associates
MULTJOBS	Dummy for multiple jobs	SPBACHELORS	Dummy for spouse's bachelor's degree

SPOUSEPRES	Dummy for spousal presence	SPMASTERS	Dummy for spouse's master's degree
SPSEX	Dummy for spouse's biological sex	SPPROFESSIONAL	Dummy for spouse's professional degree
SPWHITE	Dummy for White spouse	SPDOCTORAL	Dummy for spouse's doctorate
SPBLACK	Dummy for Black spouse		

The values collected from the data were only for the year 2020.

FAMINCOME was initially collected in ranges of income and had to be mutated to a continuous variable, so I took the midpoint of each range which yielded the approximate family income for that range. The ranges varied from ranges of just $2,500 difference, to $49,999, and finally, the last range was anything above $150,000, which the purposes of this study have been made to the range's low of $150,000. I chose to take the log of family income, because the qualitative data would give an average percentage change given the demographics or educational attainment of people in a household, as well as evaluate changes in income given changes in continuous variables like

EDUCYRS. Furthermore, income is naturally non-negative. This being said, the model could potentially be regressed without natural log of *FAMINCOME* in order to see the average dollar change in income.

Dummy variables are given the value of either 0 or 1. A ranking of 1 for all dummies except for *SEX* and *SPSEX* indicate that the dummy is satisfied under its condition and represents the qualitative attribute associated with it. For *SEX* a value of 1 indicates that the respondent is male and a 0 indicates the respondent is female. *SPSEX*'s 1 value indicates a male and 0 indicates a female.

The degree dummies for both respondent and their partner of spouse was included along with years of education to identify any sheepskin effects. The number of years to complete graduate school degrees or professional licenses can vary from person to person, so I used an average of 14 years for associate degree, 16 years for bachelor's degree, 18 years for master's degree, 20 years for a professional degree or license, and 22 years for a doctorate. The degree dummies only account for the highest degree attained and does not take into account prerequisite degrees (e.g. an individual with a bachelor's and a master's degree is only considered categorically to have a master's for the purposes of the dummies).

Hispanic and mixed races serve as the default to a respondent not identifying as any of White, Black, Native, or Asian Pacific Islander.

API includes two groups: Asian only and Hawaiian Pacific Islander only, but does not account for Asian-Hawaiian mixed.

EDUCYRS was also put into non continuous categories. Most notable was early schooling in which first grade to eighth grade education being the highest level achieved was one or two categorizations of data. I chose to average this to roughly five years of schooling for the purposes of making the data continuous and easier to understand.

I hypothesize that education, between the respondent and their partner, will be the primary driver of what controls family income. I think an ideal, secondary driver would have been workforce experience, however the data did not yield any findings on that, but *AGE* could be used as a potential proxy. Another variable that could have large effect on family income would be innate ability of the adults in the household, but there is not any data in the data set for that, nor would it be easy to unbiasedly collect said data in a self-reported survey.

Aside from educational status and achievement, *HH_NUMADULTS* and *SPEMPNOT* are pretty obvious choices for variables to have a leading effect on family income, because in order to work in the United States you have to be a certain age. This can vary from state to state; however, every adult is of working age around the country. Therefore, the more adults in a household, the more people can legally work and contribute to the family income. Going hand-in-hand with this idea is spousal employment. If the respondent's spouse is employed, the household will be making more money.

Interpretation of Estimated Variables and Model

The following section displays my regression on family income. Family income is measured in natural log dollars, therefore assume a one unit change in the respective x of a variable leads to an estimated percent change in dollar amount change in family income equal to the variable's β. Values are the estimated β's for their respective variable as well as their standard error:

	Model 1
(Intercept)	8.219**
	(2.836)
HH_SIZE	0.005
	(0.008)
HH_NUMADULTS	0.184***
	(0.013)
AGE	-0.005***
	(0.000)
WHITE	0.036
	(0.061)
BLACK	-0.231***
	(0.065)
API	0.080
	(0.073)
NATIVE	-0.246*
	(0.098)
SEX	0.113***
	(0.021)
EDUCYRS	0.064***

	Model 1
	(0.004)
MULTJOBS	0.053
	(0.033)
SPOUSEPRES	1.390
	(2.808)
SPSEX	0.017
	(0.028)
SPWHITE	0.060
	(0.080)
SPBLACK	0.124
	(0.091)
SPAPI	-0.082
	(0.095)
SPEDUC	0.000*
	(0.000)
SPEMPNOT	0.236***
	(0.022)
SPUSUALHRS	0.000
	(0.000)
BACHELORS	0.143***
	(0.024)
MASTERS	0.174***
	(0.035)
PROFESSIONAL	0.148*
	(0.068)

	Model 1
DOCTORAL	0.004
	(0.064)
SPASSOCIATES	0.163***
	(0.034)
SPBACHELORS	0.243***
	(0.026)
SPMASTERS	0.286***
	(0.033)
SPPROFESSIONAL	0.265***
	(0.068)
SPDOCTORAL	0.283***
	(0.058)
Num.Obs.	8782
R2	0.344
R2 Adj.	0.342
AIC	17966.5
BIC	18171.9
Log.Lik.	-8954.274
F	170.288

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Table 2

Of the variables used to regress family income, *NATIVE* and *PROFESSIONAL* are significant at the .05 level; the intercept and *MASTERS* variables are significant at the .01 level; and *HH_NUMADULTS, SEX, EDUCYRS,*

SPOUSEPRES, SPEMPNOT, BACHELORS, SPMASTERS, SPPROFESSIONAL, and *SPDOCTORAL* are all significant at the .001 level. All other variables failed to be significant.

As I had predicted, the number of adults in a household, spousal employment, years of education, and the type of degree achieved were key forces in predicting (at least to some extent) family income. One thing that needs explaining however is the variable *DOCTORAL* has a statistically insignificant effect on family income. Meanwhile, *SPDOCTORAL* had statistically significant effect on family income at the .001 level. This leads me to believe there is either something wrong with the data that I acquired or amputated, or some type of bias that would make it where certain people in a household were more likely to fill out the form. This could also be suggested by the difference in significance between spousal professional licenses and degrees (*SPPROFESSIONAL*) and respondent's professional degree and licenses (*PROFESSIONAL).*

The degree dummy variables decreased the estimated yearly gains of education while also identifying a possible sheep skin effect. All degrees, except for the aforementioned *DOCTORAL* variable were significant. Without the dummies, I would expect to see the estimated value of an additional year of education be upwardly biased.

Of the demographic variables, *NATIVE* and *BLACK* the only racial demographic that are significant. Furthermore, it has a large, negative β indicates, per the regression, some type of systemic cause for Blacks, Native

Americans, and Native Alaskans having lower family incomes than other races. *SEX* and *SPSEX* vary in significance. *SEX* is significant at .001 level while *SPSEX* is insignificant.

A few variables got zeroed-out due to some respondents not having spouses or partners or ones that contributed to their household income. These variables were *SPUSUALHRS* and *SPEDUC.* I ran limited models regressing each one individually on log of *FAMINCOME.* Doing this, I removed some of the observations and only included families where the respondents had a spouse or partner. The limited regression for *SPEDUC* on **Table 3** includes 4715 observations, and the limited regression for *SPUSUALHRS* on **Table 4** includes 4767 observations. All other limited models will contain 8782 observations.

	Model 1
(Intercept)	9.902***
	(0.040)
SPEDUC	0.092***
	(0.003)
Num.Obs.	4715
R2	0.201
R2 Adj.	0.201
AIC	8669.2
BIC	8688.6
Log.Lik.	-4331.601
F	1188.639

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Table 3

After running the limited model for *SPEDUC*, the model showed that spousal education was significant at the .001 level. Its beta value is greater than the unlimited model's beta value of *EDUC,* but I believe it is upwardly biased due to lack of spousal degree dummies in the model.

	Model 1
(Intercept)	11.012***
	(0.014)
SPUSUALHRS	0.009***
	(0.000)
Num.Obs.	4767
R2	0.080
R2 Adj.	0.080
AIC	9489.7
BIC	9509.1
Log.Lik.	-4741.840
F	415.899

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Table 4

One of the surprising results of the unlimited model in **Table 2** was that *AGE* had a significant, negative effect on *FAMINCOME* at a .001 significance level. I would have assumed that as *AGE* increased in value, there would be on average a positive change in *FAMINCOME*. The reason I hypothesized this was because *AGE* could serve as proxy for experience in the workforce. In **Table 5**, I regressed *FAMINCOME* on *AGE* to see if there were any changes in the results.

	Model 1
(Intercept)	11.379***
	(0.026)
AGE	-0.008***
	(0.000)
Num.Obs.	8782
R2	0.033
R2 Adj.	0.033
AIC	21325.2
BIC	21346.5
Log.Lik.	-10659.612
F	301.550

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Table 5

After running the regression in **Table 5**, there was no change in sign, nor was there a great change in a β value. Furthermore, *AGE* remained significant at a .001 level. This could be due to people who are retired having little personal income and living off of assets or could indicate that prime earing years occur early in an individual's life, on average.

Another surprising result of the unlimited model was that *DOCTORAL* was insignificant. This surprise was two-fold. Firstly, *DOCTORAL* represents the highest order degree of the dummies included in the model, and it is represented by the most years of education. Secondly, its spusal counterpart, *SPDOCTORAL*, is significant at the .001 significance level. In **Table 6** I regressed *FAMINCOME* on *DOCTORAL.*

	Model 1
(Intercept)	10.942***
	(0.009)
DOCTORAL	0.580***
	(0.057)
Num.Obs.	8782
R2	0.012
R2 Adj.	0.012
AIC	21517.3
BIC	21538.5
Log.Lik.	-10755.629
F	105.121

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Table 6

The limited model depicted in **Table 6** counters the limit model's estimation of *DOCTORAL*'s significance on *FAMINCOME.* The limited model's estimation was in agreeance with my hypothesis. One reason that it may show as insignificant in the unlimited model is that the averages used for *FAMINCOME* is capped at $150,000 in the top income group. If the values for *FAMINCOME* were truly continuous and not an average or estimation, I imagine that the effects of *DOCTORAL* would be better seen and could possibly make the findings in the unlimited model significant.

Policy and Economic Implications

The most outstanding findings from my regression were that the education is a huge part of how much income a family can earn—not only on the return of years of schooling but also attaining a degree. This agrees with common literature that the gains of education are significant across the world (Harmon, Oosterbeek, Walker). This should only incentivize policymakers to support educational infrastructure further and incentivize people to attain higher levels of education.

Aside from education however, according to the regression, people who identify as Native and Black suffer from lower family incomes on average. This is cause for concern, because it may have either systemic problems for Blacks and Natives or there is discrimination against them which suppresses their family income on average. Policymakers should investigate the difference in expected income and find the cause for the significant difference in family income.

Summary, Conclusion, and Future Research

I created a linear regression which estimated the dollar effects of surveyed families in the United States based on their family income based on their educational levels and demographics of the household.

While there are some no-brainer variables that cause family income on average to increase, it is worth noting that the effects of education and varying types of educational attainment and accomplishments of academic degrees. With the exception of Natives and Blacks, who, on average, have a family income several percentage points lower than the other races.

I would be interested in knowing if any of these numbers looked different in another year before the COVID-19 Pandemic. Wages could have changed and affected family income, others could be unaffected or even increase amount of income depending on their job, so I think it would be interesting to investigate other years and see if the β's vary to any significant level.

Appendix:

This is a list of variables included in my regression with their casual name listed first, as it would appear in text or conversation, followed by its name as a variable in the regression. Variables are listed in alphabetical order of their casual name.

Age (AGE): The age, in years, of the respondent to the survey.

Asian Pacific Islander (API): Dummy variable determining whether the respondent identifies as either Asian or Hawaiian Pacific Islander and no other race. This variable in the data set was initially two separate variables but has been combined into one demographic group. It does not include respondents who identified as Asian-Hawaiian Islander mixed.

Bachelor's Degree (BACHELORS): Dummy variable determining whether respondent earned a bachelor's degree as their highest form of education.

Black (BLACK): Dummy variable determining whether the respondent identifies as Black and no other race.

Doctoral Degree (DOCTORAL): Dummy variable determining whether respondent earned a doctorate as their highest form of education.

Family Income (FAMINCOME): amount of money in dollars a family earns. Initial data was collected in groups and not as continuous values, therefore data has been mutated by taking the midpoints of each income range and using it as the average to make a continuous variable.

Household number of adults (HH_NUMADULTS): the number of adults in a household—people who are 18 years or older.

Household Size (HH_SIZE): the number of members in the household, regardless of age or relation.

Master's Degree (MASTERS) Dummy variable determining whether respondent earned a master's degree as their highest form of education.

Multiple Jobs (MULTJOBS): Dummy variable for whether respondent has worked more than one job within the past week.

Native (NATIVE): Dummy variable for whether the respondent identifies as Native American or Alaskan Native and no other race.

Professional License (PROFESSIONAL): Dummy variable determining whether respondent earned a professional license as their highest form of education (e.g. lawyer).

Sex (SEX): Dummy variable that indicates the respondent's biological sex.

Spousal Associate's (SPASSOCIATES): Dummy variable determining whether respondent's spouse or partner earned an associate's degree as their highest form of education.

Spousal Bachelor's (SPBACHELORS): Dummy variable determining whether respondent's spouse or partner earned a bachelor's degree as their highest form of education.

Spousal Doctorate (SPDOCTORAL): Dummy variable determining whether respondent's spouse or partner earned a doctorate as their highest form of education.

Spousal Education (SPEDUC): highest level of education the respondent's spouse or partner completed, corrected for number of years.

Spousal Employment (SPEMPNOT): Dummy variable for whether respondent's spouse or partner is employed.

Spousal Master's (SPMASTERS): Dummy variable determining whether respondent's spouse or partner earned a master's degree as their highest form of education.

Spousal Presence (SPOUSEPRES): Dummy variable for whether respondent's spouse or partner is within the household.

Spousal Professional License (SPPROFESSIONAL): Dummy variable determining whether respondent's spouse or partner earned a professional license as their highest form of education.

Spousal Sex (SPSEX): Dummy variable for whether respondent's spouse or partner is biologically male or female.

Spouse is Asian Pacific Islander (SPAPI): Dummy variable for whether the respondent's spouse or partner identifies as either Asian or Pacific Islander

Spouse is Black (SPBLACK): Dummy variable for whether the respondent's spouse or partner identifies as Black and only Black.

Spouse is Native (SPNATIVE): Dummy variable for whether the respondent's spouse or partner identifies as Native American or Alaskan Native and only Native American or Alaskan Native.

Spouse's Usual Hours (SPUSUALHRS): Continuous variable reporting how many hours the respondent's spouse or partner usually work per a week, excluding those with highly variable hours.

Spouse is White (SPWHITE): Dummy variable for whether the respondent's spouse or partner identifies as white and only white.

White (WHITE): Dummy variable for whether the respondent identifies as White and no other race.

Years of Education (EDUCYRS): This continuous variable lists the number of years the respondent has completed. This does not account for different type of degrees attained by the individual.

Work Cited

Ettner, Susan L. New evidence on the relationship between income and health, Journal of Health Economics, Volume 15, Issue 1, 1996, Pages 67-85, ISSN 0167-6296, https://doi.org/10.1016/0167-6296(95)00032-1.

Harmon, Colm, Oosterbeek, Hessel, Walker, Ian. *The Returns to Education: Microeconomics.* K+Journal of Economic Surveys, Volume 17, Issue 2, 2003, Pages 115-156, https://doi.org/10.1111/1467-6419.00191

Hofferth, Sandra L. Flood, Sarah M. Sobek, Matthew and Backman, Daniel . American Time Use Survey Data Extract Builder: Version 2.8 [dataset]. College Park, MD: University of Maryland and Minneapolis, MN: IPUMS, 2020.

Kinsler, Josh and Pavan, Ronni. Family Income and Higher Education Choices: The Importance of Accounting for College Quality. Journal of Human Capital 2011 5:4, 453-477.

Villejo, S. J., Enriquez, M. T., Melendres, M. J., Tan, D. E., & Cayton, P. J. (2014). Determinants of income class in Philippine households: Evidence from the family income and expenditure survey 2009. *The Philippine Statistician*, *63*(2), 87-104.