Summer Program for Undergraduate Research (SPUR)

Wharton Undergraduate Research

9-12-2021

Econometric Analysis of Labor Income and Job Seeking Disparities in the United States

Shaolong "Lorry" Wu
University of Pennsylvania

Econometric Analysis of Labor Income and Job Seeking Disparities in the United States

Abstract

In the labor market of the United States, a wide range of socioeconomic and demographic factors impact workers' income and decisions to seek new jobs, which are two critical metrics of labor income dynamics. Studies of income dynamics have historically been examining major demographic and situational factors such as marital and family matters of Americans from the 1960s to 2000s. However, technological breakthroughs have drastically changed the landscape of the labor force and economy, and individuals face a more complex and diverse context. This paper recognizes the need to analyze the factors behind these two income dynamics metrics in the contemporary setting. This paper confirms that the wages and other types of income of the cohort (born in 1980s and 1990s) in the United States are explained by personal demographics, living habits, and family background conditions. This paper also finds a series of factors, such as region, marital status, and education, to be significant in determining whether individuals will seek a new full-time job over time. This paper limits itself to predictive and forensic analysis and leaves the question of job search motivations to authors of related areas.

Keywords

wage, job search, demographic, longitudinal

Disciplines

Applied Statistics | Business | Physical Sciences and Mathematics | Statistics and Probability

Wharton Summer Program of Undergraduate Research Paper:
Shaolong “Lorry” Wu
Advisor: Professor Paul Shaman, Department of Statistics and Data Science
09/12/2021

Table of Contents

[Summary]

Wage and job seeking behaviors have been critical measures for income dynamics. Technological breakthroughs have drastically changed the landscape of the labor force and economy, causing a new need to study the income dynamics. Previous research of individuals has been too general for the population. But for the people born in the 1980s and 1990s, they face completely different situations with marital and family matters. Given the substantial change and labor situations, this study aims to study labor income dynamics and job search with more modern context. This paper confirms the previous intuition that the wages of the younger cohort (born in 1980s and 1990s) in the United States are explained by personal demographics, living habits, and family background conditions. This paper also finds that for this cohort, region, marital status, and education play significant roles in determining labor wage and probabilities to engage in job search. Job seeking probabilities can be predicted at considerably good confidence by cross sectional and longitudinal models. This paper, however, does not discover the motivations behind job search, but rather limits itself to predictive analysis.

[I. Literature Review & Background Summary]

This paper pertains to several literatures. From the 1980s to the 2000s, the work force has become more gender-balanced with higher wages and participation rates of women (Blau 2007). Wage rates, household income, and job seeking behaviors have become critical aspects that characterize American workers' income dynamics. What demographic, personal behavioral, and situational variables determine the labor income and job seeking behaviors? This question has been discussed for a long time.

Many factors that may influence labor income have been examined. Betts 1995 finds the positive impact of parents' education and the child's schooling conditions on the future income of the child. Bradley 2002 uses the methodology of probit and OLS regression to study the impact of age, marital status, education, and health conditions on women's labor supply. Cristia 2008 studies the impact of first child on female labor. Genadek 2007 studies the influence of no-fault divorce laws on women's labor. The data used for these papers' modeling are all sample data sets at certain case studies, instead of the meta-data set for the entire United States. Song 2011 studies the labor market outcomes of having a GED.

In terms of job seeking behaviors, Wanberg 1996 establishes a longitudinal model of the demographic, personal, and situational variables predictive of job-seeking behavior, and Wanberg 1999 suggests that the two prominent motive determinants of job-search intensity are commitment to work and financial needs. Creed 2009 furthers that the motivations of job seeking behaviors are self-regulation and seeking reemployment.

This paper will seek to build more holistic regression models that predict labor income and job seeking with more modern data sets.

[II. Motivation]

Mincer 1995, Heckman 1980, and Mroz 1987 established this field of labor income and supply studies with a focus on married women in heterosexual binary families. However, the changes in social values, family structures, and labor market situations have changed the original consideration vastly. For example, same sex marriage and single families have flourished, women's status has significantly improved from her husband's vassal to independent, and the labor force is getting more educated.

The previous research that studies one unique aspect of labor income factors has its limitations. The effect of education should be studied with more categories instead of a binary of whether completed GED (Song 2011), given the increasing percentage of people getting high school education and more (the percentage is 77.2% in 2011, 78.0% in 2013, 78.3% in 2015, 83.2% in 2017). Therefore, the research of demographic, personal behavioral, and situational factors that drive behind labor income needs to be examined by a newer approach with more modern data sets.

[III. Raw Data Description]

The data set used in this study is National Longitudinal Surveys of Youth (NLSY) of Bureau of Labor Statistics. (https://www.bls.gov/nls/home.htm)

The NLSY97[1] consists of a nationally representative sample of 8,984 men and women born during the years 1980 through 1984 and living in the United States at the time of the initial survey in 1997. Participants were ages 12 to 16 as of December 31, 1996. Interviews were conducted annually from 1997 to 2011 and biennially since then. The ongoing cohort has been surveyed 18 times as of date. Data are available from Round 1 (1997-98) through Round 18 (2017-18).

The major sections of survey questions (variables) include Education, Training & Achievement Scores, Employment, Household, Geography & Contextual Variables, Parents, Family Process & Childhood, Dating, Marriage & Cohabitation; Sexual Activity, Pregnancy & Fertility; Children, Income, Assets & Program Participation, Health: Conditions & Practices, Attitudes, Expectations, Non-Cognitive Tests, Activities and Crime & Substance Use.

There are several problems that make working with the data set particularly difficult:

1. Response bias. For certain variables, such as spouse income in 2017, approximately 90% of the data sets are missing due to either justified survey skipping or unjustified skipping. It is possible that the people who are poorer are not willing to respond. Thus, the discussion and analysis we carry out are limited to the people who reported relatively consistently throughout the span of the NLSY survey.

2. Selection Bias. Respondents may respond just to the questions which they feel more confident to answer.

3. Missing data. Some of the questions are asked again and again across years and there are a certain number of data points skipped or missed (up to 30% in some variables, which we exclude from our model). This unbalance in the data set poses difficulty to classical time series models. Also, it appears evident that the missing data are not missing at random.

[IV. Processed Data Sets]

The bulk of my analysis is based on four data sets named as income_factors_2011, income_factors_2013, income_factors_2015, income_factors_2017, which are the data frame for variables surveyed in the year and selected from NLSY97 rounds 1-18. The data is made available at https://github.com/Shaolong-Lorry-Wu/NLSY97_summer21.

[1] The introductions of the data set are quoted from the official website of National Longitudinal Survey of Youth. https://www.nlsinfo.org/

The major factors that are used from my data sets: sex, spouse income in the previous year, household gross income in the previous year, hourly wage rates in the previous year, region, whether job seeking in the past three months, highest degree obtained, unemployment status (numerical, measured as the number of weeks unemployed in the last year), substance usage status (smoking and marijuana), household size, income to poverty ratios.

Table 1.1 Mean Sample Characteristics of income_factors[2] (sub data set of NLSY)

Mean of numerical variables	2011	2013	2015	2017
Hourly wage rates of primary job	17.930	21.330	-[3]	25.220
Spouse income last year	37929.000	43310.000	47327.000	53567.000
Household gross income last year	63160.000	67521.000	73962.000	82994.000
Household size	3.224	3.333	3.392	3.460
Unemployed weeks last year	1.358	2.515	1.936	1.358
Marijuana/smoke usage percentage	0.1695149(marijuana)	0.2391(smoke)	0.163(marijuana)	-[4]

Table 1.2 Distribution of categorical variables of income_factors[5] (sub data set of NLSY)

[2] There are multiple skips and invalid in the data entries: refusal (-1), don't know (-2), invalid skip (-3), valid skip (-4), non-interview (-5). Here, the sample means refer to the mean of the sample when these are excluded.

[3] The 2015 data set does not have hourly wage. For regression models, we replace it with the individual's total income that year.

[4] The 2017 data set does not have marijuana usage variable.

[5] There are multiple skips and invalid in the data entries: refusal (-1), don't know (-2), invalid skip (-3), valid skip (-4), non-interview (-5). Here, the sample means refer to the mean of the sample when these are excluded.

Distribution of categorical variables	2011	2013	2015	2017
Internet usage frequency	3754(several times a day), 733(once a day), (less than once a day)	5201(several times a day), 723(once a day), 1091(less than once a day)	4513(several times a day),922(once a day,1844(less than once a day)	-
Marital status	4144(never married), 2653(married), 114(separated), 419(divorced), 11(widowed)	3522(never married), 2934(married), 84(separated), 546(divorced), 16(widowed)	4144(never married),2653(married),114(separated),479(divorced),11(widowed)	2766(never married),3066(married),154(separated),663(divorced),23(widowed)
Region	914(Northeast), 1246(South Central), 2296(South), 1337(Southwest)	1078(Northeast),1487(South Central),2913(South),1594(Southwest)	1140(Northeast),1515(South Central),3020(South),1682(Southwest)	1023(Northeast),1368(South Central),2770(South),1499(Southwest)

[V. Research Question/Problem of Interest]

The major research interest is to explore the factors that impact the income and job seeking status of a representative cohort who were born in the 1980s in the United States, who are the mainstream of the American labor force. This research does not aim to study the problem systematically for all age groups, because such a kind of meta-analysis is not economic and does not recognize the difference between generations caused by changes in family upbringing and technology. To narrow the focus, studying the cohort of NLSY born in the 1980s is most appropriate. Since people of 1990s are just getting married and not having the diverse marital and household status and people of 2000s and younger are not mostly entering the labor force, my data sets of income_factors based on the NLSY97 cohort is the most up-to-date cohort worth analyzing. This research seeks to understand the personal and household income dynamics of people born in the 1980s from 2010 to 2020 (before Covid). Given the change in social values and technologies, this problem has been given new practical interest and meaning.

8984 respondents (4599 male, 4385 female) were surveyed 4 times respectively in 2011, 2013, 2015, 2017. The study addresses the three research questions:

1. What demographic, personal, and situational factors play significant roles in determining the job wage rates and the spouse income of individuals? Does education, marital status, or marijuana usage make a significant difference? How does the magnitude of each element's influence change across different years?
2. Is there a pattern of serial correlation in spouse income, family gross income, and in our modeling? If so, what is the best structure to characterize the pattern in time series?
3. Can we predict the odds of a given person to be searching for a job each year? Can we make interpretations for the motivations for job search?

[VI. Methodology]

Due to the limit of the summer research schedule, the analysis will begin with cross-sectional data from one time point and will expect to include time series analysis in the future. To pick one time intercept, it is necessary to avoid certain idiosyncratic time periods, such as COVID, the 2008 to 2009 financial crisis, or other peaks or troughs of the economy.

In addition to ensuring the cohorts are approximately 20 to 30 years old during surveys, we choose the period from 2011 to 2017 because this is a period of continued economic expansion. No downturn or recession in the economy allows more consistency between models and enables us to see the trends with time series.

This research specifically aims to examine how personal behavioral, demographical, and situational factors, such as race, education, and substance, impact the individual's income dynamics and wage rates in the 21st century. To model the wage, ordinary least square models are used to predict the wage rates of the person's primary job. In addition, logistic regression is used to model the probability of a person to be seeking a job. In general, this research begins with cross-sectional analysis by assuming individual years as independent. Then this research will attempt to make primitive longitudinal analysis to account for the change in importance of these factors over time and examine the trend of these factors.

[VII. Exploratory Trend Analysis]

To get an overview of the labor income dynamics, here's the trend for three major metrics of income. Their correlation across years is respectively tabulated below[6]: spouse's annual income (of the previous year), household gross annual income (of the previous year), and hourly wage rates.

To proceed from the correlation matrix, we want to examine if the correlation matrix fits with the AR [1] structure.

The first-order autoregressive [AR (1)] structure can be written as

$$u_{it} = \rho u_{i,t-1} + e_{it}$$

For which

$$E(e_{it}|x_{it}, u_{i,t-1}, x_{i,t-1}, \cdots) = 0, E\left(e_{it}^2|x_{it}\right) = E\left(e_{it}^2\right) = \sigma^2$$

With

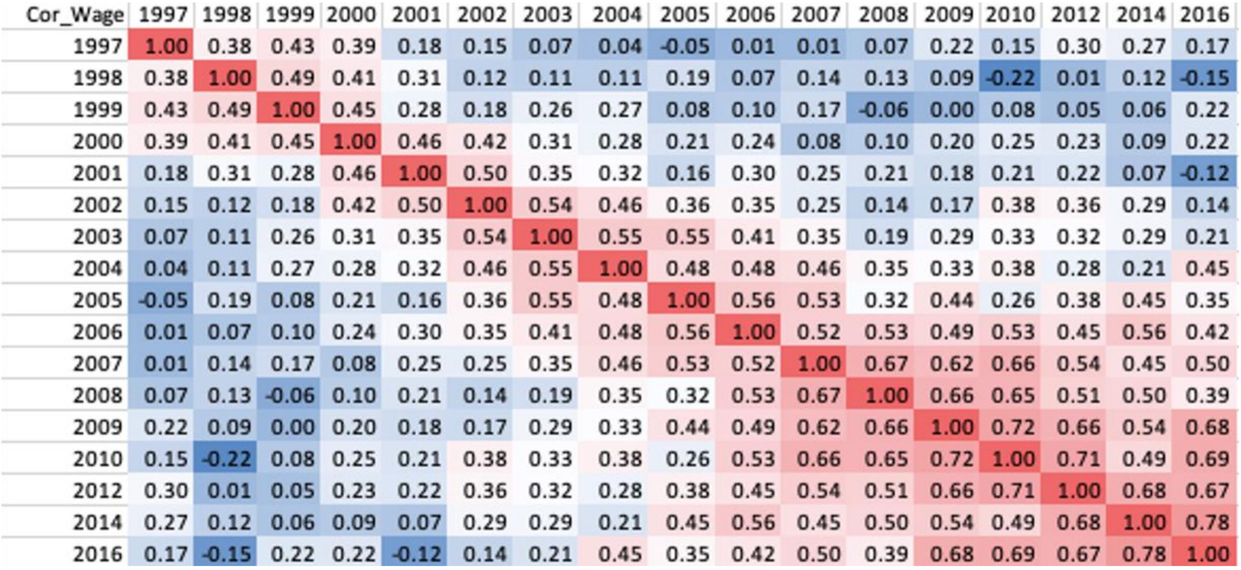
$$|\rho| < 1$$

for this [AR (1)] structure, assuming there are 4 years selected, say 2011, 2013, 2015, 2017, then the covariance matrix structure should have the following structure:

$$\Omega = \frac{\sigma^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

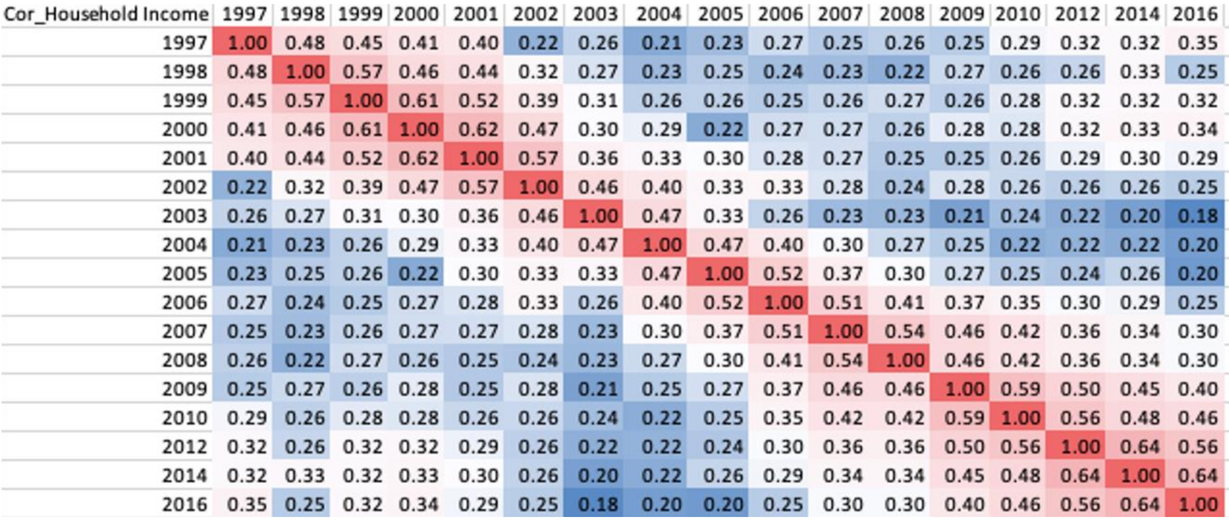
[6] Some of the correlations are very small due to the missing data points in some years.

It is not difficult to spot that AR (1) is not appropriate: if the correlation matrix indeed follows AR (1) structure, then going down from primal diagonal to subprime diagonal and down, the correlation decreases exponentially, just like ρ, ρ^2, ρ^3, $\cdots$, etc. with $|\rho|<1$. Here the correlation from one year to the next is so low compared to our expectation. For exploratory purposes, we find that AR (2) may fit the correlation matrix better. That suggests there is a positive correlation between the income in the past two years and income this year for the NLSY97 cohort.



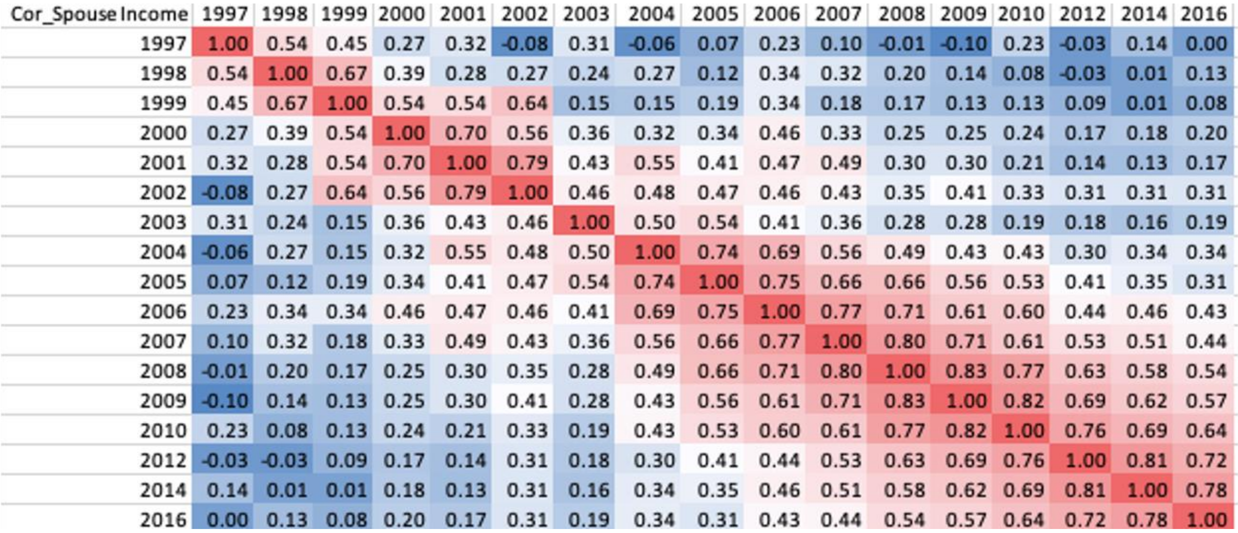
Cor_Wage	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2012	2014	2016
1997	1.00	0.38	0.43	0.39	0.18	0.15	0.07	0.04	-0.05	0.01	0.01	0.07	0.22	0.15	0.30	0.27	0.17
1998	0.38	1.00	0.49	0.41	0.31	0.12	0.11	0.11	0.19	0.07	0.14	0.13	0.09	-0.22	0.01	0.12	-0.15
1999	0.43	0.49	1.00	0.45	0.28	0.18	0.26	0.27	0.08	0.10	0.17	-0.06	0.00	0.08	0.05	0.06	0.22
2000	0.39	0.41	0.45	1.00	0.46	0.42	0.31	0.28	0.21	0.24	0.08	0.10	0.20	0.25	0.23	0.09	0.22
2001	0.18	0.31	0.28	0.46	1.00	0.50	0.35	0.32	0.16	0.30	0.25	0.21	0.18	0.21	0.22	0.07	-0.12
2002	0.15	0.12	0.18	0.42	0.50	1.00	0.54	0.46	0.36	0.35	0.25	0.14	0.17	0.38	0.36	0.29	0.14
2003	0.07	0.11	0.26	0.31	0.35	0.54	1.00	0.55	0.55	0.41	0.35	0.19	0.29	0.33	0.32	0.29	0.21
2004	0.04	0.11	0.27	0.28	0.32	0.46	0.55	1.00	0.48	0.48	0.46	0.35	0.33	0.38	0.28	0.21	0.45
2005	-0.05	0.19	0.08	0.21	0.16	0.36	0.55	0.48	1.00	0.56	0.53	0.32	0.44	0.26	0.38	0.45	0.35
2006	0.01	0.07	0.10	0.24	0.30	0.35	0.41	0.48	0.56	1.00	0.52	0.53	0.49	0.53	0.45	0.56	0.42
2007	0.01	0.14	0.17	0.08	0.25	0.25	0.35	0.46	0.53	0.52	1.00	0.67	0.62	0.66	0.54	0.45	0.50
2008	0.07	0.13	-0.06	0.10	0.21	0.14	0.19	0.35	0.32	0.53	0.67	1.00	0.66	0.65	0.51	0.50	0.39
2009	0.22	0.09	0.00	0.20	0.18	0.17	0.29	0.33	0.44	0.49	0.62	0.66	1.00	0.72	0.66	0.54	0.68
2010	0.15	-0.22	0.08	0.25	0.21	0.38	0.33	0.38	0.26	0.53	0.66	0.65	0.72	1.00	0.71	0.49	0.69
2012	0.30	0.01	0.05	0.23	0.22	0.36	0.32	0.28	0.38	0.45	0.54	0.51	0.66	0.71	1.00	0.68	0.67
2014	0.27	0.12	0.06	0.09	0.07	0.29	0.29	0.21	0.45	0.56	0.45	0.50	0.54	0.49	0.68	1.00	0.78
2016	0.17	-0.15	0.22	0.22	-0.12	0.14	0.21	0.45	0.35	0.42	0.50	0.39	0.68	0.69	0.67	0.78	1.00

FIGURE 1.0 CORRELATION OF WAGE OF NLSY97 COHORT



Cor_Household Income	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2012	2014	2016
1997	1.00	0.48	0.45	0.41	0.40	0.22	0.26	0.21	0.23	0.27	0.25	0.26	0.25	0.29	0.32	0.32	0.35
1998	0.48	1.00	0.57	0.46	0.44	0.32	0.27	0.23	0.25	0.24	0.23	0.22	0.27	0.26	0.26	0.33	0.25
1999	0.45	0.57	1.00	0.61	0.52	0.39	0.31	0.26	0.26	0.25	0.26	0.27	0.26	0.28	0.32	0.32	0.32
2000	0.41	0.46	0.61	1.00	0.62	0.47	0.30	0.29	0.22	0.27	0.27	0.26	0.28	0.28	0.32	0.33	0.34
2001	0.40	0.44	0.52	0.62	1.00	0.57	0.36	0.33	0.30	0.28	0.27	0.25	0.25	0.26	0.29	0.30	0.29
2002	0.22	0.32	0.39	0.47	0.57	1.00	0.46	0.40	0.33	0.33	0.28	0.24	0.28	0.26	0.26	0.26	0.25
2003	0.26	0.27	0.31	0.30	0.36	0.46	1.00	0.47	0.33	0.26	0.23	0.23	0.21	0.24	0.22	0.20	0.18
2004	0.21	0.23	0.26	0.29	0.33	0.40	0.47	1.00	0.47	0.40	0.30	0.27	0.25	0.22	0.22	0.22	0.20
2005	0.23	0.25	0.26	0.22	0.30	0.33	0.33	0.47	1.00	0.52	0.37	0.30	0.27	0.25	0.24	0.26	0.20
2006	0.27	0.24	0.25	0.27	0.28	0.33	0.26	0.40	0.52	1.00	0.51	0.41	0.37	0.35	0.30	0.29	0.25
2007	0.25	0.23	0.26	0.27	0.27	0.28	0.23	0.30	0.37	0.51	1.00	0.54	0.46	0.42	0.36	0.34	0.30
2008	0.26	0.22	0.27	0.26	0.25	0.24	0.23	0.27	0.30	0.41	0.54	1.00	0.46	0.42	0.36	0.34	0.30
2009	0.25	0.27	0.26	0.28	0.25	0.28	0.21	0.25	0.27	0.37	0.46	0.46	1.00	0.59	0.50	0.45	0.40
2010	0.29	0.26	0.28	0.28	0.26	0.26	0.24	0.22	0.25	0.35	0.42	0.42	0.59	1.00	0.56	0.48	0.46
2012	0.32	0.26	0.32	0.32	0.29	0.26	0.22	0.22	0.24	0.30	0.36	0.36	0.50	0.56	1.00	0.64	0.56
2014	0.32	0.33	0.32	0.33	0.30	0.26	0.20	0.22	0.26	0.29	0.34	0.34	0.45	0.48	0.64	1.00	0.64
2016	0.35	0.25	0.32	0.34	0.29	0.25	0.18	0.20	0.20	0.25	0.30	0.30	0.40	0.46	0.56	0.64	1.00

FIGURE 1.1 CORRELATION OF HOUSEHOLD GROSS INCOME OF NLSY97 COHORT



Cor_Spouse Income	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2012	2014	2016
1997	1.00	0.54	0.45	0.27	0.32	-0.08	0.31	-0.06	0.07	0.23	0.10	-0.01	-0.10	0.23	-0.03	0.14	0.00
1998	0.54	1.00	0.67	0.39	0.28	0.27	0.24	0.27	0.12	0.34	0.32	0.20	0.14	0.08	-0.03	0.01	0.13
1999	0.45	0.67	1.00	0.54	0.54	0.64	0.15	0.15	0.19	0.34	0.18	0.17	0.13	0.13	0.09	0.01	0.08
2000	0.27	0.39	0.54	1.00	0.70	0.56	0.36	0.32	0.34	0.46	0.33	0.25	0.25	0.24	0.17	0.18	0.20
2001	0.32	0.28	0.54	0.70	1.00	0.79	0.43	0.55	0.41	0.47	0.49	0.30	0.30	0.21	0.14	0.13	0.17
2002	-0.08	0.27	0.64	0.56	0.79	1.00	0.46	0.48	0.47	0.46	0.43	0.35	0.41	0.33	0.31	0.31	0.31
2003	0.31	0.24	0.15	0.36	0.43	0.46	1.00	0.50	0.54	0.41	0.36	0.28	0.28	0.19	0.18	0.16	0.19
2004	-0.06	0.27	0.15	0.32	0.55	0.48	0.50	1.00	0.74	0.69	0.56	0.49	0.43	0.43	0.30	0.34	0.34
2005	0.07	0.12	0.19	0.34	0.41	0.47	0.54	0.74	1.00	0.75	0.66	0.66	0.56	0.53	0.41	0.35	0.31
2006	0.23	0.34	0.34	0.46	0.47	0.46	0.41	0.69	0.75	1.00	0.77	0.71	0.61	0.60	0.44	0.46	0.43
2007	0.10	0.32	0.18	0.33	0.49	0.43	0.36	0.56	0.66	0.77	1.00	0.80	0.71	0.61	0.53	0.51	0.44
2008	-0.01	0.20	0.17	0.25	0.30	0.35	0.28	0.49	0.66	0.71	0.80	1.00	0.83	0.77	0.63	0.58	0.54
2009	-0.10	0.14	0.13	0.25	0.30	0.41	0.28	0.43	0.56	0.61	0.71	0.83	1.00	0.82	0.69	0.62	0.57
2010	0.23	0.08	0.13	0.24	0.21	0.33	0.19	0.43	0.53	0.60	0.61	0.77	0.82	1.00	0.76	0.69	0.64
2012	-0.03	-0.03	0.09	0.17	0.14	0.31	0.18	0.30	0.41	0.44	0.53	0.63	0.69	0.76	1.00	0.81	0.72
2014	0.14	0.01	0.01	0.18	0.13	0.31	0.16	0.34	0.35	0.46	0.51	0.58	0.62	0.69	0.81	1.00	0.78
2016	0.00	0.13	0.08	0.20	0.17	0.31	0.19	0.34	0.31	0.43	0.44	0.54	0.57	0.64	0.72	0.78	1.00

FIGURE 1.2 CORRELATION OF SPOUSE INCOME OF NLSY97 COHORT

[VIII. Theoretical Constructs of Cross-Sectional Models]

[VIII.I OLS model for labor income]

If we want to have a valid framework of explanatory variables, then we need to track the change across time. Thus, we want to avoid picking a year in the financial crisis 2008-2009 and get a continuous period where we can take multiple years. Our OLS model for labor income[7] is the following:

$$W_{it} = \alpha + \beta_1 X_{it} + \beta_1 Z_{it} + e_{it}$$

In the OLS model, α is time invariant term, and $e_{it}$is error term for individual i at time t, for which t=2011,2013,2015,2017.

X_{it} represents the level of the individuals' education for individual i at time t (no education, GED, high school, associate/junior college, bachelor's, master's, PhD, or professional degrees), and Z_{it} represents the demographic and socioeconomic regressors, including sex (binary dummy), marital status, region, usage of substance, and access/usage frequency of Internet (as a measure of financial conditions).
In previous literature, Song (2011), examined the labor market impacts of having a GED credential. However, given the efforts in improving GED education, most people tend to have completed GED according to the NLSY97 cohort data (only 746 in 2011, 647 in 2013, 616 in 2015, and 536 in 2017 out of 8984 survey participants did not complete GED certificate). Thus, using a binary variable for education is no longer sufficient. We need to accurately reflect the multiple levels of education within the sample. It's helpful to shift the research interests to comparisons of education into comparisons of three categories: having no education or GED, having completed high school or associate's/junior college, or having bachelors or above.

[VIII.II Logistic Model for job seeking probabilities]

Following Hayashi 2000[8], I use the logit model of logistic regression to measure the binary output of whether the person sought a job in three months prior to the survey.

The logit model for binary response is

$$f(y_t = 1|x'_t; \theta_0) = \Lambda\left(x'_t\theta_0\right), f(y_t = 0|x'_t; \theta_0) = 1 - \Lambda\left(x'_t\theta_0\right)$$

$$\Lambda(v) := \frac{e^v}{1 + e^v}$$

And thus, the objective function to be maximized is

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\{y_t \log(\Lambda(\mathrm{x}'_t\theta_0) + (1 - y_t)\log(1 - \Lambda(\mathrm{x}'_t\theta_0))\}$$

[7] There are three measures of income, such as spouse income, gross household income, and labor wages. Without loss of generality, we use labor wage as a response variable here to explain the procedure, which is the same if we use spouse income and gross household income as independent variables.

[8] F. Hayashi, Econometrics, 2000, Page 508

Using this logit model, the model for probability of searching for job is:

$$\alpha + \beta_1 X_{it} + \beta_2 Z_{it} + e_{it}$$

In the logit model, α is time invariant term, and e_{it} is error term for individual i at time t, for which t=2011,2013,2015,2017. X_{it} and Z_{it} are just the same with that in the OLS model. $Y_{it} = 0$ if the individual didn't seek for job in the past three months, or 1 if the individual did seek for job in the past three months. Admittedly, there's a more direct question of "did you want full time work", but that question is only asked in the 2006 survey for the NLSY97 cohort. But what we care about is not just the employment status, but the job seeking behavior. Fortunately, the job seeking behavior status was recorded well so we just use it as the binary response variable.

In all income_factors data set, the categorical variables used are the following:
1.Family background
2.Job conditions
3.Income and wealth
4.Personal demographics

The representative set of regressors includes sex, region, education, Internet usage frequency. Some of the models only have a subset of the regressors. Since I used R to make the model with factor variables, it involves some unique ways to interpret coefficients of dummy variables. For more details, please refer to the appendix.

[IX. Model Results and Interpretations]

[IX.I labor income OLS results]

Following our theoretical construct for OLS, here we present the results of the four years chosen, with estimates and power of each of the regressors:

Table 2.1 Labor Income OLS results without aggregating categories

Response variable: labor wage rates	2011		2013		2015		2017	
(total income in 2015)	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
(Intercept)	1587.430	0.000	1940.360	0.000	32968.000	0.000	1864.510	0.000
sex:woman	-473.230	0.000	-631.280	0.000	-17225.000	0.000	-656.530	0.000
region:North Central	-223.390	0.040	-236.560	0.078	-5063.000	0.097	-250.500	0.107
region:South	-142.680	0.145	-211.200	0.085	-1562.000	0.575	-371.310	0.007
region:West	16.260	0.879	124.130	0.359	1603.000	0.596	278.910	0.069
educ:GED	75.190	0.622	68.250	0.695	4095.000	0.423	154.010	0.525
educ:high school	319.160	0.014	318.530	0.038	11874.000	0.008	422.070	0.033
educ:associates	506.060	0.003	583.520	0.004	12709.000	0.015	784.690	0.000
educ:bachelor	822.610	0.000	1015.260	0.000	22962.000	0.000	1554.930	0.000

educ:master	1399.410	0.000	1194.550	0.000	29222.000	0.000	2242.960	0.000
educ:professional degree	2052.260	0.003	1427.620	0.075	23191.000	0.033	2285.270	0.000
educ:PhD	2628.260	0.000	2770.100	0.000	59400.000	0.000	5964.180	0.000
Internet:muliple times/day	-42.760	0.675	-228.780	0.093	-5698.000	0.135	NA	NA
Internet: once/day	-56.900	0.655	-356.960	0.059	-8402.000	0.089	NA	NA
Internet: 3-5days/week	-257.820	0.093	449.310	0.063	-7593.000	0.343	NA	NA
Internet: 1-2days/week	-337.690	0.061	-170.040	0.557	-13174.000	0.223	NA	NA
Internet: once/week	-372.110	0.092	-408.790	0.222	-10639.000	0.297	NA	NA
Internet: no	-66.080	0.705	-88.840	0.722	-8568.000	0.342	NA	NA
marital:married	275.180	0.000	312.010	0.000	10427.000	0.000	420.560	0.000
marital:separated	24.470	0.930	427.740	0.233	14034.000	0.024	-59.330	0.852
marital:divorced	-10.610	0.938	341.210	0.026	5369.000	0.130	-1.060	0.995
marital:widowed	-128.050	0.884	38.170	0.960	-3158.000	0.878	-361.900	0.647
marijuana:yes	101.080	0.252	-71.530	0.398	-3446.000	0.164	NA	NA

The tabulated results above are from four respective OLS models in 2011, 2013, 2015, 2017. Note that in 2015, since the data set does not have hourly wage rate as the independent variable, we use the person's annual income as the response variable. Similarly, in 2017, since the survey did not ask for whether the respondent used marijuana in the past 12 months, we use whether the respondent smoked in the past 12 months as an indicator of substance usage instead. Despite that the people who use marijuana may not necessarily smoke, this is a good proxy[9], because both marijuana and cigarettes have similar levels of addictive power and harm to the human body.

The model suggests that most of the categories in marital status and education are significant. Sex and Internet usage are not. But before moving to discussions and interpretations, we realize that in this raw model, there are too many categories. That leads to small counts in each of the categories and decreases the significance of the model due to data points' idiosyncrasies. To minimize the impact of "small counts" in some categories of the factor variables, we need to aggregate some categories together for the OLS and logistic models of 2011, 2013, 2015, 2017. This is also theoretically justifiable. For example, the difference between not completing high school and high school is much bigger than the difference between associate and bachelor's degree. So, we may compress education into two simple categories of "completed high school" or "not completed high school". Likewise, we compressed 'divorced, separated, and widowed' together into one category of 'was married'. This helps us deal with the factor variables with unbalanced counts in each category.

Table 2.2 Labor Income OLS results with categories aggregated.

Response variable: job search probability	2011		2013		2015		2017	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
(Intercept)	-1.822	0.000	-2.737	0.000	-2.626	0.000	-2.834	0.000
sex:woman	0.055	0.513	-0.021	0.860	-0.014	0.893	-0.146	0.183
region:North Central	-0.230	0.096	0.167	0.368	-0.036	0.821	-0.053	0.769

[9] In the subsample selected for 2011, 2013, 2015, 2017 models, by looking at the unique ID of every respondent, we find that co-use of both substances is more common than tobacco or marijuana use only.

region:South	-0.045	0.708	0.033	0.851	-0.083	0.562	0.043	0.786
region:West	-0.119	0.365	0.076	0.689	-0.067	0.666	0.016	0.929
educ:high school&associates	0.026	0.845	0.248	0.190	0.412	0.018	0.319	0.133
educ:bachelor&above	0.556	0.000	1.073	0.000	0.875	0.000	1.017	0.000
Internet:muliple times/day	-0.196	0.159	-0.885	0.002	-0.319	0.177	NA	NA
Internet: once/day	0.150	0.347	-0.030	0.921	-0.400	0.192	NA	NA
Internet: 3-5days/week	-0.130	0.541	0.537	0.098	0.014	0.972	NA	NA
Internet: 1-2days/week	-0.247	0.366	-0.535	0.376	-0.498	0.496	NA	NA
Internet: once/week	-0.605	0.132	-13.269	0.957	-0.860	0.239	NA	NA
Internet: no	-1.320	0.002	-1.658	0.103	-1.604	0.113	NA	NA
marital:married	-0.476	0.000	-0.134	0.275	-0.217	0.040	-0.268	0.021
marital:was married	-0.146	0.368	-0.207	0.384	-0.310	0.098	-0.406	0.041
substance usage:yes	0.328	0.002	0.432	0.000	0.349	0.004	NA	NA
number of weeks unemployed	0.012	0.074	0.045	0.003	0.060	0.000	0.008	0.726

After fixing the problem of small counts, we get a sound model with a high degree of significance. In the OLS fits of 2011, 2013, 2015, and 2017, we all have large degrees of freedom of over 4000. The counts are also more balanced here.

Since sex, region, education, marital status, and substance usage status are all factor variables with several categories, we use ANOVA[10] to test the significance of each of the variables. It turns out that sex, region, and education are significant at $\alpha = 0.01$ level and Internet usage frequency is significant at $\alpha = 0.1$ level (the p-value is mostly around 0.06).

From the model, we find that men are estimated to make approximately 4.63 dollars more than women every hour. People in North Central and South are estimated to make 2.4 and 1.6 dollars per hour less than people in the Northeast. There's no significant difference in hourly wage rates between Northeast and West. This presents the difference in economic opportunities across the United States, as the West Coast and Northeast tend to be more prosperous.

Being Married is associated with a large increase (approximately 3 dollars) hourly wage rate compared to never married. For those who are married, they also have a higher hourly wage. This effect is likely explained by two possibilities: first, people who want to get married need to financially stabilize themselves before marriage, so the married people tend to have higher wages. Second, marriage is a big encouragement for people to perform better in the workplace and thus get bonuses or pay rises.

Education is the most significant boost for wage, which agrees with our intuition. Completing high school and bachelors' degree are the two critical milestones for increase in job pay (both lead to at least 5 dollars per hour increase in wage). Comparing table 2.1 and table 2.2, we find that people with higher education degrees such as master's and PhD earn a lot more.

The above differences between categories are in fact quite similar across 2011 to 2017 by comparing the estimates of the OLS coefficients horizontally. From the data set, we also find that the wage levels increase by year.

[10] See ANOVA results in appendix of the OLS model.

[IX.II job seeking logistic results]

Likewise, in the logistic model, we also aggregate the categories.

Table 2.3 Logistic model of probability for seeking jobs (with unemployment added)

Response variable: job search probability	2011		2013		2015		2017	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
(Intercept)	-1.905	0.000	-3.469	0.000	-2.681	0.000	-2.987	0.000
sex:woman	0.053	0.580	0.152	0.342	0.008	0.942	-0.141	0.272
region:North Central	-0.152	0.313	0.408	0.095	0.115	0.516	-0.030	0.882
region:South	-0.033	0.810	0.302	0.203	-0.053	0.757	-0.014	0.940
region:West	-0.035	0.811	0.138	0.602	0.001	0.995	0.103	0.603
educ:high school&associates	0.112	0.484	0.730	0.026	0.342	0.108	0.425	0.122
educ:bachelor&above	0.558	0.001	1.410	0.000	0.905	0.000	1.056	0.000
marital:married	-0.459	0.000	-0.042	0.804	-0.178	0.147	-0.118	0.167
marital:was married	-0.072	0.695	-0.084	0.810	-0.233	0.289	-0.333	0.706
Internet:muliple times/day	-0.349	0.034	-0.823	0.042	-0.328	0.254	NA	NA
Internet: once/day	-0.161	0.441	-1.205	0.099	-0.321	0.392	NA	NA
Internet: 3-5days/week	-0.127	0.604	-0.084	0.893	0.037	0.945	NA	NA
Internet: 1-2days/week	-0.374	0.273	0.895	0.179	-14.014	0.977	NA	NA
Internet: once/week	-14.637	0.963	-14.179	0.984	-0.228	0.757	NA	NA
Internet: no	-1.280	0.013	-13.966	0.977	-14.019	0.971	NA	NA
substance usage:yes	0.295	0.013	0.388	0.021	0.369	0.007	NA	NA
weeks_unemployed_predicted	0.002	0.958	0.029	0.471	-0.019	0.594	-0.021	0.706

From exploratory data analysis, we find that education increases willingness to full time employment. Household income situation seems to play a relatively small role in it, as we see the coefficient is small. The more education the individual's parents had, the more likely the person will want a full-time job. Note that willingness to take a full-time job is only a prospective question the NLSY97 cohort faced when they were younger. When they reach their 20s and 30s, they face the actual decisions about whether to search for jobs. There is an aging effect we aim to see.

For the cross-sectional results tabulated above in table 1.3, we find that the model significance is strong for 2011, 2013, 2015, and even 2017 (which has fewer variables).

The category of West in the region has low significance, suggesting that there is no significant difference in job search willingness between people in the West states and people in the Northeast Coast, which are both very developed.

Having married once (means both currently married or was married) significantly decreases one's chance of "searching for a job in the past 3 months". We may suspect that marriage may provide economic security to individuals, or individuals would be more stable with their employment once they have married so they are less likely to lose their jobs or switch their jobs.

In addition to these results, we wonder if size of the household and income to poverty ratios could indicate the probability to be seeking jobs. Unlike other factor variables which are mostly fixed for an individual within 10 years, these are good situational variables that change across times.

Table 2.4 Logistic model of probability for seeking jobs with more variables added

Response variable: job search probability	2011		2013		2015		2017	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
(Intercept)	-1.601	0.000	-2.753	0.000	-2.378	0.000	-2.553	0.000
sex:woman	0.001	0.988	-0.012	0.920	-0.009	0.934	-0.191	0.101
region:North Central	-0.218	0.139	0.180	0.336	-0.045	0.792	-0.080	0.676
region:South	-0.042	0.749	0.048	0.784	-0.068	0.659	-0.011	0.950
region:West	-0.137	0.337	0.088	0.643	-0.017	0.919	0.030	0.870
educ:high school&associates	0.116	0.421	0.231	0.227	0.459	0.017	0.302	0.180
educ:bachelor&above	0.736	0.000	1.025	0.000	0.973	0.000	1.040	0.000
marital:married	-0.410	0.000	-0.136	0.304	-0.155	0.213	-0.094	0.490
marital:was married	-0.054	0.760	-0.208	0.384	-0.325	0.120	-0.324	0.133
Internet:muliple times/day	-0.265	0.078	-0.877	0.002	-0.188	0.432	NA	NA
Internet: once/day	0.085	0.623	-0.014	0.964	-0.211	0.495	NA	NA
Internet: 3-5days/week	-0.183	0.425	0.548	0.091	-0.113	0.812	NA	NA
Internet: 1-2days/week	-0.273	0.354	-0.522	0.388	-0.234	0.752	NA	NA
Internet: once/week	-0.493	0.227	-13.250	0.957	-0.789	0.283	NA	NA
Internet: no	-1.122	0.009	-1.642	0.107	-1.319	0.194	NA	NA
substance usage:yes	0.346	0.002	0.429	0.000	0.337	0.009	NA	NA
number of weeks unemployed	0.016	0.035	0.046	0.002	0.064	0.000	0.014	0.523
income_poverty_ratio	-0.001	0.000	0.000	0.451	0.000	0.290	0.000	0.093
household_size	-0.032	0.331	-0.009	0.837	-0.101	0.011	-0.063	0.128

Income to poverty ratio and household size are both statistically significant variables across all four years. However, the estimate of their influence is small relative to education and marital status. Income to poverty rates have little effect on the job search, and households have a small positive effect on job search.

After considering the individual variables’ significance, we want to evaluate the overall significance of the logistic models. To do that, we use chi-square statistics to calculate the overall p-value for each logistic model. The chi-square distribution’s deviance and degrees of freedom can be calculated from the null and residual distribution in the logistic regression output.

$$deviance = deviance_{null} - deviance_{residual}$$

$$df = df_{null} - df_{residual}$$

Despite the small practical influence, we find that the p-value of the model in Table 2.5 to be even smaller than the model in Table 2.3. Thus, including income to poverty ratio and household size help further improve the overall significance of the model.

[IX.III Discussion of Other Estimators]

The exploratory analysis verifies our intuition that there is a serial trend in our model that affects our interpretations of the results. This part will be discussed in the lagged structure part.

Given that our models are independent in each year, we wonder if we have some other estimators that can reflect the trend in job seeking behaviors and unemployment. For longitudinal modeling, we need to focus on the respondents who made valid responses across 2011, 2013, 2015, 2017. After this procedure, we may address how serial effects impact our modeling. Note that categorical variables always do not change with time, and this is the problem that prevents us from using fixed effects estimators. Thus, the between and within estimators would not work. We may want to try a pooled estimator, but even when that is feasible, it does not offer substantial insights by putting the four years together.

[X. Model Applications]

The analysis has forecasting applications to people’s actions in the job market. Here are two examples:

Assume a high school educated Southern woman, married, and had 5 weeks unemployed, then the probability of job search ≈ 4.8%. A married Western man who has a bachelor’s degree, is employed all year long, uses the Internet every day, has a 2-person family, and uses no marijuana has a probability of job search ≈ 16.6%.

The model does not go into details of inter-state differences. But it characterizes how the combination of factors drives people’s job search decisions. The primary positive contributor to the high probability that the man may engage in job research is his bachelor’s degree. This may be a manifestation of better socioeconomic mobility (the possibility to change jobs) for young, educated males on the West Coast. By contrast, the Southern married woman may be more likely to be sacrificing her job prospects for family reasons, even though she was not fully employed the previous year.

[XI. Serial Correlation & Autoregressive Structure]

Since we treat the income measures respectively in 2011, 2013, 2015, 2017, we wonder if there’s some serial correlation within the pattern. Naturally, to do this, there are typically three methods: calculating the residuals via cross-sections, calculate the residuals across time series, or include the response variable of the year before or several years previously to the regression model.

However, the first two are not feasible. Calculating the residuals across the cross sections would not report the pattern of the change in influence of each variable. Calculating the residuals across the time series would require the matching data points (the same subset of the NLSY97 cohort). However, we find that it’s impossible to get enough individuals whose responses are all valid for all variables across the four years. Thus, we can’t form a time series without sacrificing some variables.

For this reason, in exploring the income variables, we include the lag1, lag2, and lag3 terms of income variables. For example, in the Ordinary Least Square model for spouse income in 2016, we include spouse income in 2014, spouse income in 2012, spouse income in 2010. Likewise, in the Ordinary Least Square model for spouse income in 2014, we include spouse income in 2012 and spouse income in 2010.

Table 3.1 OLS model for spouse income in 2016

```
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            2.435e+04  2.690e+03   9.050  < 2e-16
## spoue_income2014                       4.164e-01  2.210e-02  18.842  < 2e-16
## spoue_income2012                       2.841e-01  2.546e-02  11.156  < 2e-16
## spoue_income2010                       1.542e-01  2.790e-02   5.526 3.56e-08
## sex:female                             7.356e+03  1.308e+03   5.625 2.02e-08
## region:North Central                  -6.672e+03  2.092e+03  -3.189 0.001441
## region:South                          -6.799e+03  1.938e+03  -3.508 0.000458
## region:West                           -2.141e+03  2.082e+03  -1.029 0.303737
## educ:high school&associates            3.311e+03  2.055e+03   1.611 0.107230
## educ:bachelor&above                    1.290e+04  2.155e+03   5.987 2.39e-09
## marital:married                       -3.728e+03  1.773e+03  -2.103 0.035537
## marital:was married                   -5.156e+02  2.680e+03  -0.192 0.847441
```

Table 3.2 OLS model for spouse income in 2014

```
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            2.406e+04  2.710e+03   8.878  < 2e-16
## spoue_income2014                       4.170e-01  2.211e-02  18.858  < 2e-16
## spoue_income2012                       2.843e-01  2.547e-02  11.163  < 2e-16
## spoue_income2010                       1.546e-01  2.791e-02   5.540 3.29e-08
## sex: female                            7.317e+03  1.308e+03   5.593 2.44e-08
## region:North Central                  -6.702e+03  2.092e+03  -3.203 0.001372
## region:South                          -6.818e+03  1.938e+03  -3.518 0.000442
## region:West                           -2.111e+03  2.082e+03  -1.014 0.310671
## educ:high school&associates            3.443e+03  2.060e+03   1.671 0.094849
## educ:bachelor&above                    1.307e+04  2.163e+03   6.040 1.73e-09
## marital:married                       -3.669e+03  1.774e+03  -2.069 0.038666
## marital:was married                   -4.339e+02  2.682e+03  -0.162 0.871475
## number of weeks unemployed in 2016     1.065e+02  1.205e+02   0.884 0.376978
```

There is no obvious difference for significance models if we include weeks of unemployment. We find that the lag 1, lag 2, lag 3 terms are all significant. To analyze the exact autoregressive structure of the correlation, we may output the inverse of correlation matrix of the 2014 and 2016 spouse income models[11].

Table 3.3 Inverse matrix of the correlation matrix of spouse income in 2010, 2012, 2014, 2016

[11] See appendix for code for computation of the covariance matrix.

	2011	2013	2015	2017
2011	0.42793166	-0.0345807133	0.0341346106	0.022627779
2013	-0.03458071	0.8306445934	0.0003091451	-0.004251213
2015	0.03413461	0.0003091451	0.7000389168	-0.035372231
2017	0.02262778	-0.0042512135	-0.0353722306	0.711217598

In this inverse matrix, we find that the terms off the main diagonal are relatively small. If there is some autoregressive structure, then there will be a special correlation structure where the correlation decays quickly as over years. But this inverse matrix does not accord with the autoregressive structure because the off-diagonal elements are all very small.

However, the fact that we do not find a clear autoregressive structure here is likely because the data set has many missing values. The percentage of missing value increases when we calculate pairwise correlation. We suspect that an AR (2) structure may fit well with the correlation matrix, but that requires further modeling.

[XII. Limitations and Future Directions]

Some of the trend analysis in this paper has been subject to errors brought by the problem of an unbalanced data set. In certain variables in some years, there are close to 40% of the data points missing. These cases are mostly classified as ‘valid skips’ instead of surveying errors because the survey allows people to respond or not respond to questions at their discretion. Most individuals have at least one skip in one of the variables at one of the years. Despite this limit in survey research, if we may get a more balanced data set and higher frequency interviews, our research methods can be boosted.

One natural next step would be continuing the models to the 2019 release of the NLSY97 cohort to see the trends. From my analysis, it’s easy to see the importance of avoiding years of economic downturn. Thus, we must acknowledge that Covid ends these 11 years of stable economic performance since 2009 and poses some difficulty to our future analysis. Changes in time may also lead to technological reforms and changes in social norms, which may make the model predictions no longer accurate.

Lastly, the relationship behind motivations of job seeking behaviors may be further explored. This research employs only simple linear logistic models to estimate what drives the individuals to seek behaviors. There might be an issue of simultaneity among the regressors. For example, access to the Internet may influence channels to job opportunities and thus impact the length of unemployment period.

Furthermore, the linear hypothesis does not account for the conglomerate effects of multiple variables and thus may not fully explain the impetus behind job search.

People may seek jobs to get reemployment or may seek jobs simply to improve their salaries. People largely unemployed and making a small income would be likely to be seeking a job. People with relatively good education backgrounds in more developed regions tend to have more options of changing their current job. The influence of factors, such as sex and household size, agrees with our intuition, as women and people with larger household are more risk averse to instability caused by switching jobs. A linear relationship may not be ample. A nonlinear (such as quadratic) relationship between income to poverty ratio might be more helpful. In addition to the subset I selected, the NLSY data set provides various socioeconomic measures of the individuals. Thus, future research may employ them to group the individuals for their job seeking purposes.

[XIII. Bibliography]

Betts 1995: Does School Quality Matter? Evidence from the National Longitudinal Survey of Youth Source: The Review of Economics and Statistics, May 1995, Vol. 77, No. 2 (May, 1995), pp. 231-250

Blau, Francine D., and Lawrence M. Kahn. “Changes in the Labor Supply Behavior of Married WOMEN: 1980–2000.” Journal of Labor Economics 25, no. 3 (2007): 393–438. https://doi.org/10.1086/513416.

Bradley, Cathy J., Heather L. Bednarek, and David Neumark. “Breast Cancer and Women's Labor Supply.” Health Services Research 37, no. 5 (2002): 1309–27. https://doi.org/10.1111/1475-6773.01041.

Creed, P. A., King, V., Hood, M., & McKenzie, R. (2009). Goal orientation, self-regulation strategies, and job-seeking intensity in unemployed adults. Journal of Applied Psychology, 94(3), 806–813. https://doi.org/10.1037/a0015518

Cristia, Julian P. “The Effect of a First Child on Female Labor Supply.” The Journal of human resources. 43, no. 3 (2008): 487–510.

Genadek, K. R., Stock, W. A., & Stoddard, C. (2007). No-fault divorce laws and the labor supply of women with and without children. Journal of Human Resources, XLII (1), 247–274. https://doi.org/10.3368/jhr.xlii.1.247

Heckman, James J., and Thomas E. Macurdy. "A Life Cycle Model of Female Labour Supply." The Review of Economic Studies 47, no. 1 (1980): 47-74. Accessed August 3, 2021. doi:10.2307/2297103.

Mincer, Jacob. Labor Force Participation of Married Women: A Study of Labor Supply, edited by Humphries, Jane,ed Elgar Reference Collection: International Library of Critical Writings in Economics, vol. 45. Aldershot, U.K.: Elgar; distributed in the U.S. by Ashgate, Brookfield, Vt, 1995

Mroz, Thomas A. "The Sensitivity of an Empirical Model of Married Women’s Hours of Work to Economic and Statistical Assumptions." Econometrica (1986-1998) 55, no. 4 (07, 1987): 765.

Song, Wei. “Labor Market Impacts of The GED Test Credential on High School Dropouts: Longitudinal Evidence from NlSY97,” 2011-2. GED Testing Service, GED Testing Service. Available from: American Council on Education. One Dupont Circle NW Suite 250, Washington, DC 20036. https://files.eric.ed.gov/fulltext/ED541696.pdf.

Wanberg, C. R., Kanfer, R., & Rotundo, M. (1999). Unemployed individuals: Motives, job-search competencies, and job-search constraints as predictors of job seeking and reemployment. Journal of Applied Psychology, 84(6), 897–910. https://doi.org/10.1037/0021-9010.84.6.897

Wanberg, C. R., Watt, J. D., & Rumsey, D. J. (1996). Individuals without jobs: An empirical study of job-seeking behavior and reemployment. Journal of Applied Psychology, 81(1), 76–87. https://doi.org/10.1037/0021-9010.81.1.76

[XVI. Appendix]

Appendix I. OLS of number of weeks unemployed

From OLS part, we find that number of weeks unemployed is not a significant variable, but it can be predicted significantly by other variables as below:

Response variable: number of weeks unemployed	2011		2013		2015		2017	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
(Intercept)	2.879	0.000	3.664	0.000	3.897	0.000	4.328	0.000
sex:woman	0.178	0.276	0.050	0.857	0.362	0.068	0.004	0.981
region:North Central	-0.121	0.656	-0.238	0.604	-0.165	0.616	0.076	0.804
region:South	-0.124	0.609	0.427	0.304	0.257	0.380	0.088	0.745
region:West	-0.248	0.352	-0.552	0.230	-0.251	0.434	-0.177	0.552
educ:high school&associates	-1.397	0.000	-0.833	0.017	-1.382	0.000	-1.973	0.000
educ:bachelor&above	-1.989	0.000	-2.076	0.000	-2.190	0.000	-2.661	0.000
Internet:muliple times/day	0.569	0.024	1.288	0.005	1.059	0.005	NA	NA
Internet: once/day	-0.144	0.647	1.994	0.002	1.477	0.002	NA	NA
Internet: 3-5days/week	0.836	0.025	2.896	0.000	3.177	0.000	NA	NA
Internet: 1-2days/week	1.656	0.000	3.931	0.000	2.386	0.007	NA	NA
Internet: once/week	0.919	0.063	2.147	0.054	1.738	0.046	NA	NA
Internet: no	0.475	0.218	0.173	0.821	1.441	0.016	NA	NA
marital:married	-0.856	0.000	-1.557	0.000	-1.727	0.000	-1.173	0.000
marital:was married	-0.842	0.006	-0.053	0.916	-0.822	0.013	-1.177	0.000
substance usage:yes	0.681	0.002	1.188	0.000	0.543	0.040	NA	NA

Appendix II. Interpretations of factor variables in R output

The coefficients for the model may be explained as: (where each of the subscripts represent different subcategories of the categorical variable)

Sex: $S_1 + S_2 = 0$

Region: $R_1 + R_2 + R_3 + R_4 = 0$

Education: $E_0 + E_1 + E_2 + \cdots + E_7 = 0$

Internet usage frequency: $I_1 + I_2 + \cdots + I_7 = 0$

Marital: $M_0 + M_1 + \cdots + M_4 = 0$

Marijuana: $MJ_0 + MJ_1 = 0$

The model output of the following coefficients sum up to 0. Thus, it's possible to calculate the actual individual coefficient of each category, instead of looking at their comparative difference from the model coefficients.

Coefficient 1:$Intercept + S_1 + R_1 + E_0 + I_1 + M_0 + MJ_0$
Coefficient 2: $S_2 - S_1$
Coefficient 3: $R_2 - R_1$
Coefficient 4: $R_3 - R_1$
Coefficient 5: $R_4 - R_1$
Coefficient 6: $E_1 - E_0$
Coefficient 7: $E_2 - E_0$

Appendix III. R Command for testing the overall significance of logistic regression

```
p<-1−pchisq(deviance,df)
p
```

Appendix IV. Household Income, Parents' education, and willingness for full-time employment

```
model9_logit<-glm(seek_job[sel9]~HHincome_17[sel9]+sex[sel9]+Mom_educ
[sel9]+Dad_educ[sel9]+Internet[sel9],family=binomial(link="logit"));su
mmary(model9_logit)
```

```
## Call:

## glm(formula = seek_job[sel9] ~ HHincome_17[sel9] + sex[sel9] +

##     Mom_educ[sel9] + Dad_educ[sel9] + Internet[sel9], family = bino
mial (link = "logit"))

##

## Deviance Residuals:

##     Min       1Q   Median       3Q      Max

## -1.2347  -0.9741  -0.8578   1.3621   1.6056

##

## Coefficients:

##                                 Estimate Std. Error z value Pr(>|z|)


## (Intercept)                   -8.960e-01  2.704e-01  -3.314 0.000921
 ***
```

```
## HHincome_17[sel9]              1.270e-06  1.374e-06   0.924 0.355272

## sex[sel9]2                    -7.828e-02  1.781e-01  -0.440 0.660226

## Mom_educ[sel9]associate       -2.805e-02  2.551e-01  -0.110 0.912442

## Mom_educ[sel9]bachelor&above   1.140e-01  3.020e-01   0.378 0.705786

## Dad_educ[sel9]associate        3.935e-01  2.534e-01   1.553 0.120378

## Dad_educ[sel9]bachelor&above   2.279e-01  3.045e-01   0.749 0.454075

## Internet[sel9]1                4.038e-03  2.245e-01   0.018 0.985649

## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 739.28  on 563  degrees of freedom
## Residual deviance: 733.93  on 556  degrees of freedom
## AIC: 749.93
##
## Number of Fisher Scoring iterations: 4
```

Appendix V. ANOVA for OLS model

```
anova(modelols1_2011_2)
## Analysis of Variance Table
##
## Response: job1_wage[sel1_11]
##                          Df     Sum Sq   Mean Sq F value    Pr(>F)

## sex[sel1_11]             1 1.9233e+08 192332700 31.1554  2.49e-08 *
**
```

```
## region[sel1_11]          3 7.0422e+07  23473835  3.8025 0.0097615 *
*

## educ[sel1_11]            2 8.7684e+08 438420841 71.0184 < 2.2e-16 *
**

## Internet_frq[sel1_11]    6 6.7330e+07  11221714  1.8178 0.0915133 .


## marital[sel1_11]         2 1.0598e+08  52989159  8.5835 0.0001896 *
**

## marijuana[sel1_11]       1 7.3464e+06   7346417  1.1900 0.2753705


## Residuals             5777 3.5663e+10   6173342


## ---

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(modelols1_2013_2)
```

```
## Analysis of Variance Table

##

## Response: job1_wage[sel1_13]

##                         Df     Sum Sq   Mean Sq F value    Pr(>F)


## sex[sel1_13]             1 2.3783e+08 237832927 34.9550 3.645e-09 *
**

## region[sel1_13]          3 1.2485e+08  41616710  6.1165 0.0003794 *
**

## educ[sel1_13]            2 7.5467e+08 377333687 55.4578 < 2.2e-16 *
**

## marital[sel1_13]         2 1.2096e+08  60481211  8.8891 0.0001405 *
**

## Internet_frq[sel1_13]    6 8.3832e+07  13971989  2.0535 0.0554253 .


## smoke[sel1_13]           1 6.6348e+06   6634763  0.9751 0.3234614


## Residuals             4162 2.8318e+10   6803975


## ---

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(modelols1_2015_2)
```

```
## Analysis of Variance Table
##
## Response: income_15[sel1_15]
##                          Df     Sum Sq    Mean Sq F value    Pr(>F)

## sex[sel1_15]              1 2.2180e+11 2.2180e+11 58.5869 2.373e-14
***
## region[sel1_15]           3 2.4905e+10 8.3017e+09  2.1928   0.08678
 .
## educ[sel1_15]             2 4.0750e+11 2.0375e+11 53.8185 < 2.2e-16
***
## marital[sel1_15]          2 1.3031e+11 6.5153e+10 17.2096 3.588e-08
***
## Internet_frq[sel1_15]     6 3.0626e+10 5.1044e+09  1.3483   0.23187

## marijuana[sel1_15]        1 7.1110e+09 7.1110e+09  1.8783   0.17060

## Residuals              4422 1.6741e+13 3.7859e+09

## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(modelols1_2017_2)
```

```
## Analysis of Variance Table
##
## Response: job1_wage[sel1_17]
##                    Df     Sum Sq    Mean Sq F value    Pr(>F)
## sex[sel1_17]        1 3.1204e+08  312039695  25.926 3.666e-07 ***
## region[sel1_17]     3 5.4945e+08  183150006  15.217 7.363e-10 ***
## educ[sel1_17]       2 3.3876e+09 1693795800 140.728 < 2.2e-16 ***
## marital[sel1_17]    2 2.9272e+08  146362145  12.160 5.376e-06 ***
## Residuals        5514 6.6366e+10   12035969
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix VI. Code and Data Sets

The raw data sets, processed data sets, and code of analysis can all be found at this link: https://github.com/Shaolong-Lorry-Wu/NLSY97_summer21