Michael Borsellino and Arijit Sen

Examining the Asset Gap in America's Aging Population

Wealth and income are two oft-studied subjects in the social science disciplines. The effect of each has been shown to have critical effects on health (i.e. Hajat et al, 2010; Headey & Wooden, 2004) and education outcomes (i.e. Karagiannaki, 2012), community life (i.e. Woolf et al, 2015), economic health (i.e. Carroll et al, 2011) and inequality (i.e. Piketty, 2014), among others. However, assets and liabilities, as subsets of wealth, do not have the same body of knowledge.

Existing research on the wealth gap stresses the significance of three key demographic variables: marital status, educational achievement, and race. Findings suggest that marriage and marriage tenure are key predictors of higher wealth (Yamokosi & Keister, 2006). Significantly, distinctions between married, remarried, previously married, and never married can explain much of wealth heterogeneity among the older population in the United States (Wilmoth & Koso, 2002; Wu et al, 2014). Education, as expected, is also a key indicator of wealth. Those with higher education achievement consistently have higher wealth (Altonji & Doraszelski, 2005; Pollack et al, 2013). Finally, in our society, Whites have been proven to have higher wealth than Blacks and Hispanics (Taylor et al, 2011; Oliver & Shapiro, 2006; Smith, 1995).

This analysis is interested in understanding demographic differences in assets and asset growth amongst aging residents of the United States. Using this approach, we test two hypotheses: 1) marital status, education level, and race explain total assets in America's over-fifty population, and 2) marital status, education level, and race influence the likelihood that a household is in the top quartile of total assets. The next section details the data and methods utilized in this analysis.

Methodology

Health and Retirement Study

This analysis relies on the University of Michigan Health and Retirement Study (HRS) dataset. The HRS is a longitudinal panel study that explores questions about aging through a representative sample of approximately 20,000 American residents. The survey has been administered every other year since 1992 for a total of 13 samples through 2016. The original study sought to "follow age- eligible individuals and their spouses as they made the transition from active worker into retirement" (HRS – Sample Evolution). The goal of the study has since expanded to inform the discussion around America's aging population and national retirement. It seeks to accomplish this by linking health and disease histories with functional status, disability and economic factors, and retirement and health services utilization (HRS - About).

The Study relies on a multi-stage area probability sample design to select participants. The first stage involves probability proportionate to size selection of Metropolitan Statistical Areas and non-Metropolitan Statistical Area counties. This is followed by a sampling of area segments within these primary selected areas. The third stage is the systematic sampling of all housing units in the selected area segments. The final stage is the selection of age-eligible persons within selected housing units. Because it looks at housing units, age-eligible persons in nursing homes or other forms of untraditional housing are not included in the sample population. Further, the HRS oversamples minorities and residents of Florida; it also seeks to ensure the inclusion of women by sampling the spouses of all selected persons. The preferred sampling method is through a one-hour or more telephone interview in English or Spanish, though in some cases face-to-face interviews are done if telephone is inconvenient or medically impossible. In the case of death, exit interviews are conducted with the surviving spouse. There is a single questionnaire for all cohorts to promote consistency and to prevent obsolescence (HRS – Sample Evolution). The HRS has an aggregate 73% baseline response rate (HRS – Sample Sizes). This

analysis relied on HRS data from 2000 through 2006. There were a total of 45,093 observations included.

Variables

The dependent variable is Total Assets. Specifically, this analysis is interested in how the assets of study participants grow over time. Therefore, we observed Total Assets in 2000, the base year, 2002, 2004, and 2006 (see Table 1, below). There are three independent variables (Table 2, below): Marital Status, Education, and Race Category. Marital Status has three possible values: 1) married, 2) previously married, and 3) never married. Education has four possible values: 1) less than high school, 2) high school graduate, 3) some college, and 4) college graduate and above. Race also has four possible values: 1) Hispanic, 2) White, 3) Black, and 4) other. A key component of this analysis was the use of interaction terms between demographic categories and years since the base year of 2000. This helps us to identify how demographic variables influence total assets over time.

Table 1. Descriptive Statistics; Dependent Variable

Total Assets	Mean	Standard Deviation	Frequency
2000	329.43	962.54	10,082
2002	330.87	886.40	10,230
2004	377.77	1,424.79	12,223
2006	465.99	2,145.50	12,558

Table 2. Descriptive Statistics; Independent Variables

	Hispanic				White				Black				Other			
	<HS	HS	SC	C+	<HS	HS	SC	C+	<HS	HS	SC	C+	<HS	HS	SC	C+
Married	1164	432	336	176	2312	6048	4148	5696	688	684	516	416	76	100	116	216
Previously Married	1480	360	304	120	3960	6732	3512	2840	2216	1436	872	540	164	124	104	132
Unmarried	96	16	20	24	128	372	280	504	196	164	104	56	8	20	36	8

*<HS-Less than high school; HS-High school graduate; SC-Some College; C+-College graduate and above

To get a sense of the patterns in the data, we relied on Lowess smoothing to find variations in trends among households based on each demographic predictor. The first four figures in Appendix A are the mean profiles of those plots which paint a clear picture of how household assets change over time for each demographic category. For example, in Figure 1 we

see that married households show increasing assets over time, while the trend is mostly flat for previously married and never married households. In Figure 2, each step up in education is associated with higher assets in 2000, and those with college degrees show an increasing trend..

For testing the second hypothesis, it was similarly important to get an idea of any apparent patterns in the data before commencing with the analysis. Plotting subgroup properties gives us a sense of any potential marginal effects that we should expect to see. Figures 5, 6, and 7 in Appendix A show these subgroup plots. In Figure 5, we see that there are more married households in the top quartile of assets compared to those that never married.. In Figure 7, there are more White households in the top quartile of assets compared to Black and Hispanic households.

Longitudinal Analysis

Both objectives in this analysis are centered around change in growth of assets over time. As a result, a longitudinal analysis was used to capture this change for different household typologies. In order to test the first hypothesis, it was required to identify the best possible overall model for the trend in growth of assets controlling for demographic categories. A marginal linear with correlated errors using an unstructured marginal variance-covariance matrix was used for this purpose. Since we are interested in changes across all households, not changes in individual households, this model allows us to disregard potential random effects at the individual household level. For our second objective, estimating how demographic groups in the top quartile of assets changed over time, a generalized estimating equation (GEE) was used. Similar to the first model, this model was fit with an unstructured working correlation matrix. The GEE is a suitable model for non-normal outcomes, in this case a binary variable representing whether a household is in the top quartile of assets.

For both models, reference categories within each variable were defined to be: Previously Married (2) within Marital Status, Less than High School (1) within Education, and Hispanic (1) within Race Category. These reference categories are important for interpreting results in the next section. Finally, this analysis was completed primarily in Stata, though R was used to generate several plots. The complete syntax is available in Appendix B.

Results

Table 3, below, shows the results of the marginal linear model. As for fixed effects in 2000, marriage, all education categories (high school graduate, some college, college graduate and above), and being White all had positive and significant effects on asset growth compared to the reference categories. The magnitude of these significant effects can be interpreted through the posted coefficient. For example, being married in 2000 increased total assets by 160.83 compared to the reference category, holding all other variables constant. Time had no significant effect on asset growth and no single predictor had a negative and significant effect on asset growth. The interaction terms provided a somewhat different picture. Marriage had a positive and significant effect only in 2006. Being a college graduates and above had a positive and significant effect in 2004 and 2006. Both of these results are expected and consistent with the literature. "Other" race also had a positive and significant effect in 2006, though it is conceptually unclear why.

Table 3. Marginal Linear Model with Unstructured Variance-Covariance Matrix

Total Assets 2000	Coefficient	Standard Error	z	P> |z|
Married	160.83	18.74	8.58	0.00
Never Married	63.02	50.58	1.25	0.21
High School	73.10	24.32	3.01	0.00
Some College	145.60	27.69	5.26	0.00
College	451.06	28.15	16.02	0.00
White	201.94	34.55	5.84	0.00
Black	-19.68	39.47	-0.50	0.62
Other	63.91	72.33	0.88	0.38
2002	22.84	33.30	0.69	0.49
2004	-0.69	44.15	-0.02	0.99
2006	-34.15	65.80	-0.52	0.60
Married x 2002	-4.44	18.53	-0.24	0.81
Married x 2004	46.82	25.29	1.85	0.06
Married x 2006	148.45	38.19	3.89	0.00
Never Married x 2002	-68.59	50.99	-1.35	0.18
Never Married x 2004	-74.51	65.53	-1.14	0.26
Never Married x 2006	-88.20	96.49	-0.91	0.36
High School x 2002	-7.04	23.99	-0.29	0.77
High School x 2004	-9.49	33.11	-0.29	0.77
High School x 2006	-16.46	50.15	-0.33	0.74
Some College x 2002	1.12	27.45	0.04	0.97
Some College x 2004	-8.73	37.20	-0.23	0.81
Some College x 2006	36.41	56.00	0.65	0.52
College x 2002	-35.54	27.88	-1.27	0.20
College x 2004	128.20	37.80	3.39	0.00
College x 2006	271.17	57.03	4.76	0.00
White x 2002	-7.31	34.63	-0.21	0.83
White x 2004	15.36	45.63	0.34	0.74
White x 2006	63.39	67.85	0.93	0.35
Black x 2002	2.44	39.53	0.06	0.95
Black x 2004	-11.00	52.25	-0.21	0.83
Black x 2006	-9.71	77.76	-0.12	0.90
Other x 2002	-7.78	72.99	-0.11	0.92
Other x 2004	29.98	95.34	0.31	0.75
Other x 2006	277.42	140.27	1.98	0.05
Intercept	-48.32	33.29	-1.45	0.15

Using Wald chi-square tests, we were able to test the true value of the above parameters based on the sample estimate. The results showed that, of the fixed effects, marital status, education level, and race are all statistically significant. Among the interaction terms, the interaction between marital status and time proved to be statistically significant, as did the interaction between education and time. This model also yielded high error variances in 2004 and 2006 compared to in 2000 and 2002 (see Table 4, below). This supports the decision to not specify covariances to improve model fit, as the covariances vary wildly. Finally, after

comparing the predicted and actual values of total assets (Figure 8, Appendix A), we found that the error terms are not normally distributed. This suggests that a categorical outcome variable is more suited for analysis. This is attempted in the latter half of this section through the GEE specification.

Table 4. Marginal Linear Model, Variance and Covariance

Unstructured Residuals	Estimate	Standard Error
Variance(e1)	850,633.8	11,715.4
Variance(e2)	722,998.5	9,958.3
Variance(e3)	1,938,500.0	24,728.0
Variance(e4)	4,475,627.0	56,603.5
Covariance(e1,e2)	390,893.4	8,456.9
Covariance(e1,e3)	577,986.2	13,167.1
Covariance(e1,e4)	665,676.1	20,164.2
Covariance(e2,e3)	428,506.4	11,554.4
Covariance(e2,e4)	614,015.0	18,646.9
Covariance(e3,e4)	1,270,078.0	28,769.9

Table 5, below, shows the results of the generalized estimating equation for modeling binary outcomes. For fixed effects in the 2000, being married or never married as well as all education categories (high school graduate, some college, college graduate and above) had positive and significant effects on the log-odds of being in the top quartile of assets compared to the reference categories. For example, being a college graduate and above, compared to the reference level of education, and all other variables held constant, increased the log-odds of being in the top quartile of assets by 1.61. Being Black had a negative and significant effect on the log-odds of being in the top quartile compared to the reference category. And for all households with reference categories, a one-wave (two years) change reduced the log-odds of being in the top quartile of assets. The magnitude of this effect, for example, is -0.25, meaning that a one-wave change reduces the log-odds of being in the top quartile of assets by 0.25 for a reference household (Hispanic, previously married, and less than high school education). As for the interaction terms in this model, the interaction between being married and time ensured a statistically significant increase of 0.05 in log-odds of being in the top quartile of asset value

compared to the base effect of time (-0.25 for a reference household); the interaction between having never been married and time is associated with a negative and statistically significant effect. Statistically significant, positive effects are also seen for all interactions between education and time compared to the reference category, and for the interaction between White household and time compared to the reference category. The statistically significant intercept suggests that a reference household in 2000 has a -1.58 log-odds of being in the top quartile of assets, which makes sense given that all reference categories are associated with lower asset levels on average.

Table 5. Generalized Estimating Equation with Unstructured Correlation Matrix

Top Quartile	Coefficient	Standard Error	z	P>|z|
Married	0.54	0.04	13.66	0.00
Never Married	0.52	0.10	5.45	0.00
High School	0.59	0.06	10.20	0.00
Some College	1.07	0.06	17.44	0.00
College	1.61	0.06	25.97	0.00
White	-0.03	0.07	-0.47	0.64
Black	-0.33	0.09	-3.82	0.00
Other	-0.02	0.14	-0.13	0.90
Years Since 2000	-0.25	0.02	-14.20	0.00
Married x Years Since 2000	0.05	0.01	6.22	0.00
Never Married x Years Since 2000	-0.06	0.02	-2.98	0.00
High School x Years Since 2000	0.05	0.01	3.51	0.00
Some College x Years Since 2000	0.03	0.01	2.01	0.04
College x Years Since 2000	0.07	0.01	5.06	0.00
White x Years Since 2000	0.14	0.02	8.48	0.00
Black x Years Since 2000	-0.04	0.02	-1.91	0.06
Other x Years Since 2000	0.06	0.03	1.83	0.07
Intercept	-1.58	0.07	-21.42	0.00

Each sample estimate above was subjected to a Wald chi-square test; all fixed effect estimates and interaction estimates proved to be statistically significant. We then examined the working correlation matrix (Table 7, below). It is apparent that the correlations weaken as time progresses; this could have an effect on the standard errors observed in Table 5. As a result, we estimated the model with a first-order autoregressive structure (Table 6, below). This improved the model fit, as evidenced by the lower quasi-information criterion (QIC) (Table 8, below). It

also changes some of the interpretations: White is now a significant predictor, as is the interaction between Other races and the time variable. The interaction between Some College and time is no longer significant. However, the statistical significance of the sample estimates, apparent from the Wald Tests, remains unchanged.

Table 6. Generalized Estimating Equation with First-Order Autoregressive Correlation Matrix

Top Quartile	Coefficient	Standard Error	z	P>|z|
Married	0.56	0.04	14.16	0.00
Never Married	0.66	0.10	6.89	0.00
High School	0.61	0.06	10.58	0.00
Some College	1.11	0.06	18.21	0.00
College	1.63	0.06	26.50	0.00
White	-0.16	0.07	-2.26	0.02
Black	-0.34	0.08	-4.06	0.00
Other	-0.06	0.14	-0.42	0.68
Years Since 2000	-0.30	0.02	-14.72	0.00
Married x Years Since 2000	0.05	0.01	5.45	0.00
Never Married x Years Since 2000	-0.10	0.02	-4.22	0.00
High School x Years Since 2000	0.05	0.01	3.28	0.00
Some College x Years Since 2000	0.03	0.02	1.71	0.09
College x Years Since 2000	0.07	0.02	4.63	0.00
White x Years Since 2000	0.18	0.02	9.45	0.00
Black x Years Since 2000	-0.04	0.02	-1.83	0.07
Other x Years Since 2000	0.07	0.04	2.05	0.04
Intercept	-1.44	0.07	-19.93	0.00

Table 7. Working Correlation Matrices (Left – Unstructured; Right – First-Order Autoregressive)

	c1	c2	c3	c4
r1	1.00			
r2	0.72	1.00		
r3	0.29	0.35	1.00	
r4	0.26	0.32	0.50	1.00

	c1	c2	c3	c4
r1	1.00			
r2	0.52	1.00		
r3	0.27	0.52	1.00	
r4	0.13	0.27	0.52	1.00

Table 8. QIC Comparison

	p	Trace	QIC	QIC_u
Unstructured	18	25.642	55522.372	55507.088
Autoregressive	18	26.748	55451.963	55434.468

Discussion

Both models were statistically significant. These results indicate that marital status, education level, and race all have significant effects on total assets. These demographic variables also influence the likelihood that a household belongs to the top quartile of assets among America's over-fifty population. This is consistent with the extant literature on wealth, as well as our hypotheses. Based on the the fact that the error terms of the marginal distribution are not normally distributed, future studies should attempt to find a more suitable model that results in normally distributed error terms. We suggest a penalized spline model, though this is not the only option.

There are limitations to this study that must be addressed. First is the design of this study, which is narrow and exploratory in nature. While not a drawback per se, there is room for improvement. A comprehensive consideration of all available independent variables would certainly add to this research in the future. This research also opens to door to examine trends in liabilities as a component of wealth; this is not a drawback, rather an avenue for future research. A complete picture of household equity would certainly inform the discussion around America's aging population and national retirement. A second limitation is the failure to use sample weights. The HRS data provides two weights for each cohort: household weight and respondent weight. These sample weights account for differences in selection probabilities, a byproduct of the multi-stage area probability sample design. Each HRS wave has different sample weights that should be accounted for in future studies.

References

Altonji, J. & Doraszelski, U. (2005). The role of permanent income and demographics in Black/White differences in wealth. *The Journal of Human Resources.*

Carroll, C., Otsuka, M., & Slacalek, J. (2011). How large are housing and financial wealth effects? A new approach. *Journal of Money, Credit, and Banking, 43*(1), 55-79.

Hajat, A., Kaufman, J., Rose, K., Siddiqi, A., & Thomas, J. (2010). Long-term effects of wealth on mortality and self-rated health status. *American Journal of Epidemiology, 173*(2), 192-200.

HRS. (2017). About | The Health and Retirement Study.

HRS. (2008). Sample Evolution: 1992-1998.

HRS. (2017). Samples Sizes and Response Rates.

Karagiannaki, E. (2012). The effect of parental wealth on children's outcomes in early adulthood. *Centre for Analysis of Social Exclusion | London School of Economics.*

Oliver, M., & Shapiro T. (2006). *Black wealth, white wealth*. New York, NY: Routledge.

Piketty, T. (2014). *Capital in the twenty-first century*. United States: President and Fellows of Harvard College.

Pollack, C., Cubbin, C., Sania, A., Hayward, M., Vallone, D., Flaherty, B., & Braveman, P. (2013). Do wealth disparities contribute to health disparities within racial/ethnic groups? *Journal of Epidemiology and Community Health, 67*(5), 439-445.

Smith, J. (1995). Racial and ethnic differences in wealth in the Health and Retirement Study. *Journal of Human Resources, 30,* 158-183.

Taylor, P., Kochlar, R., Fry, R. (2011). *Wealth gaps rise to record highs between whites, blacks and hispanics*. Washington, DC: Pew Research Center.

Wilmoth, J., & Koso, G. (2002). Does marital history matter? Marital status and wealth outcomes among preretirement adults. *Journal of Marriage and Family, 64*(1), 254-268.

Woolf, S., Aron, L., Dubay, L., Simon, S., Zimmerman, E., & Luk, K. (2015). How are income and wealth linked to health and longevity? *The Urban Institute.*

Wu, S., Asher, A., Meyricke, R., & Thorp, S. (2014). Age pensioner profiles: A longitudinal study of income, assets and decumulation. *ARC Centre of Excellence in Population Ageing Research.*

Yamokosi, A., & Keister, L. (2006). The wealth of single women: Marital status and parenthood in the asset accumulation of young Baby Boomers in the United States. *Feminist Economics*, *12*(1-2), 167-194.

Appendix A

Figure 1. Years versus Total Assets by Marital Status

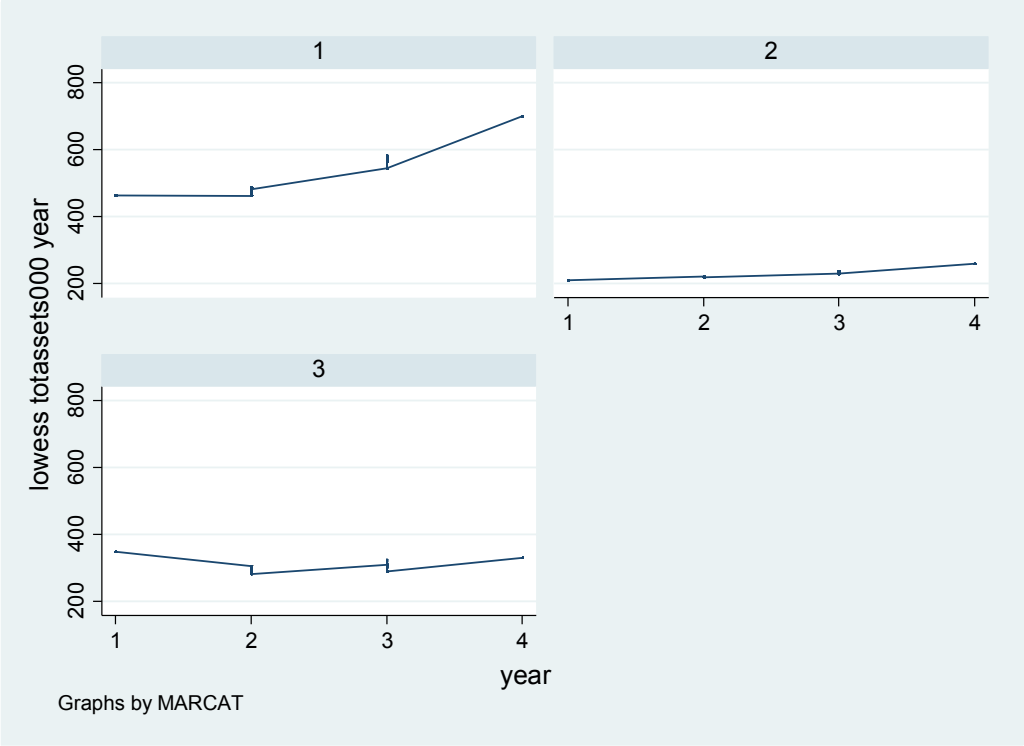


Figure 2. Years versus Total Assets by Education Level

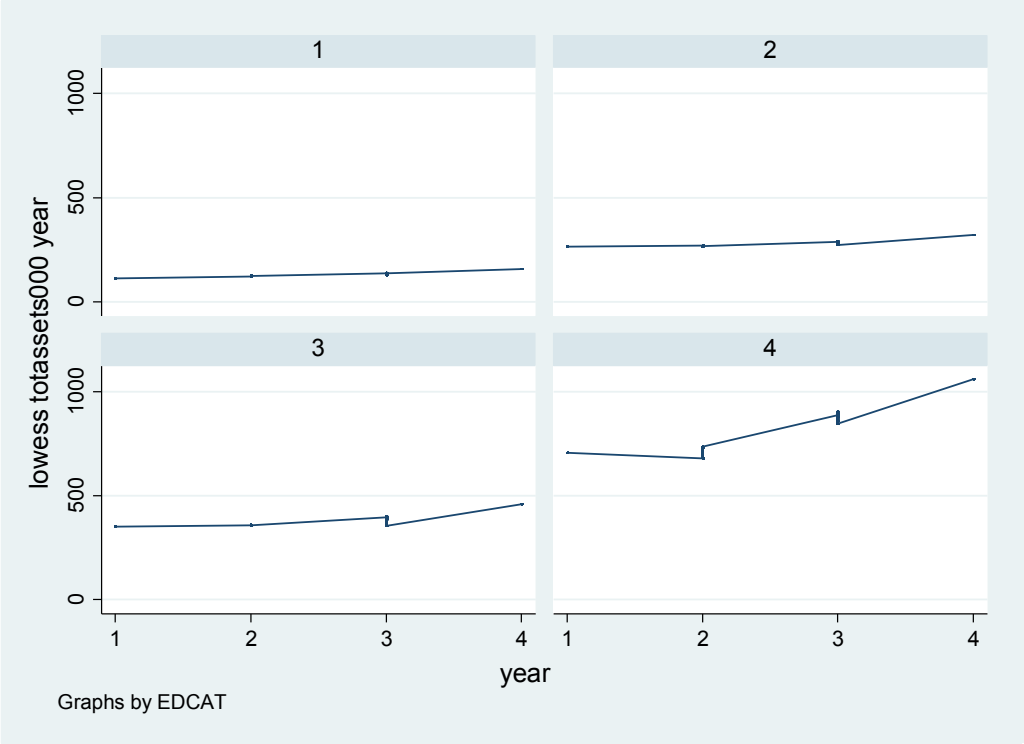


Figure 3. Years versus Total Assets by Race Category

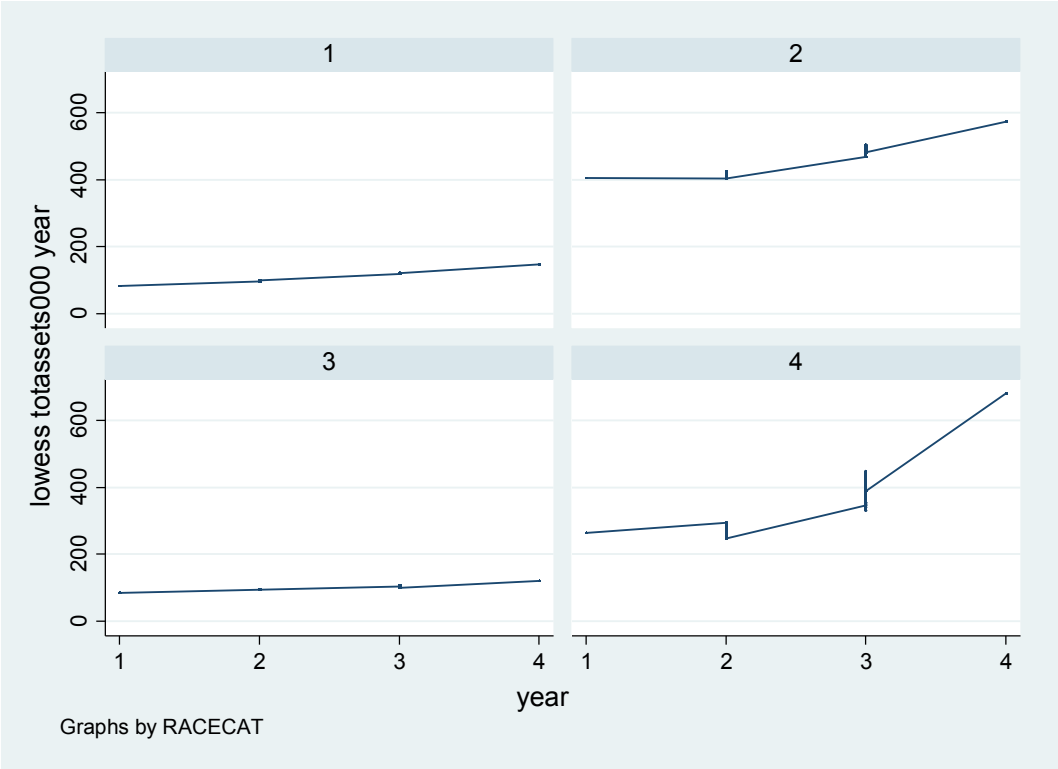


Figure 4. Years versus Total Assets without categorical restriction

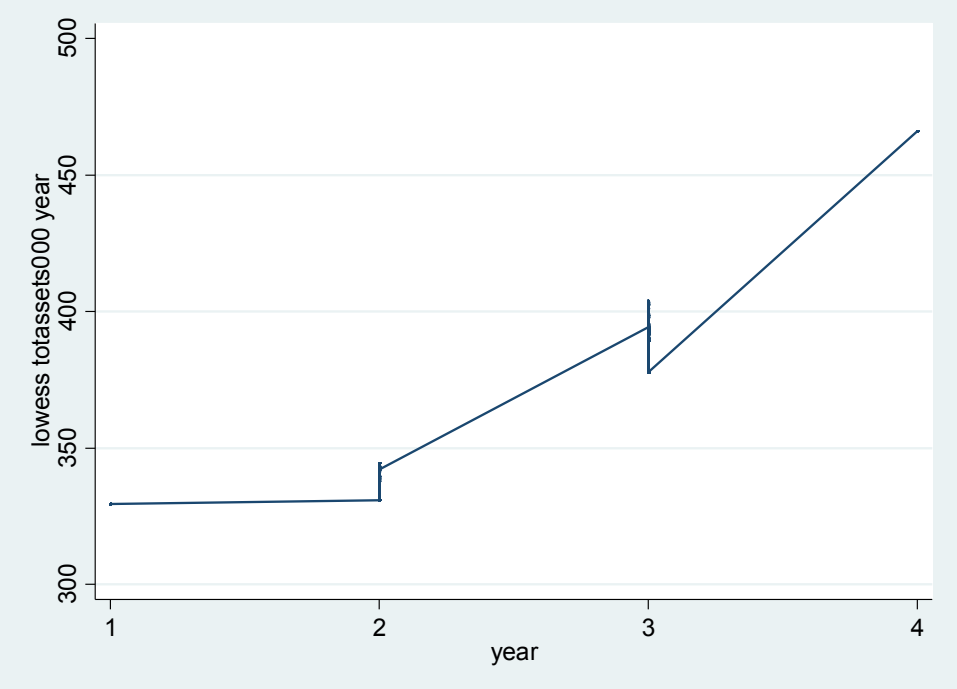


Figure 5. Year versus Proportion of Households in the Top Quartile of Total Assets by Marital Status

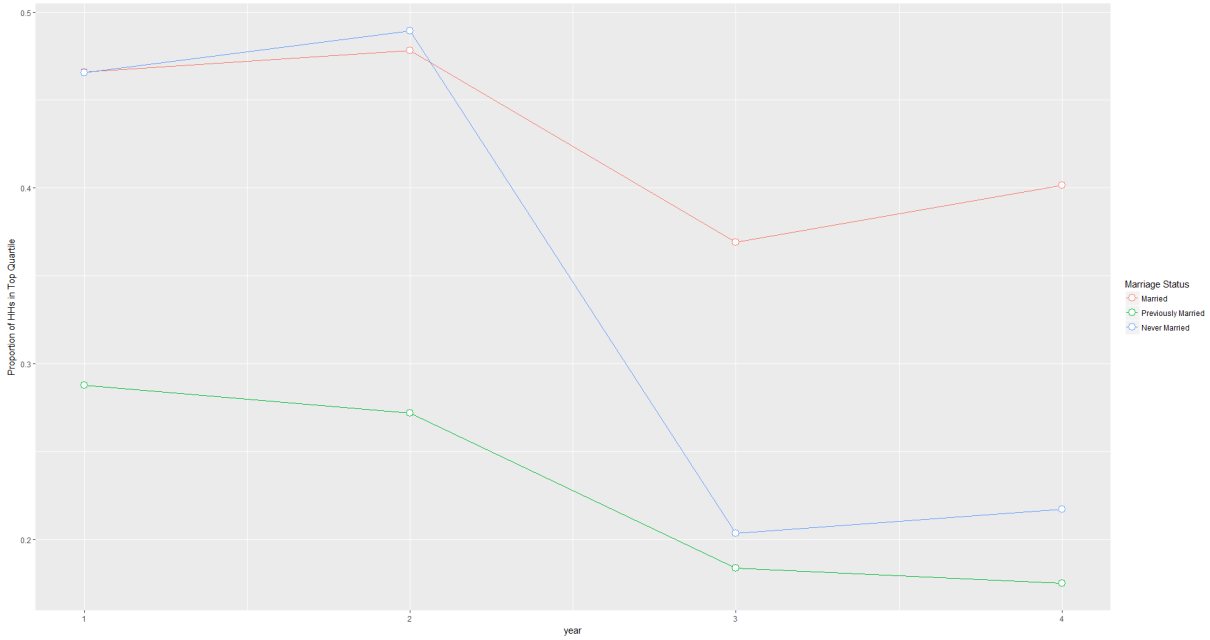


Figure 6. Year versus Proportion of Households in the Top Quartile of Total Assets by Education Level

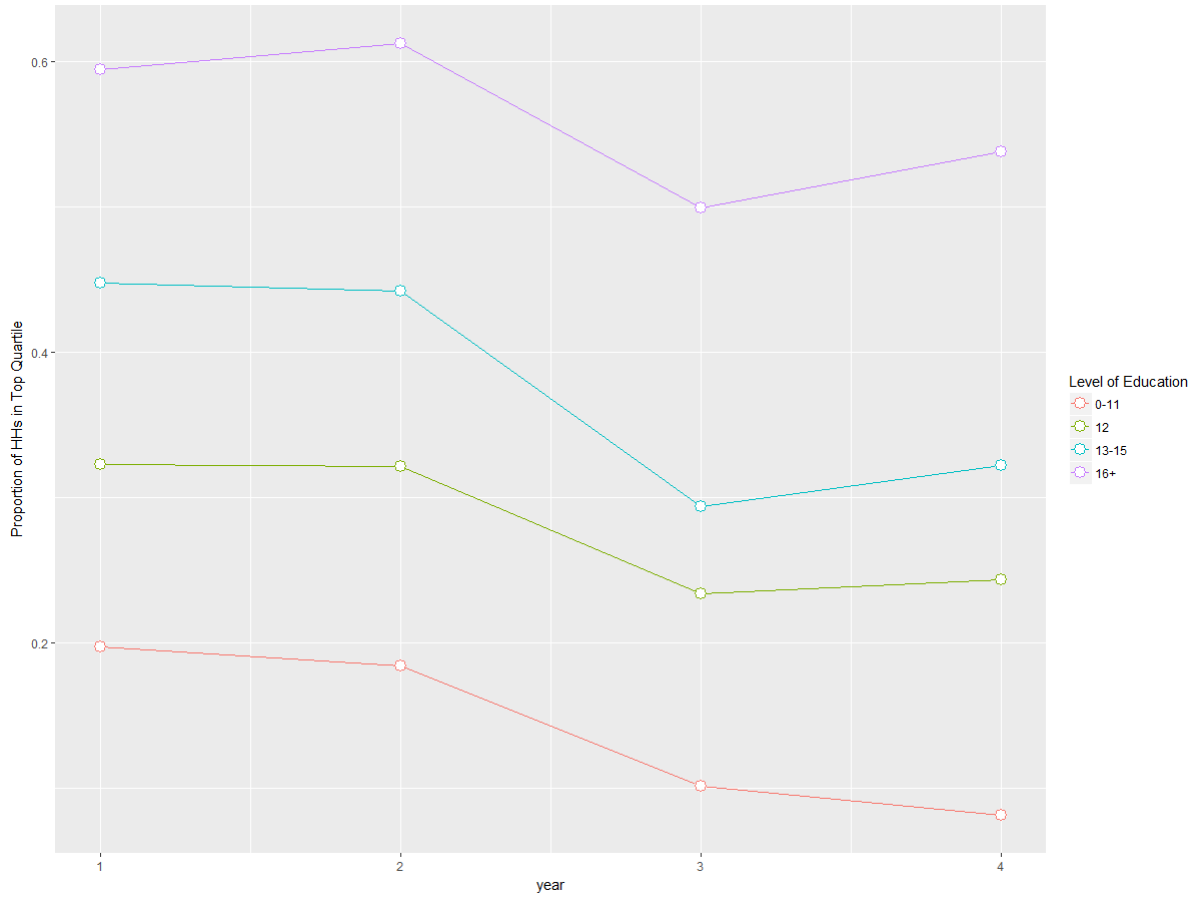


Figure 7. Year versus Proportion of Households in the Top Quartile of Total Assets by Race Category

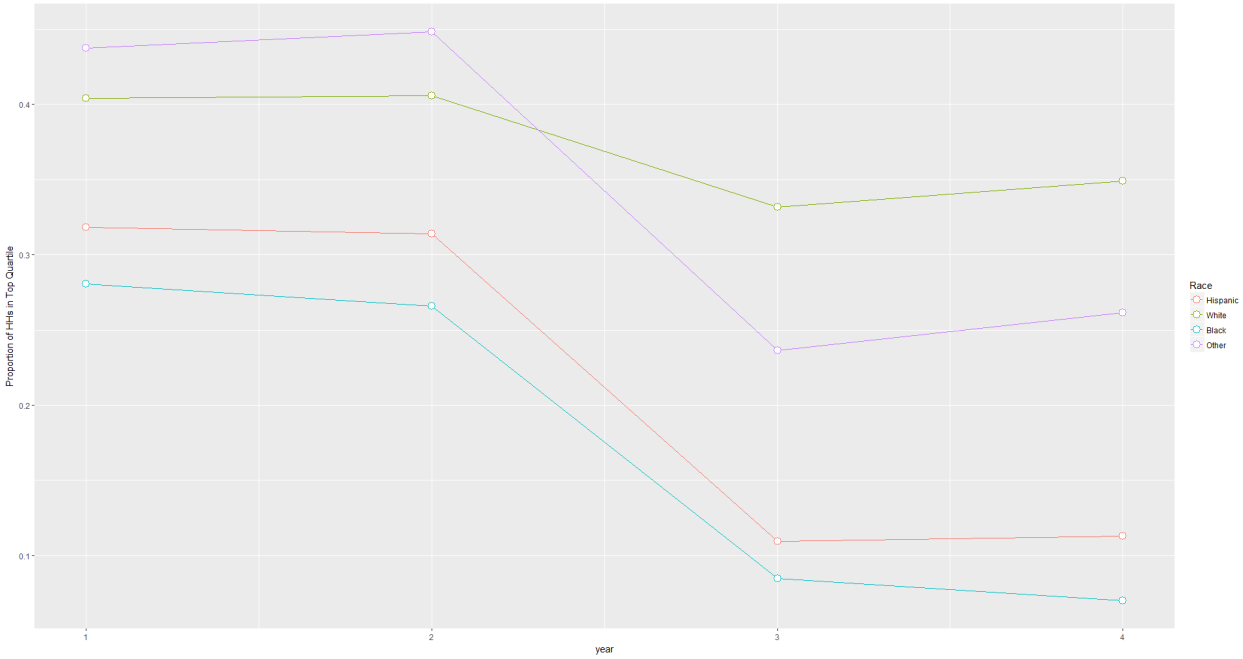
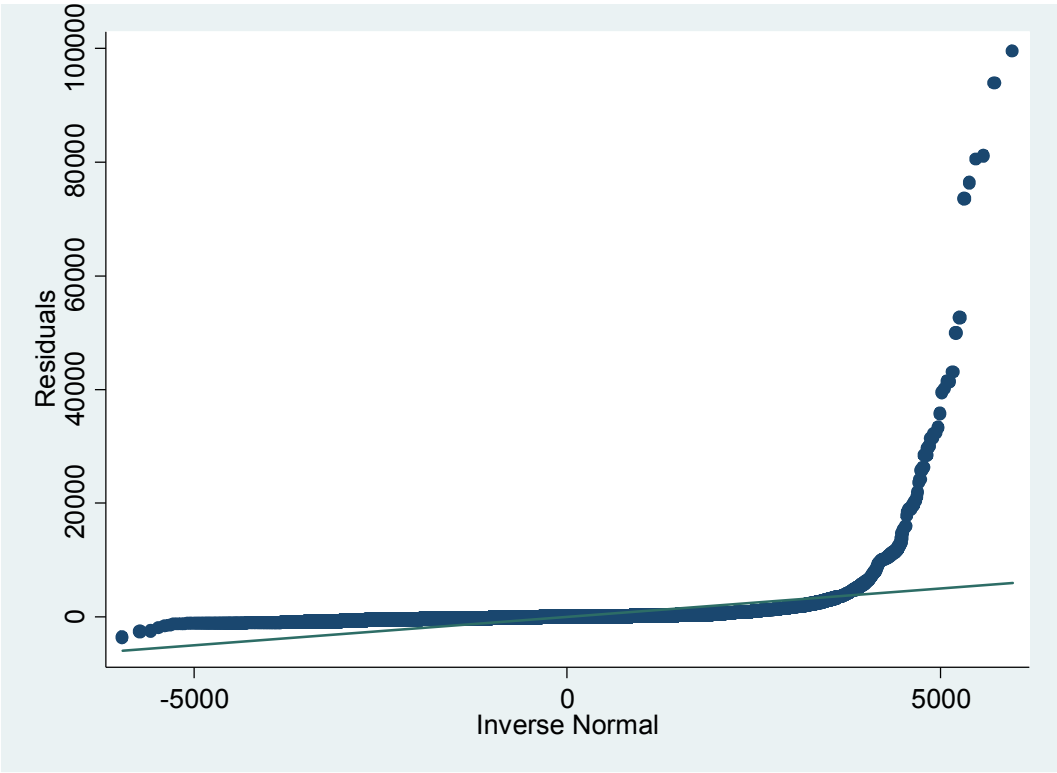


Figure 8. Residual QQ Plot of Total Assets



Appendix B

```
* Load Dataset
use "C:\Users\ariji\OneDrive\Documents\3rd Sem\Statistical
Modelling\Assignment4\hrs_reduced2.dta"

* Summarize dependent variable by year for descriptive statistics
tab yrssince00,sum( totassets000)

* Summary of demographic variables
table MARCAT EDCAT RACECAT,contents(freq)

* Lowess Plot of outcome variables- first wrt MARCAT (1= Married; 2 = Previously Married;
3= Never Married)
twoway (lowess totassets000 year), by( MARCAT)

* Lowess Plot of outcome variable – by EDCAT; 1= Less than H.S. 2= High School.3 =
Attended College. 4= College Graduate and above
twoway (lowess totassets000 year), by( EDCAT)

* Lowess Plot of outcome variable, by RACECAT; 1=Hispanic, 2=White, 3=Black, 4=Other
twoway (lowess totassets000 year), by( RACECAT)

* Here is a Lowessoutput plot without categorical restrictions
twoway( lowess totassets000 year)

* Here are the relevant Sphagetti Plots that uses all observations, but are less informative due to
clutter.
twoway (line totassets000 year), by( MARCAT)
twoway (line totassets000 year), by( EDCAT)
twoway (line totassets000 year), by( RACECAT)
twoway (line totassets000 year)

* Model fitting: our first objective is to find the best possible overall model for the trend in
growth of assets for each demographic category (Marital Status, Education, and Race)

* We fit the model using correlated errors using an unstructured variance-covariance matrix but
we do not care about random effects as we are not interested in trends for a given household
mixed totassets000 ib2.MARCAT ib1.EDCAT ib1.RACECAT ib0.yrssince00
ib2.MARCAT#ib0.yrssince00 ib1.EDCAT#ib0.yrssince00 ib1.RACECAT#ib0.yrssince00||
HHIDPN: , noconst residuals(unstructured,t(year)) variance reml

* Mimicking the Type-III tests we find in SAS to check for overall importance of fixed effects
(note that the testing specification is different, Stata runs Wald Chi-square test, and SAS runs a
F-test)

test 2.yrssince00 4.yrssince00 6.yrssince00
```

```
* Now for marital status
test 1.MARCAT 3.MARCAT

* Now for education
test 2.EDCAT 3.EDCAT 4.EDCAT

* Finally, RACECAT
test 2.RACECAT 3.RACECAT 4.RACECAT

* Next we test for the interaction terms – First, Year and Marital status
test 2.yrssince00#1.MARCAT 2.yrssince00#3.MARCAT 4.yrssince00#1.MARCAT
4.yrssince00#3.MARCAT 6.yrssince00#1.MARCAT 6.yrssince00#3.MARCAT

* Now Year and Education interaction
test 2.yrssince00#2.EDCAT 2.yrssince00#3.EDCAT 2.yrssince00#4.EDCAT
4.yrssince00#2.EDCAT 4.yrssince00#3.EDCAT 4.yrssince00#4.EDCAT 6.yrssince00#2.EDCAT
6.yrssince00#3.EDCAT 6.yrssince00#4.EDCAT

* Now Year and Race interaction
test 2.yrssince00#2.RACECAT 2.yrssince00#3.RACECAT 2.yrssince00#4.RACECAT
4.yrssince00#2.RACECAT 4.yrssince00#3.RACECAT 4.yrssince00#4.RACECAT
6.yrssince00#2.RACECAT 6.yrssince00#3.RACECAT 6.yrssince00#4.RACECAT

* Calculation of residuals and generating Q-Q Plot
predict resid, residuals
qnorm resid

* Now we define a top 25% group and a bottom 75% group. First we need to figure out the
quartile values
summarize totassets000, detail

* Clearly the 75% value is 384. So we create binary variable which takes the value 1 if asset
value is above 384 and takes the value 0 if asset value is below or equal to 384.
gen top25=0
replace top25=1 if totassets000>384
* To meet our second research objective – i.e. estimating whether differences being in the TOP
25 for particular demographic groups are changing over time, we estimate a GEE model suited
for binary outcome

* First we have to plot subgroup proportions for the top quartiles. This is done way easily in R
compared to Stata

# Loading Stata file
library(haven)
hrs_reduced2 <- read_dta("hrs_reduced2.dta")
# Generating aggregated datafiles, so we get proportions for each wave and
```

```
demographic category
aggdata <- aggregate(hrs_reduced2$top25, by=list(hrs_reduced2$EDCAT,hrs_reduced2$year),
FUN=mean, na.rm=TRUE)
colnames(aggdata) <- c("edcat","year","prop")
aggdata2 <- aggregate(hrs_reduced2$top25,
by=list(hrs_reduced2$MARCAT,hrs_reduced2$year), FUN=mean, na.rm=TRUE)
colnames(aggdata2) <- c("marcat","year","prop")
aggdata3 <- aggregate(hrs_reduced2$top25,
by=list(hrs_reduced2$RACECAT,hrs_reduced2$year), FUN=mean, na.rm=TRUE)
colnames(aggdata3) <- c("racecat","year","prop")

* Back to Stata. First we have to declare the dataset as panel
xtset HHIDPN yrssince00, delta(2)

* Now we run the GEE specification
xtgee top25 ib2.MARCAT ib1.EDCAT ib1.RACECAT yrssince00 ib2.MARCAT#c.yrssince00
ib1.EDCAT#c.yrssince00 ib1.RACECAT#c.yrssince00, family (binomial 1) link (logit)
corr(unstructured)

* Results from Wald tests
test 1.MARCAT 3.MARCAT
test 2.EDCAT 3.EDCAT 4.EDCAT
test 2.RACECAT 3.RACECAT 4.RACECAT
test c.yrssince00#1.MARCAT c.yrssince00#3.MARCAT
test c.yrssince00#2.EDCAT c.yrssince00#3.EDCAT c.yrssince00#4.EDCAT
test c.yrssince00#2.RACECAT c.yrssince00#3.RACECAT c.yrssince00#4.RACECAT

* Now we look at the working correlation matrix
estat wcorrelation

* We fit the model using a AR1 structure
xtgee top25 ib2.MARCAT ib1.EDCAT ib1.RACECAT yrssince00 ib2.MARCAT#c.yrssince00
ib1.EDCAT#c.yrssince00 ib1.RACECAT#c.yrssince00, family (binomial 1) link (logit) corr(ar1)

*We look at the working correlation matrix
estat wcorrelation

*Results from Wald Tests
test 1.MARCAT 3.MARCAT
test 2.EDCAT 3.EDCAT 4.EDCAT
test 2.RACECAT 3.RACECAT 4.RACECAT
test c.yrssince00#1.MARCAT c.yrssince00#3.MARCAT
test c.yrssince00#2.EDCAT c.yrssince00#3.EDCAT c.yrssince00#4.EDCAT
test c.yrssince00#2.RACECAT c.yrssince00#3.RACECAT c.yrssince00#4.RACECAT
```

* We now need to compare the QICs. This can be done by using a third party QIC Stata package
ssc install qic

* This package is old and requires all interactions and factor variable values to be classified into their own variables before they can be put into the syntax
gen MARCAT1=1.MARCAT
gen MARCAT3=3.MARCAT
gen EDCAT2=2.EDCAT
gen EDCAT3=3.EDCAT
gen EDCAT4=4.EDCAT
gen RACECAT2=2.RACECAT
gen RACECAT3=3.RACECAT
gen RACECAT4=4.RACECAT
gen M1Y= 1.MARCAT#c.yrssince00
gen M3Y=3.MARCAT# c.yrssince00
gen E2Y=2.EDCAT# c.yrssince00
gen E3Y=3.EDCAT# c.yrssince00
gen E4Y=4.EDCAT# c.yrssince00
gen R2Y=2.RACECAT# c.yrssince000
gen R3Y=3.RACECAT# c.yrssince00
gen R4Y=4.RACECAT# c.yrssince00

* First the QIC for the unstructured model
qic top25 MARCAT1 MARCAT3 EDCAT2 EDCAT3 EDCAT4 RACECAT2 RACECAT3 RACECAT4 yrssince00 M1Y M3Y R2Y R3Y R4Y E2Y E3Y E4Y, i(HHIDPN) t(yrssince00) family (binomial 1) link (logit) corr(unstructured)

* Then the QIC for the AR1 model
qic top25 MARCAT1 MARCAT3 EDCAT2 EDCAT3 EDCAT4 RACECAT2 RACECAT3 RACECAT4 yrssince00 M1Y M3Y R2Y R3Y R4Y E2Y E3Y E4Y, i(HHIDPN) t(yrssince00) family (binomial 1) link (logit) corr(ar1)