TV hours across generations

Godsee Joy, Karanveer Lamba, Deric Liang

Agenda

1. Introduction (1 min)
2. Data Wrangling and Feature Selection (2 min)
3. EDA Summary (3 min)
4. Model Building (3 min)
5. Conclusion and Limitations (1 min)

1) Introduction

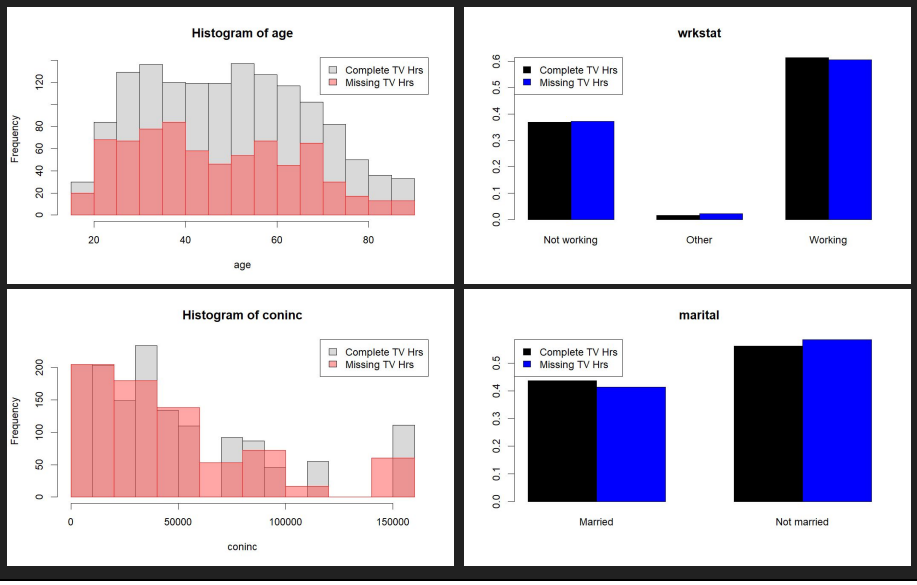
- Research Question: Does a person's age generation relate to the number of hours spent watching TV after accounting for other factors relating to TV hours?
- Dataset: 2018 General Social Survey
- Literature:
  - Older adults watch more TV - adults 65+ threefold more TV time than young adults
  - Effect size relates to living alone
  - Decline of TV hours for younger folks
  - Higher income and education levels negatively affect TV watch time

- GSS started 1972 - mostly biannually conducted, latest data is 2022
- Used by reputable sources like Associated Press
- In person interviewing going to households
  - The sample is a multi-stage area probability sample to the block or segment level. At the block level, however, quota sampling is used with quotas based on sex, age, and employment status.
  - They've iterated their approach over the years as well
  - IID not a concern with this data
- Literature: concerns here for health

2) Data Wrangling and Feature Selection

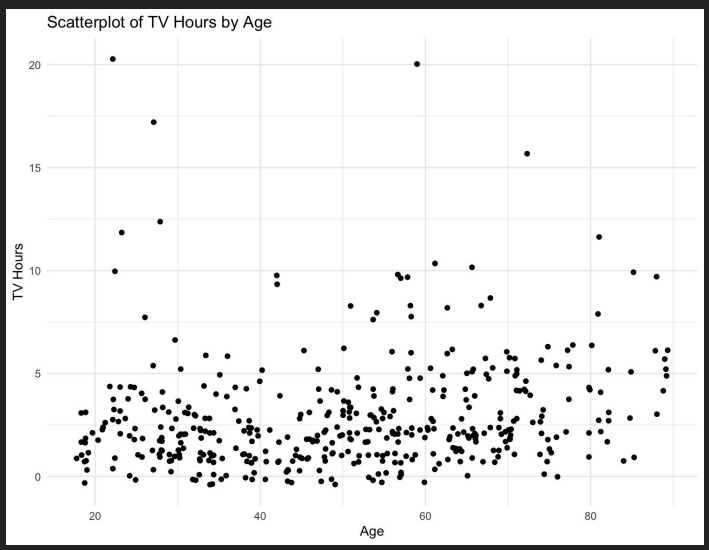
- Features
  - Age
  - Liu, Xiangyu study
    - Education
    - Income
  - Time constraints
    - Marital status
    - Employment status
- 2,348 respondents -> 1,421 observations removing missing values
- Transformations
  - Age -> Generations per research question
  - Log of Income per heavy right-skew
  - Consolidated marital status & employment status

2) Data Wrangling and Feature Selection



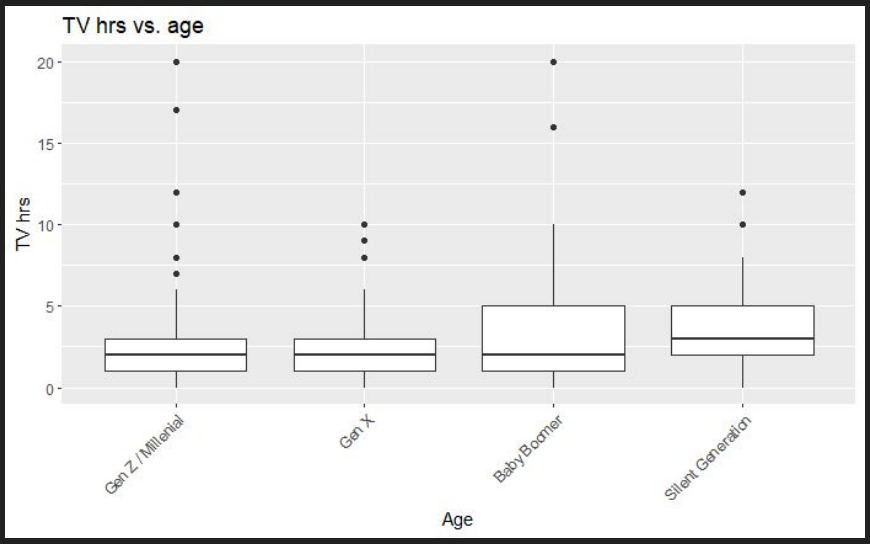
- About a third of our data was missing for tv hours but check the distributions across our variables of interest and saw roughly similar distributions across all
- Still left with around 1400 samples, sufficiently large sample size

3) EDA Summary



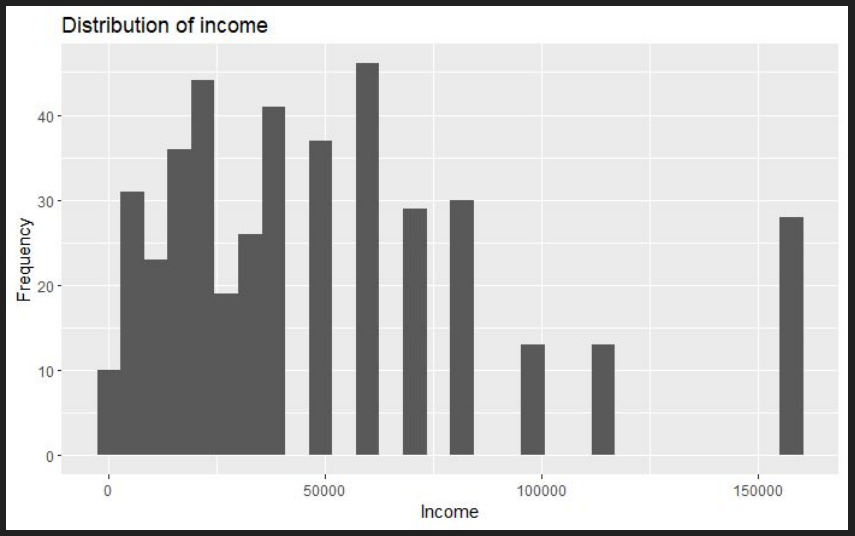
- We split the dataset into test and train and all EDA was done on the train dataset
- When we looked at raw age variable against tv hours, there was not a clear linear relationship
- We thought perhaps there could be age group differences that are less easy to see and interpret when the data is disaggregated
- While some of the values above 15 hours seem like potential outliers, when we looked at those responses across other variables, there were not clear indicators of these being response errors so kept them in
- Additionally, for those aged 89 or above, the survey coded those respondents into one group “89 or above” making it difficult to use this as a metric variable
- This all led us to think about how we could group age into categories for analysis

3) EDA Summary



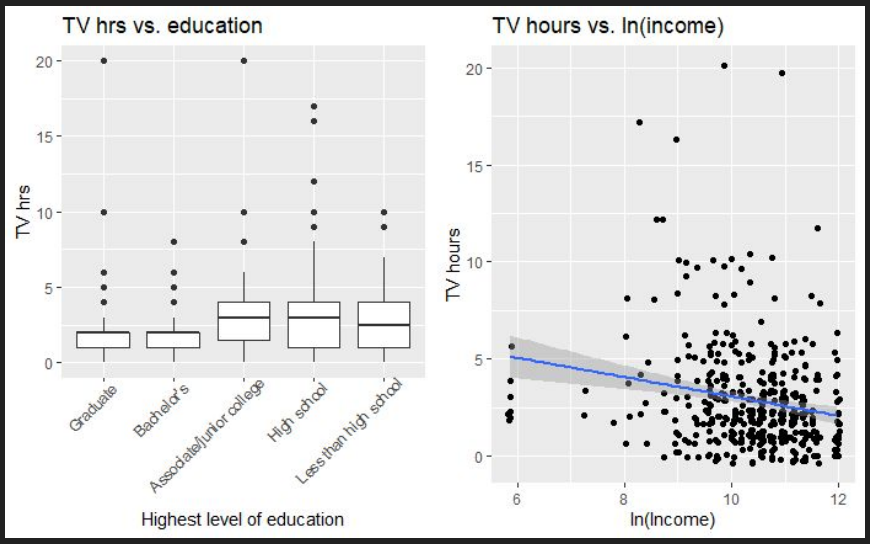
- We looked at creating age groups by decades and also generations
- We ended up sticking with generational groups to keep sample sizes within each category larger and for easier interpretation
- Here we see what was masked in the previous plot, in that:
  - Median # of hours of tv watched are higher for the Baby Boomer and Silent Generation than younger generations
  - Here we see more clearly potential outliers in the responses that exceed the 4th quartile. Though we decided not to remove these, due to lack of clear evidence of them being respondent errors,
  - Survey itself is not clear about how tv watching is defined - e.g., actively vs passively, or on a tv vs cell vs tablet
  - GSS conducts this in person going to homes - even more less likely that these are sampling errors or drawn from different populations
  - Additionally, one paper from the literature found that the avg American spends 2.5 hours watching TV a day and our avg for our training set was 2.9 hours, so even with these higher values, the mean overall is not affected too drastically

3) EDA Summary



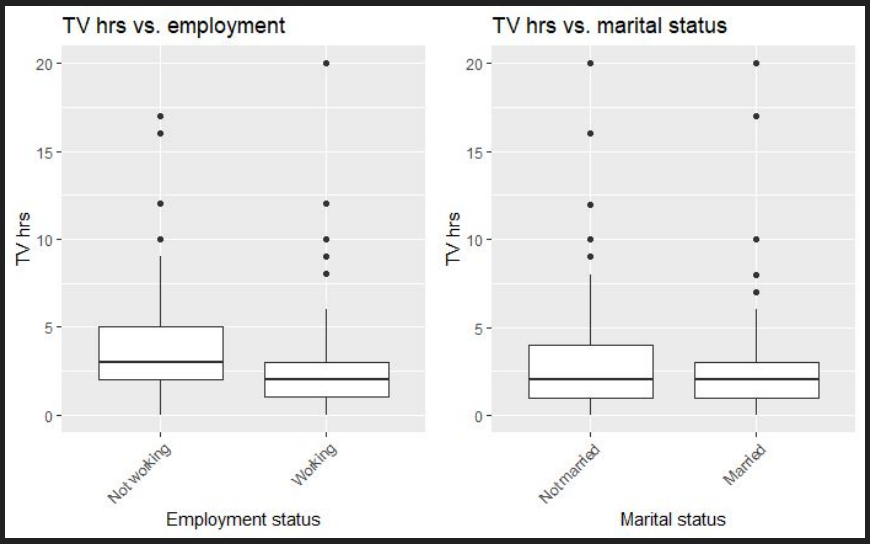
- The distribution of income is right skewed
- To handle the skewed distribution we decided to log the income variable for the purpose of the regression analysis
- Typically economics literature will also log transform income due to its prominent skew
- Income here is measure as inflation-adjusted family income (income in constant dollars)

3) EDA Summary



- With these plots we observed the following:
  - Fewer hours of TV watched as the level of education increases
  - Fewer hours of TV watched as the income increases
- This aligns generally with literature as well

3) EDA Summary



- These two were from our group's hypotheses in terms of measuring possible constraints on time
- We see that not working respondents watch more tv than working respondents
- For marital status, we see that the middle 50% for the married is more tightly distributed than the not married group
    - This is not looking at whether individuals live together or not though, which could further explain why there are not substantial differences between these groups

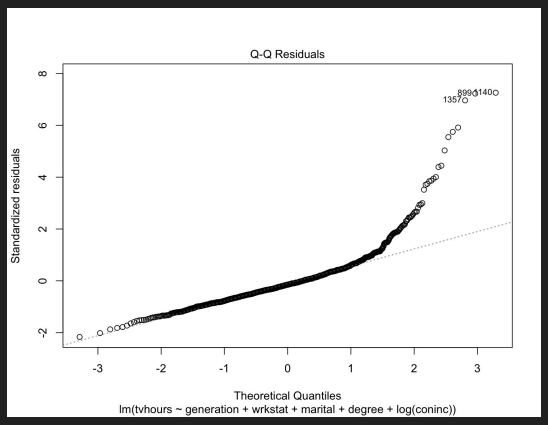
4) Model Building

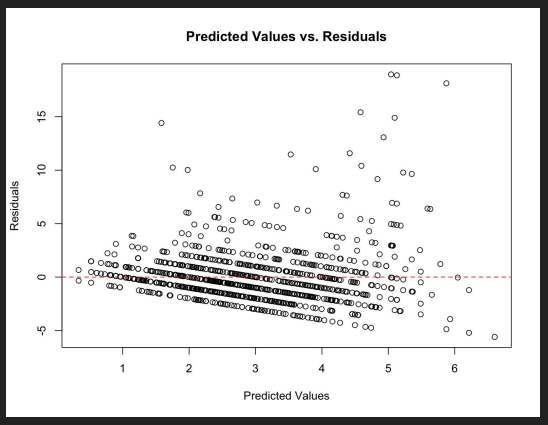
Table 1: Regression models for age on TV hours per day

	Dependent variable:					
	TV hours					
	(1)	(2)	(3)	(4)	(5)	(6)
Age	0.034***					
	(0.005)					
Generation-Gen X		0.242	0.251	0.362*	0.448**	0.560***
		(0.193)	(0.190)	(0.194)	(0.197)	(0.196)
Generation-Baby Boomer		1.387***	1.071***	1.155***	1.244***	1.378***
		(0.247)	(0.234)	(0.238)	(0.237)	(0.242)
Generation-Silent		1.676***	0.911***	0.929***	0.963***	0.991***
		(0.310)	(0.333)	(0.331)	(0.327)	(0.327)
Not working			1.318***	1.292***	1.136***	0.898***
			(0.207)	(0.203)	(0.199)	(0.179)
Not married				0.661***	0.562***	0.284*
				(0.165)	(0.161)	(0.160)
Education-Bachelor					0.250	0.188
					(0.202)	(0.204)
Education-Associate					1.155***	0.895**
					(0.366)	(0.379)
Education-High School					1.254***	0.936***
					(0.204)	(0.209)
Education-Less than High School					1.611***	1.065***
					(0.331)	(0.369)
ln(Income)						−0.432***
						(0.122)
Constant	1.251***	2.272***	1.927***	1.501***	0.599**	5.507***
	(0.254)	(0.142)	(0.143)	(0.186)	(0.238)	(1.364)
Observations	995	995	995	995	995	995
R^2	0.044	0.056	0.099	0.112	0.145	0.166
Adjusted R^2	0.043	0.053	0.095	0.107	0.138	0.157
Note:						*p<0.1; **p<0.05; ***p<0.01

- We started with our most basic model: tv hours = age
- And then added our generation variable
- We iteratively added covariates one at a time, assessing via F tests between the simpler and more complex models which additional variables were improving our model
- Directionally, the relationship between tv hours and income and education match what we saw in the literature
- For the generation coefficients, these also match trends from the literature:
    - Older generations have higher tv watch times on a statistically and practically significant standpoint
    - Assuming individuals are awake for 16 hours, not counting time for eating, self-care or working time, these coefficients translate to:
        - 3.5, 8.6 and 6.2 percent increases in waking tv hours, compared to Generation Z/ Millennials, all else being equal

4) Model Building





- Given our sufficiently large sample size, we did not need to check for the more strict CLM assumptions
- However, to additionally assess potential limitations in our model and goodness of fit, we ran a few diagnostic checks on the residuals
- We plotted a QQ plot of the residuals to assess normality and observed presence of outliers which are causing the residuals to deviate from normality
- Some evidence for heteroscedasticity?
    - We use robust standard errors to mitigate the effects of this on our statistical significance tests
- This aligns with findings from our EDA
    - We saw potential outliers for tv hours which are likely contributing to these deviations

5) Conclusions and Limitations

- Conclusions
  - Even accounting for potential constraints on time, such as working, higher incomes/education, and being married, we still observe generational differences in our data of # of tv hours watched per day (older generations watching more)
- Limitations
  - Older data - 2022 GSS data available but has high % of missing values
  - Small Gen Z sample - have to combine with Millenials
  - Anomalies in the family income variable
  - There may be other covariates which were not captured in the GSS dataset for e.g., time spent on outdoor activities, seasonality of viewership etc.

- Watching a lot of tv could be perceived as socially less optimal and respondents may under-report their true behavior
  - Experimental studies would better capture true behaviors or data from TV cable providers
- Seasonality:
  - Check date for 2018 survey - TV hours can vary by large events going on (winter olympics, sports events, news events, fifa world cup) - recency bias
- Directions for future research:
  - Qualitative studies to better understand the “why” behind these generational differences
  - Additional data like leisure time activities pursued (e.g, physically active respondents, other time intensive hobbies that cut across generations)