Introduction

This paper evaluates the relationship between self-reported health (SRH) conditions and socioeconomic variables (income, race, etc.). Self-reported health reflects one's standards for health, and several studies regard it as an accurate representation of one's health status (Baker 2001, as cited in Haron, 2010). Variance in self-reported health arises due to differences in medical history, genetics, or "inequalities in external factors in physical environment and living conditions" (Haron, 2010). Analyzing these external factors is crucial in creating equal opportunities for a healthy lifestyle.

A person's self-reported health provides more insight into "underlying physical and mental abnormalities" than other measures of health like body weight or blood pressure (Takahashi, 2020). For example, a low SRH is associated with mortality, morbidity, cancer, psychiatric diseases, and dementia (Takahashi, 2020). SRH provides meaningful insight as a risk factor for doctors and health specialists. Politically, SRH demonstrates the importance of welfare programs (TANF, SNAP, Medicaid) to equalize opportunities for health. This paper aims to determine the most important external factors affecting SRH to prioritize what reforms are needed to create a healthier society. We aim to understand the relevant significance of external factors on one's SRH by answering the following: How important are factors like income, education, gender, and race/ethnicity on one's self-reported health condition?

Theory/Conceptual Framework

We hypothesize that one's self-reported health is a function of the chosen exogenous variables modeled in Equation 1:

$SRH_i =$

$f(Income_i, Gender_i, Education_i, Education_{mother}, Education_{father}, Race_i, Ethnicity_i)$The variables are theoretically explained as follows:

- I (income) = As a person's income increases, they are more likely to afford appropriate health care, proper living conditions, sufficient nutrition, medicine, and other gains to better health. Depending on income level, these advantages or disadvantages can affect one's lifestyle, and, as a result, their perception of health.
- G (gender) = Females, traditionally care-givers for the family, may have a more critical eye for health and judge their status accordingly. Additionally, studies have shown women to be far more risk and feedback averse than men, specifically when making choices (Eckel, Grossman 2002). Thus, women may rate their health lower than it truly is. Additionally, males may rate themselves higher, given their proven tendency to be more overconfident than females of the same ability.
- E (education) = Education allows an individual to support themselves by acquiring a job and having income to protect their health. For someone with low levels of education, they may be unable to have a job and lack the ability to foster a healthy lifestyle. Also, high levels of schooling provide individuals with health knowledge, impacting SRH.
- R (race) = Racial injustices in society can impact the SRH of marginalized groups mainly through stress, physical harm, wage/job discrimination. Also, some physical diseases

impact specific races, such as sickle cell anemia in African-Americans or hepatitis B in Asian-Americans. While Asian-Americans make up 6% of the U.S. population, they account for 58% of the 862,000 Americans living with chronic hepatitis B (CDC, 2020). Race is also linked to a lack of access to healthcare or poor-quality healthcare (Funk, 2022).

- C (ethnicity) = Similar to race, ethnic variations may cause people in marginalized groups to experience discrimination which may manifest physically or mentally. These effects may lower one's self-reported health if they feel at risk based on discrimination.

Literature Review

Self-reported health (SRH) is a concept formed in adolescence that stays relatively consistent throughout young adulthood. The measure depends on people expressing their "internal feelings" and how they evolve (Cislaghi, 2009). Many factors correlate to higher SRH levels, such as religiosity in the Caribbean Islands and social capital in Portugal (Reyes-Ortiz, 2007; Silva, 2016). Some factors are especially relevant across studies in the United States and internationally.

Lunderberg (2008) found that education positively correlated to SRH rates and negatively correlated to the number of chronic illnesses of individuals. Haron (2010) similarly reported that 71% of his sample that reported tertiary education rated their SRH as "good," with a steep decline as years of schooling decreased.

Cislaghi (2019) found correlations between higher SRH and income levels across several countries. An article by Smith and Kington (1997) found that older individuals reporting excellent health had "2.5 times more income and five times more wealth" than people of the same age who reported poor health. Research suggests that higher-income people have access to more healthcare options, whereas "resource-deprived people" experience financial and racial barriers to systems that can improve their SRH. Such inequalities can result in social and emotional deprivation and lower overall SRH (Cislaghi, 2019). Haron (2010) also found that those that perceived their financial situation as "very bad" tended to report "poor" on their SRH. Haron's research concludes that income is a main dividing factor in SRH, but other factors such as race, age, and marital status are still important.

Brondolo (2011) found a statistically significant association between lifetime exposure to racism and overall SRH. Exposure to social exclusion and threats/harassment were the main contributors to poor SRH in her sample. Perceived racism remained significant when controlling for age, gender, education, and employment factors. Therefore, race as an endogenous variable may be more significant than prior research suggests.

Jatrana (2021) also found that gender influences SRH reporting, but not as much as previous research reported. Using data from the Survey of Family, Income, and Employment (SoFIE), Jatrana first regressed gender with self-reported psychological distress (gender differences in health). Then, she regressed those same variables, this time with socioeconomic factors. The results suggest that exposing social and economic factors to men and women might differ and that these factors mediate or explain the association of gender with health. In the first regression, the odds of reporting moderate-to-high levels of psychological distress were 1.23 (95% CI 1.14 to 1.32) times higher for women than for men in Model 1 (which controls for demographic factors). However, Model 2 adjusted for socioeconomic variables and reduced the gender odds ratio for predicting moderate-to-high levels of psychological distress (1.10; CI 1.02, 1.19). Thus, Jatrana contends that socioeconomic variables mediate just over half of the association of gender with self-reported psychological distress. Jatrana concludes that significant gender differences exist in self-assessed health, as women were more likely to report moderate-to-high psychological distress.

Data and Descriptive Analysis

The paper uses data from the General Social Survey, which reports the responses of domestic U.S. citizens to various survey questions. We selected a date range of 1975 to 2018 based on the data availability for the endogenous variables in the model.

The exogenous variable is self-reported health. The GSS asks the following question related to SRH: "Would you say your own health, in general, is excellent, good, fair, or poor?" (GSS). To convert to a numeric variable, we assigned the following values to each category: poor (1), fair (2), good (3), and excellent (4).

The independent variables are defined in the previous section. They are coded and described as follows:

1. Gender: Recoded as dummy female variable in Stata (1 = female, 0 = male)
2. Rincome for the previous year (of respondent):
    - LT $1000 = 1, $1000 TO 2999 = 2, $3000 TO 3999 = 3, $4000 TO 4999 = 4, $5000 TO 5999 = 5, $6000 TO 6999 = 6, $7000 TO 7999 = 7, $8000 TO 9999 = 8, $10000 – 14999 = 9, $15000 – 19999 = 10, $20000 – 24999 = 11, $25000 OR MORE = 12
3. Education (of respondent): years
4. Paternal education (respondent father's education): years
5. Maternal education (respondent mother's education): years
6. Race: black – 1, other – 2, white – 3
    - When used in the stat regression, i.race is used because it is a categorical variable.

7. Ethnicity: Included as fixed effects

    - Used xtset command to indicate fixed effects each unique ethnicity in survey

The most pertinent weakness of using the General Social Survey is the unavailability of endogenous variables of interest. Some variables we expected to be meaningful for SRH contained large amounts of missing values. We excluded them to avoid large sampling error.

Since the endogenous variables are categorical (besides the quantitative education variables), there is little relevance for outliers. Below, the paper addresses the most pertinent tabulations to describe the dataset quality and address limitations.

Based on the frequencies, nearly half (48.8%) of respondents self-reported their health to be good. Only a small proportion of the respondents reported poor or fair health, which is a limitation of the dataset since self-reported health ratings appear to be overall higher.

Table 1: Frequencies and Percent of Total for Each Level of Self-Reported Health

Health rating	Frequency	Percent
1 (poor)	572	2.15
2 (fair)	3,998	15.03
3 (good)	12,979	**48.79**
4 (excellent)	9,052	34.03
Total	26,601	100.0

According to Table 2 below, 67.7% of respondents are male, and 32.3% are female, representing a significant gender bias in the dataset. For race displayed in Table 3, 80% of respondents are white compared to 13.7% for black respondents. These biased frequencies in the dataset are a source of weakness as they cause the dataset to be more representative of certain groups.

Table 2: Frequencies and Percent of Total for Female Dummy Variable		
Female	Frequency	Percent
0 (no)	24,752	**67.70**
1 (yes)	11,808	32.30
Total	36,560	100.0
Table 3: Frequencies and Percent of Total for Each Category of i.Race		
Race category	Frequency	Percent
1 (black)	5,099	13.76
2 (other)	2,300	6.21
3 (white)	29,649	**80.03**
Total	37,048	100.0

The paper includes the respondent's education, father, and mother as variables in the regression. To verify the appropriateness of using three different measures, we obtained the VIF (variance inflation factor) to measure the correlation of each measure. The results are below in Table 4. Since the factors are close to 1, we are confident that the three measures are uncorrelated and can be included**.**

Table 4: Results from VIF Test on 3 Education Variables	
Education Variable	VIF
educ (respondent)	1.75
paeduc (father)	1.71
maeduc (mother)	1.25

According to Figure 1 below, the respondent income variable is left-skewed, meaning most respondents have high-income levels. While this is a weakness of the data, the original design of the survey question in the GSS is poor. The highest option for respondents was a yearly income of at least $25,000, which most people satisfied. As time progressed, more participants satisfied the $25,000 income bracket which led to our sample distribution having a left skew.

Based on Figure 2, there is a more significant proportion of respondents with 12–16 years of education. Thus, the dataset reflects a highly educated group of individuals, with the majority having at least a GED.

Based on Figure 2, there is greater proportion of respondents with 12–16-years of education. Thus, the dataset reflects a highly educated group of individuals with the majority having at least a GED.

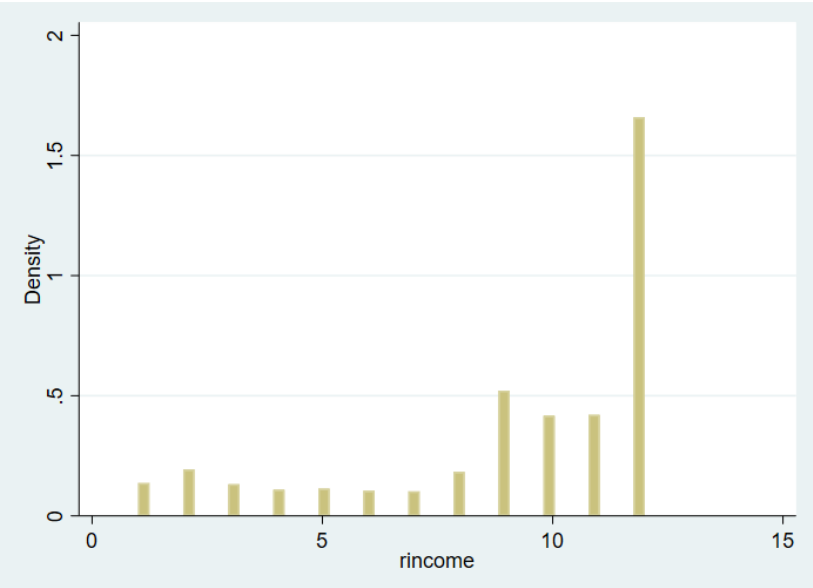

Figure 1: Density Curve for Respondent Income

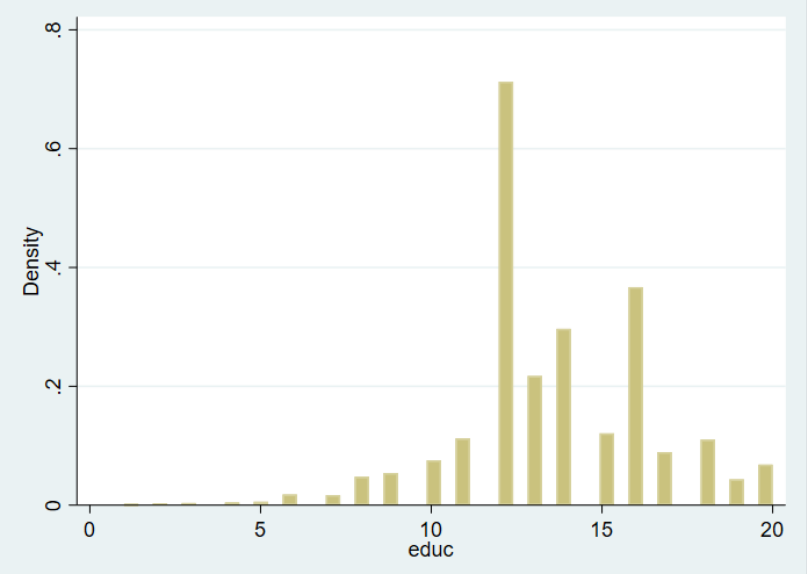

Figure 2: Density Curve for Respondent Education

1975-2018 Regression Discussion

Table 1 below contains the results from the preliminary regression for the full period of years 1975-2018. The coefficients for each right-hand-side variable give the probability of having good self-reported health ratings.

Factors of SRH	
	(1)
VARIABLES	Self Reported Health
rincome	0.005***
	[0.002]
maeduc	0.008***
	[0.002]
paeduc	0.008***
	[0.002]
educ	0.039***
	[0.002]
female	-0.086***
	[0.012]
race - other	-0.013
	[0.044]
race - white	0.064**
	[0.031]
Constant	2.387***
	[0.041]
Observations	17,630
Number of ethnicnumber	43
R-squared	0.039
Adj R-squared	0.0363
Standard errors in brackets	
*** p<0.01, ** p<0.05, * p<0.1	

Table 5: Results from the preliminary regression, full years 1975-2018

Based on the coefficient magnitude, being female has the largest impact on SRH, decreasing it by –0.086. Being white has the subsequent most considerable effect as it raises one's SRH in the dataset by 0.064, with the reference race as black. All variables produced statistically significant effects except when the respondent's race was neither black nor white.

Respondent education (educ) had the most tremendous statistical significance and yielded larger clinically relevant effects than the others. With a t-statistic of 16.97 (highest of variables), it is the most statistically significant variable in this regression. In gauging its clinical relevance, when education is increased by one standard deviation (3.08), SRH is increased by 0.12 (Table 6). The standard deviation of SRH is 0.84, so the education variable does have some clinical effect on increasing SRH.

Table 6: Effects for Each RHS Variable by Increasing One Standard Deviation	
Variable	Effect on SRH
Respondent income	0.0167
Respondent education	**0.12**
Mother's education	0.002
Father's education	0.027
Female	-0.040
Race (white)	0.024
Race (not white/black)	-0.00299

The following clinically relevant variable is being female, with the second highest t-statistic. The t-statistic for females is –7.28, which is less than half that of education but still greater than the next highest t-statistic. Therefore, the regression reveals the immense impact of being female in having lower SRH ratings. Given the literature discussed, this is expected since females tend to be more critical of their health.

1996 Regression

Initially, the paper contained an exercise variable in the regression model. If one has a healthier lifestyle in terms of fitness, they will likely have higher SRH. However, the General Social Survey only had data for respondent exercise for 1996. Given the suspected effect of exercise, we performed a singular regression on 1996 so that exercise could be included. The results are presented below in Table 7.

Factors of SRH	
	(1)
VARIABLES	Self Reported Health 1996
rincome	0.022**
	[0.010]
maeduc	0.023*
	[0.013]
paeduc	0.024**
	[0.009]
educ	0.041***
	[0.012]
exercise	0.158**
	[0.078]
female	-0.121*
	[0.065]
2.race	0.099
	[0.259]
3.race	-0.076
	[0.158]
Constant	1.920***
	[0.242]
Observations	554
Number of ethnicnumber	39
R-squared	0.111
Adj R-squared	0.0302
Standard errors in brackets	
*** p<0.01, ** p<0.05, * p<0.1	

Table 7: Results from 1996 regression, adding exercise as RHS variable

According to Table 7, exercise yields the largest coefficient (0.158) and has a statistically significant effect on self-reported health, which we expected. Being female also had a substantial

effect on lowering SRH with exercise controlled, based on its coefficient of –0.121 in Table 7. However, neither had clinically relevant effects since moving one standard deviation increased SRH by 0.05 (exercise) and decreased SRH by –0.05 (female). Since the standard deviation of SRH in 1996 was .788, neither of those effects is relevant enough to be considered clinically impactful.

Respondent education, income, and father's education are all statistically significant variables with exercise included in the regression. While income and father's education do not produce clinically significant effects (0.06 and 0.08 respectively), education does. Compared to the full 1975-2018 regression, education still has the highest t-statistic (3.29), lowest p-value (0.001), and most clinical relevance (0.115 increase). This demonstrates how important education is in effecting self-reported health, even when exercise is included in the regression.

Table 8 below contains the results of the 1996 regression without the exercise variable. This allows analysis of the effect of adding the exercise variable to the regression.

Factors of SRH	
	(1)
VARIABLES	Self Reported Health 1996
rincome	0.016**
	[0.007]
maeduc	0.021**
	[0.009]
paeduc	0.018***
	[0.007]
educ	0.044***
	[0.009]
female	-0.090**
	[0.046]
2.race	-0.059
	[0.178]
3.race	0.079
	[0.118]
Constant	1.930***
	[0.173]
Observations	1,124
Number of ethnicnumber	42
R-squared	0.085
Adj R-squared	0.0438

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Table 8: Results from 1996 regression, excluding exercise

Compared to the coefficients in Table 7, we concluded that the coefficients in 1996 changed minimally with the addition of exercise, representing their robustness. In Tables 7 and 8, the coefficient changes are each less than 0.05. Regarding a self-reported health score on a 1-4 rating, no change was enough to change clinical relevance. For example, the female variable had the most notable change of any variable as it decreased by 0.03, which is still very minimal.

Table 9: Coefficients, T-statistics, and P-values Before & After Exercise (1996)						
	Coefficient		T-statistic		P-value	
Variable	Before	After	Before	After	Before	After
Respondent income	0.015	0.022	2.24	2.18	0.025	0.030
Respondent education	**0.044**	**0.041**	**5.05**	**3.29**	**0.000**	**0.001**
Mother's education	0.020	0.023	2.48	1.80	0.013	0.072

Father’s education	0.018	0.024	2.75	2.53	0.006	0.012
Female	-0.090	-0.121	-1.97	-1.87	0.050	0.062

Table 9 displays the effects of adding exercise on three statistics, and excludes the race variables due to statistically insignificant results ($p > .05$). Regarding changes to the t-statistic, respondent education decreased the most, but that was still only by 1.76 (5.05 to 3.29). Despite its decrease, it is still the highest t-statistic, so even when controlling for exercise, education still has a critical impact. While all the p-values increase slightly and there is less statistical significance, most variables still have statistically significant effects. For example, in Table 9, the education p-value increased by 0.001, which is still extremely low. The only variables with notable changes to p-values are mother's education and female, as both became statistically insignificant. Education also continues to have the largest clinical effect (0.115). Moreover, the addition of exercise did not impact the clinical relevance of all variables.

Limitations and Future Exploration

Table 10 shows the r-squared value from the 1975-2018 regression and 1996 regression. For the main 1975-2018 regression, a 0.039 r-squared value is concerningly low. It demonstrates how little the variables chosen explain the variation in self-reported health scores. Therefore, our predicted ordinary least squares regression line has a tremendous variation of actual values.

Table 10: R-Squared Values for Each Regression Performed	
Regression	R-Squared
1975-2018	**0.039**
1996 w/o Exercise	0.085
1996 w/ Exercise	**0.111**

After exercise was added in 1996, the r-squared value increases, as Table 10 shows. While it is a minimal increase, it demonstrates the impact of exercise and justifies the inclusion of the variable in the regression. Since it increases the r-squared value, it allows our regression to explain more of the variance, which is a benefit.

Another limitation of our study is the lack of clinical significance. As Table 6 presents, there are only minor impacts after increasing each RHS variable by one standard deviation. Only the respondent education variable yielded a large enough effect to be relevant to the standard deviation of self-reported health, as discussed above. While all but one variable had statistical significance, it is a limitation that only one was clinically relevant.

The regression model omits several variables relevant to self-reported health, causing bias on the coefficients. Nutrition was unavailable in the General Social Survey but impacts self-reported health since a healthy diet reduces health complications. Additionally, the regression omits variables such as BMI measure or history of disease data as they had a large number of missing variables in the GSS. Nevertheless, if an individual is overweight or has a history of

disease, they likely have worse health and would self-report lower. We also suspect that if an individual is a doctor or health professional, they may be more critical of their health and self-report lower. Thus, an occupation dummy variable (doctor or not) would be insightful. Lastly, we believe the relative health of an individual's family to be influential. If an individual is relatively healthier compared to the people they live with, they may report higher than is true.

Conclusion

Our regression results determined that the main factors impacting SRH were education and gender. While all but one of the endogenous variables we looked at yielded statistically significant results, education was the only variable to yield clinically relevant results. Even when we added exercise, the variable education effect was still strong.

For future implications, this reveals the importance of education in rating one's health and how various inequalities in education worldwide can contribute to different health viewpoints. Our research highlights the importance of providing equal access to educational opportunities for students, regardless of background, as it may improve self-reported health ratings in the long term.

References

Brondolo, Elizabeth & Hausmann, Leslie & Jhalani, Juhee & Pencille, Melissa & Atencio-Bacayon, Jennifer & Kumar, Asha & Kwok, Jasmin & Ullah, Jahanara & Roth, Alan & Chen, Daniel & Crupi, Robert & Schwartz, Joseph. (2011). Dimensions of Perceived Racism and Self-Reported Health: Examination of Racial/Ethnic Differences and Potential Mediators. Annals of behavioral medicine : a publication of the Society of Behavioral Medicine. 42. 14-28. 10.1007/s12160-011-9265-1.

Centers for Disease Control and Prevention. (2020), "People Born Outside of the United States and Viral Hepatitis." https://www.cdc.gov/hepatitis/populations/Born-Outside-United-States.htm

Cislaghi, B., & Cislaghi, C. (2019). Self-rated health as a valid indicator for health-equity analyses: Evidence from the Italian health interview survey. *BMC Public Health*, *19*(1), 533. https://doi.org/10.1186/s12889-019-6839-5

Eckel, Catherine C., and Philip J. Grossman, "Sex and Risk: Experimental Evidence," in Handbook of Experimental Economics Results, C. Plott, and V. Smith, eds. (Amsterdam, The Netherlands: Elsevier Science B.V./North-Hol- land, forthcoming, 2002a).

Funk, Cary. "Black Americans' Views about Health Disparities, Experiences with Health Care." *Pew Research Center Science & Society*, Pew Research Center, 1 Mar. 2023, https://www.pewresearch.org/science/2022/04/07/black-americans-views-about-health-disparities-experiences-with-health-care/.

Haron, S. A., Sharpe, D. L., Masud, J., & Abdel-Ghany, M. (2010). Health Divide: Economic and Demographic Factors Associated with Self-Reported Health among Older Malaysians. Journal of Family and Economic Issues, 31(3), 328–337.

Lundborg. (2008). The Health Returns to Education: What Can We Learn from Twins? IDEAS Working Paper Series from RePEc.

Jatrana, Santosh. "Gender Differences in Self-Reported Health and Psychological Distress among New Zealand Adults." *Demographic Research*, vol. 45, 2021, pp. 693–726., https://doi.org/10.4054/demres.2021.45.21.

Reyes-Ortiz, Carlos A., Martha Pelaez, Harold G. Koenig, and Thomas Mulligan. "Religiosity and self-rated health among Latin American and Caribbean elders." *The International Journal of Psychiatry in Medicine* 37, no. 4 (2007): 425-443.

Silva, P. A. D. (2014). Individual and social determinants of self-rated health and well-being in the elderly population of Portugal. *Cadernos de saude publica*, *30*, 2387-2400.

Smith, J. P., & Kington, R. (1997). Race, socioeconomic status and health in late life. In L. Martin & B. Soldo (Eds.), Racial and ethnic differences in health of older Americans (pp. 106–162). Washington, DC: National Academy Press.