Racial Differences in Self-Reported Health Probabilities in the United States

Cameron Speakes
Department of Economics
Florida State University
Tallahassee, FL 32306-2180

Abstract

This study utilizes a logit model to show differences in the probabilities of self-reporting a higher health categorization for different racial groups in the Centers for Disease Control and Prevention's annual Behavioral Risk Factor Surveillance System from 2011 – 2016. It is demonstrated that minorities consistently have a lower probability of reporting a positive health status to the CDC over this period, yet these negative effects do seem to diminish across time.

1. INTRODUCTION

Identifying racial disparities and finding solutions to them is integral to the equitable development of a community, so economists and social scientists often seek to empirically show where these occur so that society may take corrective measures. An area of major study has been disparities in the healthcare industry, such as differences in infant mortality in hospitals persisting today (National Vital Statistics Reports, 2020) or deaths due to COVID-19 (McLaren, 2021). Healthcare has been a hot button topic for the past decade, with the implementation of the Affordable Care Act sparking renewed debate on the structure of the United States' healthcare system as a whole. An area of concern in this debate is the overall health of Americans and how different racial groups perceive their own health.

Since 1985, the CDC has conducted an annual telephone survey across the country, called the Behavioral Risk Factor Surveillance System (BRFSS), and made the data open to the public. Participants of this survey reveal a myriad of data points regarding their personal characteristics, such as self-described health status, condition, lifestyle, and demographics. Participants in the survey may choose to classify their own "general health status" into five categories: excellent, very good, good, fair, or poor.

While these categories do not provide us with a precise measure of health for respondents, they do provide the opportunity for a limited dependent variable study by using a logit model to observe effects on an outcome, $generalhealth_i$. The qualitative nature of the data point itself, and the fact that there is no singular standardized measure in the first place, creates a lack of statistical precision while attempting to measure the health of a person, making a limited dependent variable model necessary for any study using this outcome.

1.1 Application of Limited Dependent Variable Methods

The logistic function was first invented by Pierre François Verhulst in the 19^{th} century to model population growth but was not used widely in statistics until Pearl and Reed (1920) independently came up with the function to model the population of the US. In the '30s, the probit model was created by Bliss, which in turn gave rise to the logit as an alternative by Berkson.

The logit model is a specification of the binary choice model where $\varepsilon_i \sim iid\ sec^2$. The binary choice model itself is used in regression when the outcome, Y_i is an unobservable qualitative data point, and instead we observe a binary variable J_i which is equal to 1 when Y_i is positive. We estimate the binary choice model by,

$$P(J_i = 1) = P(\varepsilon_i < \frac{X_i\beta}{\sigma}) = F(\frac{X_i\beta}{\sigma})$$

which implies that,

$$P(J_i = 0) = 1 - F(\frac{X_i\beta}{\sigma}).$$

Since the variable J_i is a binary, its density function is,

$$f(J_i) = [F\left(\frac{X_i\beta}{\sigma}\right)]^{J_i} [1 - F\left(\frac{Xi\beta}{\sigma}\right)]^{1-J_i}$$

However, the variables β and σ are unidentified. $\delta = \sigma^{-1}\beta$ is identifiable though. This allows us to use $F(X_i\delta)$, which in the logit case is $\Psi(X_i\delta)$ for the logistic function. Thus, the log-likelihood function for the logit model is,

$$lnL(\delta) = \sum_{i=1}^{n}\{J_i\, ln[\Psi(X_i\delta)] + (1 - J_i)ln[1 - \Psi(X_i\delta)]\}$$

and the score equations,

$$\frac{\partial lnL(\delta)}{\partial \delta} = \sum_{i=1}^{n} \{ J_i \frac{\psi(X_i\delta)}{\Psi(X_i\delta)} - (1 - J_I) \frac{\psi(X_i\delta)}{1 - \Psi(X_i\delta)} \} X_i'$$

The dependent variable in this study, *generalhealth*$_i$, is a binary indicator = 1 if respondents report “excellent” or “very good” health and = 0 otherwise. In this manner, one can predict the likelihood of respondents reporting a positive perception of their own health.

2. MODEL

The logit model is constructed to show the probability that someone reports a positive perception of their own health. Specifically, using the data in the following specification,

$$P\ (\ generalhealth_i = 1\ ) = \ \Psi\ (\ \alpha + \beta_{race}Race + \beta_{personal}X + \varepsilon_i\ )$$

where β_{race} is a vector of coefficients for each group in the $Race$ vector, $\beta_{personal}$ is a vector of coefficients for personal characteristics correlated with health, and *X* is the vector of those characteristics. The *X* vector consists of the following data points included in the BRFSS that may correlate with personal health: state, health plan, age group, marital status, number of children, education level, income group, sex, exercise habits, disability status, veteran status, and BMI.

2.1 Regressors

The perception of your own health is a complex variable dependent upon a wide web of personal characteristics and experience. The primary characteristic in this study is race and ethnicity, and participants have been grouped into four categories: White, Black, Asian/Native

Hawaiian/Pacific Islander/American Indian/Alaska Native, and Hispanic. It is well known that different racial groups report different healthcare outcomes, and this study is designed to further that line of inquiry by measuring each category's probability of reporting a positive perception of health.

Another major variable that may impact your perception of health is your healthcare plan, with those not insured more likely to be worried about their health and report poorer outcomes in general. This is a major sticking point in this study, with the aforementioned ACA put into legislation in 2010 and into full effect in 2013. The timing of these laws had a large impact on insurance rates, take-up, and level of coverage for Americans in the 2010s, which is the exact timing of this study.

Respondents are also grouped by their state of residency to eliminate geographic trends and tendencies in health reporting. Age group, broken into five-year increments by the CDC, is also a major factor in health, as well as marital status, the number of children in the household, education level the respondent has attained, their income group, sex or gender, exercise habits, disability status, veteran status, and BMI.

3. RESULTS

The results from estimating the logit model are given in Table A, beginning in 2011, approximately two years prior to the full implementation of the Affordable Care Act. Coefficients, standard errors, z-scores, and p-values are reported for each of the three minority groups, using the White group as the comparison.

For the logit model, coefficients are not interpreted as effects on treatment, rather they are used to test for significance. For marginal effects, the vector of standardized coefficients are taken as follows,

$$\frac{\partial P(generalhealth_i \ = \ 1)}{\partial X_i{}'} = \frac{\partial \Psi(X_i\delta)}{\partial X_i{}'} = \ \psi(X_i\delta)\delta$$

These marginal effects are more appropriate for estimating a one-unit change in the regressor, in this case the binary for each racial group switching from 0 to 1. These estimates are reported in Table B along with standard errors obtained through the Delta Method and their corresponding p-values.

In 2011, respondents to the BRFSS were 8.12 percentage points less likely to report a positive health status than those in the White category if they were Black, 11.67 percentage points if they were Asian/Native Hawaiian/Pacific Islander/American Indian/Alaska Native, and 10.35 percentage points if they were Hispanic. Each effect is statistically significant, with a p-value of 0.00.

3.1 Results Consistency

These effects are maintained with consistency throughout the sample time period, hovering between 6-8, 8-11, and 8-10 percentage points for each group respectively. It should be considered that the sample in 2016 is nearly half the size of the other years due to missing observations in the dataset. However, it is unmistakable that minority groups consistently over time have a lower probability of reporting a positive health status.

Also of note, though, is the apparent downward trend. While mild and susceptible to margin of error, it is shown that throughout the six-year period the negative effects shrink by 1-2 percentage points for each group: an encouraging sign. Any causal inferences cannot be made by this study, however.

4. CONCLUSION

How people perceive their own health is an important quality of life indicator and is of concern to those invested in public health and, to an extent, the overall economy. The Affordable Care Act was aimed at Americans without adequate healthcare coverage, and by logical extension, high quality health outcomes. This study has shown that over the time period before and after the full implementation of the legislation, minority racial groups in the United States have consistently had a lower probability of reporting a positive health status.

While this is a clear disparity in perceived health, no causal implications can be made here. What can be said is that there appears to be a positive trend in these disparities, with each people group showing mildly improved probabilities over the span of the analysis. Based on the results here, the next logical progression for the research would be to deconstruct the outcome from a binary into ordinal groups and utilize an ordered logit model to dig deeper into the problem. This model was intended to establish a benchmark for further study by those interested.

5. APPENDIX: BRFSS DATA

The data for this study was that of the Centers for Disease Control and Prevention and their annual Behavioral Risk Factor Surveillance System from the years 2011 through 2016.

This model is constructed using each wave of data, with 307,043 or more observations in each sample year other than 2016. That year, the sample only consists of 157,293 observations due to missing values in the dataset. The following variables were utilized: genhlth, _race, hlthpln1, _ageg5yr, marital, children, educa, income2, sex, exerany2, qlactlm2, veteran, _bmi5cat, and _state.

The variable children was recoded to be a categorical variable indicating 0 children living in the home, 1 child, 2 children, or more than 3. The variable genhlth was recoded from an ordinal category into a binary indicating a positive health perception. The loss of statistical precision from this approach is noted in the paper. The variable income2 is an ordinal category splitting annual incomes into groups of less than $10,000, between $10,000 and $15,000, between $15,000 and $20,000, between $20,000 and $25,000, between $25,000 and $35,000, between $35,000 and $50,000, between $50,000 and $75,000, and finally greater than $75,000. The variable educa is an indicator for level of education of the respondent, categories being: Never attended school or only kindergarten, Grades 1 through 8 (Elementary), Grades 9 through 11 (Some high school), Grade 12 or GED (High school graduate), College 1 year to 3 years (Some college or technical school), or College 4 years or more (College graduate). The variable _bmi5cat are categories of the Body Mass Index, Underweight, Normal Weight, Overweight, or Obese. The variable marital is an indicator of the respondent's marital status, categories being: Married, Divorced, Widowed, Separated, Never Married, or a member of an unmarried couple. The variable _race was recoded to group Asian/Native Hawaiian/Pacific Islander/American Indian/Alaska Native peoples together due to their small proportions. The other racial groupings are the same.

The variable hlthpln1 is a binary indicator for whether the respondent has any type of healthcare coverage, be it private, employer provided, or government provided. The variable sex is an indicator for the respondents' gender at birth. The variable exerany2 is a binary indicator for whether the respondent has had any type of exercise in the last 30 days. The variable veteran is a binary indicating US military service. Finally, _state is a variable indicating the state or US territory which the respondent resides primarily.

In total, 14 variables were used from the CDC's BRFSS and any observations with missing data points were dropped from the analysis.

6. TABLES

Table A: Logit Model Estimation

		Year	Group	Coefficient	Std. Error	z	P>|z|
Log likelihood:	-205784.12	2011	_cons	-2.50	0.156	-16.01	0.000
Number of obs:	368,449		2	-0.42	0.016	-27.45	0.000
			3	-0.61	0.022	-27.64	0.000
			4	-0.54	0.017	-30.98	0.000
Log likelihood:	-205103.68	2012	_cons	-2.68	0.155	-17.32	0.000
Number of obs:	370,463		2	-0.38	0.015	-25.87	0.000
			3	-0.59	0.021	-27.64	0.000
			4	-0.52	0.017	-29.99	0.000
Log likelihood:	-200355.23	2013	_cons	-3.07	0.163	-18.84	0.000
Number of obs:	360,822		2	-0.39	0.016	-24.92	0.000
			3	-0.55	0.022	-25.52	0.000
			4	-0.50	0.018	-28.15	0.000
Log likelihood:	-196080.41	2014	_cons	-2.78	0.149	-18.62	0.000
Number of obs:	352,103		2	-0.38	0.016	-24.16	0.000
			3	-0.55	0.022	-25.30	0.000
			4	-0.49	0.018	-27.39	0.000
Log likelihood:	-172104.11	2015	_cons	-3.09	0.179	-17.19	0.000
Number of obs:	307,043		2	-0.39	0.017	-22.85	0.000
			3	-0.56	0.023	-24.57	0.000
			4	-0.46	0.019	-24.71	0.000
Log likelihood:	-88650.227	2016	_cons	-2.80	0.205	-13.66	0.000
Number of obs:	157,293		2	-0.33	0.020	-16.12	0.000
			3	-0.44	0.035	-12.34	0.000
			4	-0.45	0.026	-17.39	0.000

Table B: Marginal Effects

Year	Group	Marginal	Std. Error	z	P>|z|
2011	2	-0.0812	0.0030	-27.36	0.000
	3	-0.1167	0.0042	-27.91	0.000
	4	-0.1035	0.0033	-31.02	0.000
2012	2	-0.0727	0.0028	-25.70	0.000
	3	-0.1117	0.0040	-27.78	0.000
	4	-0.0994	0.0033	-29.93	0.000
2013	2	-0.0735	0.0030	-24.76	0.000
	3	-0.1045	0.0041	-25.59	0.000
	4	-0.0946	0.0034	-28.06	0.000
2014	2	-0.0734	0.0031	-24.00	0.000
	3	-0.1046	0.0041	-25.33	0.000
	4	-0.0927	0.0034	-27.25	0.000
2015	2	-0.0746	0.0033	-22.70	0.000
	3	-0.1073	0.0044	-24.64	0.000
	4	-0.0884	0.0036	-24.58	0.000
2016	2	-0.0635	0.0040	-16.07	0.000
	3	-0.0840	0.0068	-12.38	0.000
	4	-0.0869	0.0050	-17.41	0.000

Table C: Sample Proportions

	2011	2012	2013	2014	2015	2016
General Health						
1	50.08	49.78	49.95	50.49	50.23	49.70
0	49.92	50.22	50.05	49.51	49.77	50.30
Healthcare						
1	88.75	88.44	88.72	91.87	92.71	92.75
0	11.25	11.56	11.28	8.13	7.29	7.25
Age						
1	4.43	5.23	5.58	5.27	5.55	5.55
2	4.28	4.52	4.70	4.33	4.53	4.80
3	5.40	5.52	5.59	5.16	5.25	5.30
4	5.85	5.80	5.77	5.54	5.63	5.67
5	6.91	6.75	6.46	6.23	5.95	5.67
6	8.09	7.69	7.43	7.12	6.94	6.88
7	10.08	9.84	9.68	9.45	9.14	8.72
8	11.04	10.62	10.78	10.77	10.60	10.38
9	11.41	10.99	11.01	11.47	11.42	11.42
10	9.60	9.90	10.27	11.04	11.30	11.64
11	7.76	7.84	8.21	8.75	8.80	9.11
12	6.21	6.03	6.18	6.39	6.47	6.57
13	8.94	9.25	8.33	8.46	8.43	8.28
Sex						
1	39.14	40.31	40.93	41.53	42.35	43.32
2	60.86	29.69	59.07	58.47	57.65	56.68
Veteran						
0	87.21	87.40	87.47	86.60	86.87	86.83
1	12.79	12.60	12.53	13.40	13.13	13.17
Marital Status						
1	53.51	52.49	51.87	53.65	53.20	52.50
2	14.19	14.08	14.41	13.69	13.55	13.75
3	13.78	13.57	13.47	13.21	12.88	12.76
4	2.13	2.17	2.18	2.01	2.05	2.10
5	13.85	15.06	15.37	14.86	15.44	15.80
6	2.55	2.64	2.70	2.58	2.88	3.09
Any Exercise						
0	26.93	25.00	27.38	24.39	26.63	25.51
1	73.07	75.00	72.62	75.61	73.37	74.49

Table C: Sample Proportions

	2011	2012	2013	2014	2015	2016
Num. of Children						
0	73.09	73.13	73.46	73.94	74.01	74.51
1	10.87	10.95	10.87	10.59	10.59	10.61
2	9.87	9.75	9.53	9.30	9.28	8.97
3+	6.17	6.17	6.14	6.17	6.12	5.91
Education Level						
1	0.12	0.13	0.14	0.15	0.14	0.15
2	2.86	2.86	2.73	2.57	2.54	2.51
3	5.80	5.95	5.74	5.31	5.11	5.17
4	29.13	29.45	29.20	28.49	28.03	28.20
5	27.16	27.10	27.42	27.26	27.41	27.53
6	34.93	34.51	34.76	36.22	36.76	36.45
Income Level						
1	5.99	6.17	6.04	5.39	5.10	4.90
2	6.46	6.46	6.37	5.83	5.41	5.39
3	8.05	8.33	8.30	7.76	7.40	7.63
4	9.94	9.88	9.93	9.54	8.95	9.37
5	11.90	11.58	11.62	11.27	10.84	10.88
6	15.02	14.64	14.63	14.60	14.38	14.41
7	15.84	15.64	15.52	15.81	16.06	16.04
8	26.80	27.30	27.59	29.80	31.86	31.38
Activity Limited						
0	71.14	75.27	75.72	75.01	75.09	74.32
1	28.86	24.73	24.28	24.99	24.91	25.68
BMI Category						
1	1.69	1.73	1.78	1.68	1.66	1.69
2	34.02	33.76	33.33	32.69	32.44	31.81
3	36.27	36.15	35.94	36.23	36.29	36.11
4	28.02	28.35	28.96	29.40	29.61	30.39
Race						
1	80.83	79.45	79.86	80.04	79.28	78.98
2	7.96	8.77	8.31	7.86	8.10	8.49
3	3.56	3.93	3.98	4.06	4.18	4.11
4	7.65	7.85	7.86	8.04	8.44	8.42

References

Cramer, J.S., (2002) “The Origins of Logistic Regression”. Tinbergen Institute Working Paper No. 2002-119/4.

Ely DM, Driscoll AK, (2020) “Infant mortality in the United States, 2018: Data from the period linked birth/infant death file” National Vital Statistics Reports, vol. 69, no. 7, National Center for Health Statistics.

Maddala, G. S. (1983). Limited-Dependent and Qualitative Variables in Econometrics. New York: Cambridge University Press.

McLaren, John. (2021) "Racial Disparity in COVID-19 Deaths: Seeking Economic Roots with Census Data" The B.E. Journal of Economic Analysis & Policy, vol. 21, no. 3, pp. 897-919.