University of North Carolina at Asheville
Journal of Undergraduate Research
Asheville, North Carolina
May 2014

Predicting Blood Pressure Using Demographics and Consumer Behavior

James Sharpe
Mathematics
The University of North Carolina at Asheville
One University Heights
Asheville, North Carolina 28804 USA

Faculty Advisor: Dr. Jimin Lee

Abstract

High blood pressure is a leading cause of cardiovascular disease, with billions of dollars spent in the U.S. every year to treat hypertension. Hypertension has even been classified as an epidemic by some researchers and doctors. As such, it is important to understand who is at risk of developing high blood pressure, so that, hopefully, preventative steps can be taken. Using data acquired from the National Health and Nutrition Examination Survey (NHANES), we developed a regression model to predict systolic blood pressure from demographics and consumer data such as age, gender, education level, marital status, money spent on groceries and restaurants and how far away participants live from a grocery store. We converted the categorical variables we used – marital status, education – into dummy variables. We found that all of the demographics variables used contributed significantly to blood pressure, and that three of the behavior variables did. A logistic regression model was also developed to determine how likely someone with the given attributes is to develop high blood pressure. From the logistic regression, we calculated odds ratios to see if different groups of people were more or less likely to develop high blood pressure relative to each other.

1. Introduction

Cardiovascular disease is a major cause of deaths in the United States.[1] It is estimated that by 2030, health care cost to treat cardiovascular disease will be about 818 billion dollars in the U.S. alone .[1] High blood pressure, or hypertension, is a leading cause of cardiovascular disease and in many cases is preventable, as some factors leading to chronic hypertension include: weight, diet, amount of physical activity, smoking and prevalence of other diseases such as diabetes.[1]

But the problem is not just limited to the United States. In the year 2000, it was estimated that 26.4% of the world wide population has hypertension.[2] The prevalence of hypertension is even higher if certain countries or regions of the globe are examined. For example, the rate of hypertension in Latin America and the Caribbean of males is 44.5%.[2] The results are even more alarming in Germany, where 60.2% of all males included had hypertension.[2] Race also plays a role in hypertension. In the United States, African-Americans have a prevalence rate that is 50% higher than their Caucasian counterparts .[3]

Given the severity of this problem, and its global presence, it is important to better understand the factors contributing to hypertension. The purpose of this study was to attempt to determine some of these factors that contribute to hypertension, and how these factors contribute. This study focused on demographic variables and behavior variables, specifically consumer behavior, such as spending and eating habits. These factors were of interest because these are things that could potentially be controlled or modified.

This paper examines the effects of several factors on blood pressure using regression analysis and logistic regression analysis. We will first discuss the data, and briefly summarize the modifications that we made to the data sets before we performed the analyses. The analysis and the results of our analyses will be discussed in detail. We will also offer a discussion of our results.

2. Methods

2.1. Data

The data used was acquired from the National Health and Nutrition Examination Survey (NHANES) in August of 2013 (NHANES). A total of six datasets were used: “Demographic Variables and Sample Weights” from 2007-2008 and 2009-2010, “Blood Pressure” from 2007-2008 and 2009-2010 and “Consumer Behavior” form 2007-2008 and 2009-2010.[4-9] These datasets were merged using SAS.

The dataset had 86 variables, but, for this study, only eleven of them were examined. Those variables were: gender, marital status (MarStat), Age, education (Edu2), income (ah1), amount of money spent at grocery stores per month in dollars (Money), amount of money spent at restaurants per month in dollars (Rest), time spent cooking (Time), Distance in minutes away from the nearest grocery store (Distance), number of meals families ate together per week (Fammeals) and systolic and diastolic blood pressure (sbp and dbp respectively).

Modifications to this dataset were performed before it was used. Some responses of some variables were taken out for various reasons. An example of this was individuals who responded 77 – which stands for “refused” - and 99 – which stands for “Don’t know”- for MarStat were removed. This was done because of the low number of people who chose these responses and so fewer dummy variables would be created later. Numeric variables were also modified. For example, the variable Money ranged from 0 to 3000, however, participants who refused to answer that question were coded as 777777 and participants who didn’t know were assigned the code 999999. These values were taken out because they greatly skewed the data, and because of the low number of observations with those values. Other variables that had responses taken out in a similar fashion were: Edu2, ah1, Rest, Time, Fammeals and Distance. It is also important to note that income was originally a categorical variable with 16 categories. This was condensed down to 2 categories, people who made below the median income and people who made at or above the median income, so that fewer dummy variables could be created, making later analyses easier. Also, this study only examined individuals aged 20 to 60.

Dummy variables were created for categorical variables, including: MarStat and Edu2. This was done because, although each category is represented by a number, the categories have no real numeric relationship to each other; the numbers selected to represent different categories is arbitrary.

2.2. Research Methods

Once the dataset was obtained, a regression analysis was used to test for significance between blood pressure and demographics and behavior. A logistic regression was also performed to predict the probability that an individual, given certain characteristics, will develop high blood pressure, and to produce odds ratios to compare different groups of people with each other. Both models were executed in SAS, and SAS’s stepwise selection command was used to find the best model. Additionally, variance inflation factors (VIFs) were used to detect if multicollinearity existed among the variables.

The form of the regression analysis that we used was:

$$E(sbp) = \beta_0 + \beta_1 Age + \beta_2 Gender + \beta_3 e_1 + \beta_4 e_2 + \beta_5 e_3 + \beta_6 e_4 + \beta_7 m_1 + \beta_8 m_2 + \beta_9 m_3 + \beta_{10} m_4 + \beta_{11} m_5 + \beta_{12} ah_1 + \beta_{13} Money + \beta_{14} Rest + \beta_{15} Time + \beta_{16} Distance + \beta_{17} Fammeals \quad (1)$$

The logistic regression used was:

$$logit(\hat{\pi}) = \ln(odds) = \ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$$
$$= \beta_0 + \beta_1 Age + \beta_2 Gender + \beta_3 Edu2_1 + \beta_4 Edu2_2 + \beta_5 Edu2_3 + \beta_6 Edu2_4 + \beta_7 Rest + \beta_8 Fammeals \quad (2)$$

Note that this logistic model is the best model, as determined by stepwise selection. Where $\hat{\pi}$ is the probability of success, P(s). Success was defined in terms of failures; that is P(s) = 1- P(f), where P(f) is the probability of failure. P(f) was defined as: P(f) = P(dbp < 98 and sbp < 142). That is, if both the prediction for systolic blood pressure and diastolic blood pressure are less than 142 and 98 respectively, then that individual will not be considered to have high blood pressure.

From the logistic regression, odds were calculated. They were calculated as follows:

$$odds = \frac{\hat{\pi}}{1-\hat{\pi}}$$
$$= \beta_0 + \beta_1 Age + \beta_2 Gender + \beta_3 Edu2_1 + \beta_4 Edu2_2 + \beta_5 Edu2_3 + \beta_6 Edu2_4 + \beta_7 Rest + \beta_8 Fammeals \quad (3)$$

Notice that odds are the ratio of the probability of success of an event divided by the probability of failure of an event. Odds can be interpreted as the likelihood that an individual will develop high blood pressure. Odds greater than 1 means that someone is more likely to develop high blood pressure than not; the higher that number is, the more likely they are to develop high blood pressure. An odds less than 1 means that person is more likely to not develop high blood pressure than they are to develop high blood pressure; the lower the number is, the less likely they are to develop high blood pressure. And an odds of 1 means that a person is just as likely to develop high blood pressure than they are to not develop high blood pressure; their probability of success and probability of failure would be equal.

Given the odds, we can calculate odds ratios to compare different people, or groups of people, with each other using this formula:

$$odds\ ratio = \frac{\beta_0 + \beta_1 Age + \beta_2 Gender + \beta_3 Edu2_1 + \beta_4 Edu2_2 + \beta_5 Edu2_3 + \beta_6 Edu2_4 + \beta_7 Rest + \beta_8 Fammeals}{\beta_0 + \beta_1 Age + \beta_2 Gender + \beta_3 Edu2_1 + \beta_4 Edu2_2 + \beta_5 Edu2_3 + \beta_6 Edu2_4 + \beta_7 Rest + \beta_8 Fammeals} \quad (4)$$

For example, if we wanted to examine the odds ratio for males versus females, we would take the odds for males and divide it by the odds for females, like so:

$$\frac{odds\ for\ male}{odds\ for\ female} = \frac{e^{02448*1}}{e^{0.2448*0}} = 1.28.$$

Notice that the 0.2448 is the parameter estimate for gender. This, along with all other parameter estimates, can be found in tables 1 and 4. Also note that if you only want to compare males and females, all variables would remain the same except for gender, and thus they would cancel like so:

$$odds\ ratio = \frac{e^{\beta_0+\beta_1 Age+\beta_2 Gender+\beta_3 Edu2_1+\cdots+\beta_8 Fammeals}}{e^{\beta_0+\beta_1 Age+\beta_2 Gender+\beta_3 Edu2_1+\cdots+\beta_8 Fammeals}}$$

$$= \frac{e^{\beta_0} * e^{\beta_1 Age} * e^{\beta_2 * 1} * \ldots * \beta_8 Fammeals}{e^{\beta_0} * e^{\beta_1 Age} * e^{\beta_2 * 0} * \ldots * \beta_8 Fammeals}$$

$$= \frac{e^{\beta_2 * 1}}{e^{\beta_2 * 0}}.$$

All terms in blue are repeated in both the denominator and the numerator, so they cancel.

3. Results

The initial regression analysis contained all the 11 variables. Since all the VIFs were below 10, it was concluded that multicollinearity did not exist among any of the variables, which means that none of the variables are highly correlated with each other (table 1). The hypothesis testing was done in SAS; the test statistic used was the F value. With an F-value of 123.81, corresponding to a p-value < 0.0001, it was concluded that the model was a good model; at least one of the variables used was a significant predictor for blood pressure. The R^2 value was 0.2294. This means that 22.94% of the variance in systolic blood pressure can be explained by the variables used. The results are displayed in table 1, and the dummy variables for education and marital status are illustrated in tables 2 and 3.

Table 1. Parameter estimates of each variable used in the initial regression analysis. Also included are the standard errors for those estimates. SAS performed hypothesis tests to determine if each individual variable contributed significantly to predicting blood pressure, hence, the t-values and p-values. Note that ah_1 is a dummy variable created for income. Ah_1=0 means that that individual's annual household income was less than the median income for the United States and ah_1=1 means that that individual's annual household income was greater than the median income.

i	Variable	B_i	Standard Error	t value	p value	VIF
0	Intercept	104.94902	1.12584	93.22	< 0.0001	0
1	Age	0.46618	0.01291	36.12	< 0.0001	1.42980
2	Gender	4.19444	0.37743	-11.11	< 0.0001	1.04732
3	e_1	-1.48319	0.70099	-2.12	0.0344	2.05633
4	e_2	-1.68783	0.66725	-2.53	0.0114	2.37055
5	e_3	-1.85855	0.67074	-2.77	0.0056	2.62826
6	e_4	-4.68388	0.75335	-6.37	< 0.0001	2.50112
7	m_1	2.97024	0.98363	3.02	0.0025	1.12004
8	m_2	0.94781	0.75369	1.26	0.2086	1.07691
9	m_3	-0.14783	1.14628	-0.13	0.8974	1.04398
10	m_4	2.61682	0.63698	4.11	< 0.0001	1.29611
11	m_5	2.19340	0.69373	3.16	0.0016	1.15587
12	ah_1	-0.96534	0.43217	-2.23	0.0255	1.32159
13	Money	-0.00051	0.00071	-0.72	0.4738	1.07091
14	Rest	-0.00228	.00109	-2.09	0.0364	1.16485
15	Time	-0.00502	.00417	-1.20	0.2291	1.02004
16	Distance	0.02376	.00942	2.52	0.0117	1.01161
17	Fammeals	-0.0663	.04115	-1.61	0.1070	1.08400

Table 2. Dummy variables for education that were manually created through SAS programming. Note that "less than 9th grade" education was used as the baseline; all other categories were compared to this baseline. From table 1, e_2 has a value of -1.68783. This means that people who have a high school diploma or GED have a systolic blood pressure that is, on average, 1.68 points lower than someone who did not attend high school.

Dummy Variable (e_1, e_2, e_3, e_4)	Category
(0,0,0,0)	Less than 9th Grade
(1,0,0,0)	9-12th Grade with no diploma
(0,1,0,0)	High School Graduate/GED
(0,0,1,0)	Some College of AA degree
(0,0,0,1)	College Graduate or above

Table 3. Dummy variables for marital status that were manually created through SAS programming. For the regression analysis and logistic analysis, "Married" was the baseline, and all other categories were compared to this.

Dummy Variable(m_1, m_2, m_3, m_4, m_5)	Category
(0,0,0,0,0)	Married
(1,0,0,0,0)	Widowed
(0,1,0,0,0)	Divorced
(0,0,1,0,0)	Separated
(0,0,0,1,0)	Never married
(0,0,0,0,1)	Living with partner

From table 1, it is clear that some variables, such as money and time, do not significantly contribute to the model. Thus, we use a stepwise selection procedure, carried out by SAS, to only keep variables that contribute significantly to the prediction of systolic blood pressure. The results of this procedure are displayed in table 4.

Table 4. Results of the regression analysis after stepwise selection removed non-significant variables. Note that SAS used a level of significance of $\alpha=0.15$ to determine if a variable stayed in the model. Also note that the variables Money and Time were removed, as they were not significant at the $\alpha=0.15$ level.

i	Variable	B_i	Standard Error	t value	p value
0	Intercept	104.16989	0.96710	107.71	< 0.0001
1	Age	0.46756	0.01258	37.18	< 0.0001
2	Gender	4.06794	0.37066	-10.97	< 0.0001
3	e_1	-1.40956	0.68643	-2.05	0.0401
4	e_2	-1.70146	0.64994	-2.62	0.0089
5	e_3	-1.75926	0.65510	-2.69	0.0073
6	e_4	-4.72968	0.71603	-6.61	< 0.0001
7	m_1	2.66745	0.96856	2.75	0.0059
8	m_2	0.68079	0.73523	0.93	0.3545
9	m_3	-0.39040	1.11449	-0.35	0.7261
10	m_4	2.67748	0.62140	4.31	< 0.0001
11	m_5	2.14509	0.68183	3.15	0.0017
12	ah_1	-1.02296	0.42170	-2.43	0.0153
13	Rest	-0.00228	0.00103	-2.22	0.0263
14	Distance	0.02387	0.00897	2.66	0.0078
15	Fammeals	-0.06309	0.04012	-1.57	0.1158

The results from the logistic regression, as seen in equation (2), are displayed in table 5. SAS's stepwise selection command was also used to procure the best model – the model that did not include variables that were not significant – for the logistic analysis. The adequacy of the logistic model was checked via a Hosmer and Lemeshow Goodness-of-Fit test. This fitness test yielded a Chi-Square test statistic of 5.7384, which corresponds to a p-value of 0.6765. With this information, we concluded that the model was a good model.

Table 5. Parameter estimates for the variables used in the logistic regression. Note that the level of significance for the stepwise selection procedure here is $\alpha=0.05$, not $\alpha=0.15$ like the regression analysis above.

Parameter	Estimate	Standard Error	Wald Chi-Square	P-value
Intercept	-5.9705	0.3216	344.7390	< 0.0001
Age	0.0834	0.0061	189.8992	< 0.0001
Gender	0.2488	0.1146	4.7108	0.0300
Edu2 1	0.0996	0.1399	0.5066	0.4766
Edu2 2	0.2417	0.1137	4.5195	0.0335
Edu2 3	0.0341	0.1072	0.1013	0.7502
Edu2 4	0.0515	0.1042	0.2442	0.6212
Rest	-0.0008	0.0004	4.0446	0.0443
Fammeals	-0.0509	0.0159	10.1874	0.0014

4. Conclusion

4.1. Regression Analysis

For the regression analysis, it was found that females have a systolic blood pressure that is about four points lower than males. Individuals who earn greater than median income are expected to have a systolic blood pressure that is about1 point less than someone who earns lower than the median income. There is also a negative correlation between systolic blood pressure and education level; the higher education attained, the lower expected systolic blood pressure will be. College graduates had a systolic blood pressure that was 5 points lower than people who did not finish high school. This could be because college graduates, overall, make more money and have lower unemployment than someone who does not have a college degree (U.S. bureau of labor statistics).[10] This would also validate the earlier claim that people with higher incomes have lower systolic blood pressures.

Also, age has a positive correlation with high blood pressure, as seen in tables 1 and 4. To visualize this, a scatterplot was created that plotted age versus probability of developing high systolic blood pressure (figure 1).

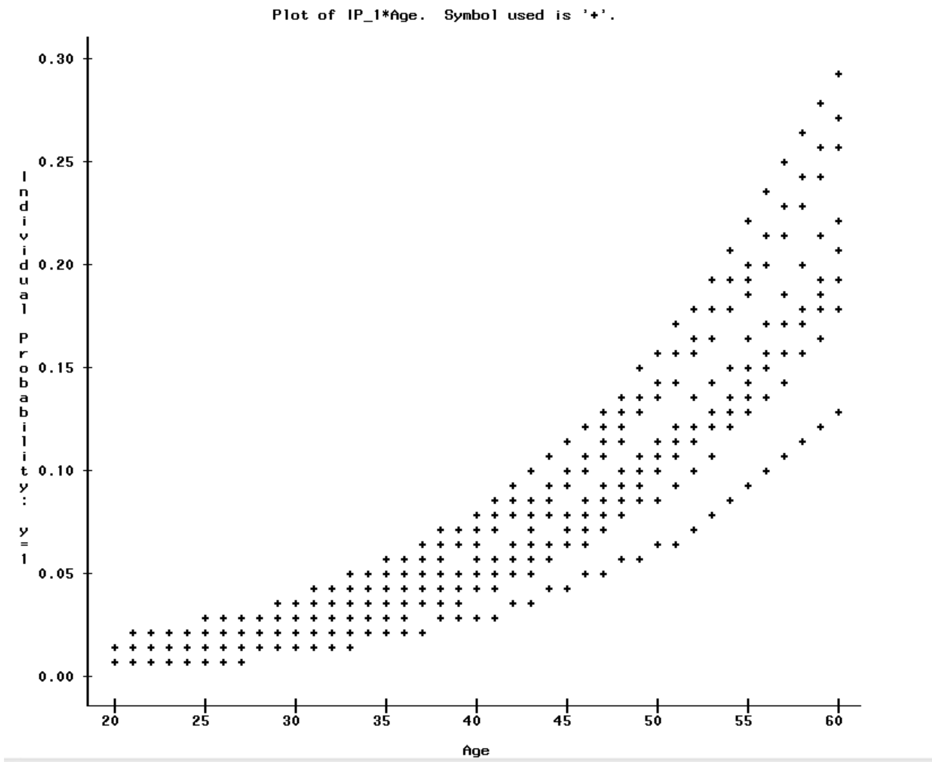

Figure 1. Scatterplot of age versus probability of developing high blood pressure.

From this plot, it is observed that as age increases, the likelihood that someone will develop high blood pressure increases drastically.

Blood pressure is also correlated with marital status. People who are widowed, divorced or who have never married tend to have higher blood pressures than people who are married (table 4). Surprisingly, people who are separated had the lowest systolic blood pressure. It is important to note that age may have an interactive effect with marital status though. On average, we would expect the age of someone who is widowed to be higher than the age of someone who is married or never married.

Even more interesting was that people who lived farther away from grocery stores had higher systolic blood pressures. Also of interest was that people who ate more family meals together tended to have lower systolic blood pressures (table 4). Fammeals was measured as the number of meals a family ate together per week. From the regression analysis, we expect that each meal eaten with family will reduce systolic blood pressure by 0.0509 points (table 4). In contrast, the more money spent at restaurants, the lower a person's systolic blood pressure will be. For every dollar spent at a restaurant, once could expect their systolic blood pressure to fall by 0.0008. These two results seem to contradict each other, but maybe not. Maybe family meals are healthier, and maybe healthy meals at restaurants are more expensive. However, this is just a speculation; more research would be needed to find an answer.

4.2. Logistic Regression Analysis

From the logistic analysis, odds ratios comparing many groups of people were calculated (table 6).

Table 6. Odds ratios for various groups of people. Odds ratios were calculated via SAS, but can easily be calculated by using equation (4); also included in the SAS output are 95% Wald confidence limits for the odds ratios.

Group Comparisons	Odds Ratio	95% Wald Confidence Limits
Male vs. female	1.282	(1.0246, 1.605)
Less than 9th grade education vs. college graduate	1.693	(1.086, 2.693)
9th-12th grade with no diploma vs. college graduate	1.952	(1.316, 2.894)
High school diploma vs. college graduate	1.586	(1.084, 2.320)
Some college/AA degree vs. college graduate	1.613	(1.112, 2.341)

The odds ratio for male versus female was calculated to be 1.282. This means that the odds for males developing high blood pressure is 28.2% higher than the odds for females, meaning that males are more likely to develop high blood pressure. This parallels the results seen in the regression analysis concerning gender differences.

The positive trend with education and blood pressure we observed during the regression analysis is also present in the logistic analysis. College graduates are the least likely to develop high blood pressure; the odds for someone who never made it to the ninth grade developing high blood pressure is 69.3% higher than the odds for a college graduate (table 6). Even more startling is that the odds for someone who never finished high school developing high blood pressure is almost twice as high – 95.2% – as the odds of a college graduate (table 6). The odds for someone with a high school diploma and the odds for someone with some college/AA degree developing high blood pressure is about 60% higher than that of a college graduate (table 6). This trend might be because college graduates get jobs that they enjoy better than people who do not go to college, which may make them happier. The jobs of college graduates might also not be as stressful, or they may make more money, on average, which means they are happier than people who do not go to college. It would be interesting to study this in more detail, for example comparing different levels of college degrees – bachelor's, master's and doctorates – or even comparing different majors.

Surprisingly, the trend with marital status seen in the regression analysis was not seen in the logistic regression analysis. However, the trends of money spent at restaurants and number of family meals per week were seen in the logistic regression analysis. The more money an individual spends at a restaurant, the lower probability that that person is to develop high blood pressure. Similarly, the more family meals eaten per week, the lower the probability of developing high blood pressure.

It is interesting that in the logistic regression analysis, income was not significant. Income is a quantitative variable, but NHANES has it recorded as a categorical variable with 17 categories. For our study, we split it into two categories to make it easier to analyze, but it would be interesting to see what role income would play if it had been recorded as a quantitative variable.

For this study, we focused on systolic blood pressure. But other studies suggest that there are differences in systolic and diastolic blood pressure when attempting to understand prevalence of hypertension.[3] The rise in systolic blood pressure increases faster than the rise in diastolic pressure as people age.[3] Additionally, diastolic blood pressure seems to stop rising, or even decline, in later years, whereas systolic blood pressure keeps increasing.[3] It would be interesting to do a similar study that looked at diastolic blood pressure to see if there are any notable differences in prediction of blood pressure.

5. References

1. Heidenreich, Paul A., Justin G. Trogdon, Olga A. Khavjou, Javed Butler, Kathleen Dracup, Michael D. Ezekowitz, Eric A. Finkelstein, Yuling Hong, S. Claiborne Johnston, Amit Khera, Donald M. Lloyd-Jones, Sue A. Nelson, Graham Nichol, Diane Orenstein, Peter W.F. Wilson, and Y. Joseph Woo. "Forecasting the Future of Cardiovascular Disease in the United States: A Policy Statement From the American Heart Association." *Circulation* 123.8(2011): 933-44.

2. Kearney, Patricia M, Megan Whelton, Kristi Reynolds, Paul Muntner, Paul K. Whelton, and Jiang He. "Global Burden of Hypertension: Analysis of Worldwide Data." *The Lancet* 365(2005): 217-23.

3. Whelton, Paul K. "Epidemiology and the Prevention of Hypertension." *The Journal of Clinical Hypertension* 6.11(2004): 636-42.

4. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2007-2008, accessed: 2013, <http://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics&CycleBeginYear=2007>.

5. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2007-2008, accessed: 2013, <http://www.cdc.gov/nchs/nhanes/nhanes2007-2008/CBQ_E.htm>.

6. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2007-2008, accessed: 2013, <http://www.cdc.gov/nchs/nhanes/nhanes2007-2008/BPX_E.htm>.

7. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2009-2010, accessed: 2013 <http://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics&CycleBeginYear=2009>.

8. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2009-2010, accessed: 2013 <http://www.cdc.gov/nchs/nhanes/nhanes2009-2010/BPX_F.htm>.

9. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2009-2010, accessed: 2013 <http://www.cdc.gov/nchs/nhanes/nhanes2009-2010/CBQ_F.htm>.

10. U.S. Bureau of Labor Statistics, "Earnings and Unemployment Rates by Educational Attainment." 2014. Web. 26 Feb.