Econometric Predictions From Demographic Factors Affecting Overall Health

Stacey, Brian

Southern New Hampshire University

25 November 2015

Econometric Predictions From Demographic Factors Affecting Overall Health

Brian Stacey

Southern New Hampshire University

Executive Summary

Efforts to accurately predict health outcomes with a focus on informing policy makers of where to best spend limited resources have been made in the past. This paper builds on the efforts of those studies in an attempt to build an accurate predictor of health from readily available data. The American Time Use Survey (2010, 2012, and 2013) provides the majority of the data from which this model is built, and it is then tested via several methods.

The analysis finds that the existing freely available data is significant in its predictive power, however is missing too many predictors to reduce the confidence interval about each individual prediction to a point of bearing meaningful fruit. That does not eliminate the usefulness of the study however, as by reducing the confidence required and accepting that the data is used for predicting societal means, the model is able to accurately predict average outcomes. This paper further attempts to analyze state level date to provide a geographic target for public funds expenditures, and accomplishes this through the analysis of various risk factors by region.

Notable in this analysis is an attempt to correct for self-reporting errors. The literature review did not reveal any previous attempts to do so using a similar methodology (beyond recognizing that such errors exist and using robust methods to account for them), making this attempt possibly unique. The correction did not result in significantly different estimates, however that may be a result of the minimal resources applied to this small aspect of the analysis.

Description

Numerous efforts have been made to increase overall health levels in the United States. The Centers for Disease Control and Prevention (CDC) has stated that increased physical activity results in better overall health (Fatoye, F., 2013) and funds research and education toward that goal (CDC Health Disparities and Inequalities Report., 2013). This paper attempts to define the relationship between several key factors and overall health, as well as assess the predictive power of individual characteristics with regard to overall health and physical activity.

Efforts to increase overall health within the population should be targeted to those groups that can most benefit from an increase in such, and minimize the financial impact on the US budget. To that end, this analysis was undertaken using previously obtained data, to evaluate where the money is best spent, and which groups and regions of the country are most in need of this change.

This analysis was written with individuals responsible for creating and funding public policy in mind (lawmakers and the agencies that carry out those policies as defined in legislation) as they are in the position to implement these recommendations and thus affect the negative health outcomes described. To appeal to a broader base of readers the detailed analyses are included in attachments with a description of the findings included in the text of the paper. In this way the reader can gain an understanding of the issue and relevant findings without the need to delve into the mathematical analysis.

The goal of the model developed in this analysis is to provide a framework for evaluating how changes to demographic indicators can affect overall health with the goal of informing policy makers where best to spend public funds for education and research to affect those indicators that can have the most positive effect.

Literature Review

Numerous studies have been conducted in the past; from Abu-Omar, K., & Rütten, A. (2008) to Winkleby, M., Jatulis, D., Frank, E., & Fortmann, S. (1992), and everything in between. The findings within which have been consistent in that they assert and prove that education and income are indicators of health, likewise increases in physical activity improves overall health on average. With education, the "correlation is strong and significant even after controlling for different measures of socio-economic status, such as income and race, and regardless of how health is measured (morbidity rates, self-reported health status or other measures of health)." (Lleras-Muney 2005).

Marmot, M. (2002) offers the opinion that education may be an indicator of health in that "education affects health precisely because those with more education have higher incomes. It could, however, be because education is a better indicator than is income of some of the social factors, linked to social position, that are important for health." Lleras-Muney (2005) concluded that educational attainment is causal, and that an increase in education of one year can reduce the probability of dying in the next ten years by as much as 3.6%, however she was unable to prove the specifics of the causal path. Her findings were significant enough to suggest that increased education may be the most cost-effective means of improving overall adult health.

Since educational attainment and labor force status have both been conclusively shown to positively correlate with income, some underlying causality between the three should be investigated. Winkleby, et al (1992) assessed the interrelationship between the three in an overall

assessment of socioeconomic status (SES) and concluded again that of the SES indicators, increased education has the best payoff for overall health. In this analysis education, and income are included, but employment status is specifically excluded; it has been found to not be a significant predictor, and the information that could be gleaned from that predictor is largely represented in income (zero income implies unemployment).

This paper reviews the relationship between those predictors and overall health, and adds predictors in the areas of household size, marital status, number of kids in the house, pain medication, physical disabilities, and the degree to which the respondent reports being well rested; and makes similar conclusions. Where not specified, the default null hypothesis for regression within this analysis is that the relationship being investigated is not significant. This analysis is not unique in the area of predictors, nor in the area of conclusions, however it does attempt to find geographic regions that would most benefit as well as where the most benefit could be realized, and attempts to build a predictive model for overall health indicators where no widely used econometric model existed before.

Data

Data available from public governmental sources is abundant with regard to time use and medical expenditures for the national level, however is limited at the regional and state level. The Centers for Disease Control and Prevention publish regular tables of data including rates of reporting of numerous diseases and rates of reporting of physical activity. (Centers for Disease Control and Prevention, 2013). This CDC data is limited in that it is self-reported and only accounts for limited disease types. The Bureau of Labor Statistics: American Time Use Survey data, available from the ATUS website (ATUS Tables, 2014) and from the ATUS-X extract builder website (American Time Use Survey-X extract builder, 2015), has more detailed time studies with regard to leisure activities, working hours, etc. but only has Likert scale data for health of respondents. The Bureau of Labor Statistics (Databases, Tables & Calculators by Subject, 2015) and Bureau of Economic Analysis (National Data, 2015) both have vast quantities of data regarding demographics, income, expenditures, etc. however this data is limited to those topics and does not have meaningful data on health or activity levels. Finally the US Census (United States Census Bureau, 2014) includes health and activity levels, but is self-reported and limited in scope. All of these sources are cross-sectional data with the ATUS also producing time series for a subset of their sample (panel data).

None of the data available is perfect, it lacks the granularity to make meaningful detailed analysis and it is predominantly self-reported. Further analysis within this field would require data collected from a targeted study of a cross section of persons and include time use related to leisure time physical activity and specific health indicators (health care expenditures, time in medical offices, blood pressure, etc.). This future data should be collected by observation and not

be self-reported, to eliminate the bias inherent in such data.

The majority of this analysis is based upon the 2010, 2012, and 2013 American Time Use Survey (panel data). Summary statistics of the relevant precursors and outcomes have been evaluated for distribution, and multiple linear regression analysis performed to attempt to find a best fit model for predicting outcomes.

The primary limitation observed within the data is within the response variable; health is reported on a Likert scale and as such is ordinal at best, however ordinality does not lend itself to multiple linear regression in the OLS form. An ordered logit regression of the health data versus the final predictors chosen (see Appendix B) reveals that the predictive power is present, and the model is sound. The logit model is limited in its predictive power though (stochastic outcomes for health based on inputs are not the desired output) since a band of likely health outcomes is (mean predicted y +/- t*se) more useful in assessing the influence of the predictors, the response variable is assumed to be interval (as the Likert scale is intended to simulate); although, by using the concept of cut-off values the multiple logit model can provide discrete outputs that follow the Likert scale used in the initial data input. Both models are included in Appendix B for the reader to choose from; this analysis will focus on the OLS model in most cases.

The initial set of data includes 37088 observations over three years. The observations containing null values for the 11 variables being evaluated were removed. Methods were evaluated for extrapolating expected values for those empty observations, however with the size of the data set and the fact that most of the missing values were binary the decision was made to exclude them.

This left 16191 useful observations.

During the process of the below analysis of descriptive statistics for normality and outliers it was discovered that the $2884.61 value of weekly earnings was used as a “catch all” for all values above that threshold. This created a situation where the data is right skewed with a large set of outliers at the right tail (masking the true mean and median and increasing standard deviation). Those values were removed as they do not represent actual values but are a discrete category within an otherwise continuous variable. This resulted in 15619 useful observations. Those observations are used to develop the model.

After removing the non relevant observations, each of these continuous variables was evaluated for normality using an Anderson-Darling test; all were found to be normally distributed. This was the expected distribution and is indicative of a randomly distributed process. This distribution reduces the complexity of the subsequent analysis as the data closely fits the assumptions generally made in these types of analyses, eliminating the need to move to more complex techniques.

Methods

The primary method of analysis used for this paper was multiple linear regression. This method was chosen as it reveals the presence of a relationship, while also presenting the user with a regression equation that can be used (with some qualification) to predict likely outcomes of further data. As one of the purposes of this research is to provide a predictive tool, this attribute is of some importance.

Since the underlying purpose of this analysis is to build a predictive model, and the confidence intervals resulting from OLS and Robust OLS are too wide to provide meaningful prediction the data was reevaluated using an Ordered Logistic Regression (logit). The data limitations discussed above further limit the methods available and raise questions with the validity of the OLS method. The ordered logit eases those concerns by reinforcing the assumption that the multiple linear regression model is sufficient. The logit regression shows p-values for each coefficient well under the 0.05 level and an overall goodness of fit of 100%, both suggesting that the model chosen adequately predicts the response.

The parameters and thresholds in the ordered logit model are not intuitive in their interpretation (R forces a zero intercept and fits the model based on that assumption). John K. Kruschke (2014) developed a function for use in R to transform ordered logit (or probit) regressions into a form that sets the threshold levels at intuitive values (approximately one half the distance between the two adjacent levels) and fits a non-zero intercept and coefficients from there.

Using the thresholds and coefficients given from the Kruschke model, predicted values for Y (General Health) were obtained and regressed against actual values using R. The results show an R^2 of ~13% which indicates that this model has less predictive power than the simple OLS, however yields more intuitive results.

The decrease in the standard error in the logit model over the OLS model causes a subsequent decrease in the confidence intervals (CI) about the predictions. This reduction isn't large enough to cause the CIs to reduce enough to make the predicted range smaller than the total range. Since the gain in this model does not offset the loss in predictive power, this model is abandoned.

Further data and method limitations are those typical seen with linear regression; multicollinearity and heteroscedasticity being chief among them. In the case of multicollinearity it is expected within this data. The degree to which that is true is assessed further in Appendix B. Each of these predictors has an effect on the outcome, some of those effects overlap and are likely caused by a non-evaluated underlying cause. The things that tend to cause people to have more education also tend to cause them to have higher income, for example; however there are other aspects of those predictors independent of this underlying cause, and it is those aspects that create the separate influence on overall health. The analysis reveals that; as expected, earnings and education are correlated (R^2=17.9% for Earnings and Education and 14.4% for lnIncome and Education). This relationship does not diminish the quality of the model beyond the potential to cause larger variances. As such, the multicollinearity is ignored hereafter.

The OLS model was then evaluated for heteroscedasticity (in Appendix B). A Breusch-Pagen test for heteroscedasticity reveals that at the 95% confidence level the null hypothesis that B1=B2=0 is rejected, indicating the presence of heteroscedasticity. To account for heteroscedasticity, White's robust standard errors were calculated for the base OLS model (results as follows).

Call: rlm(formula = Y.N ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11, data = atus.data)

Residuals:

Min	1Q	Median	3Q	Max
-2.54406	-0.62323	-0.01177	0.62815	3.01040

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	3.070993e+00	1.331436e-01	23.065267	1.033741e-117
X1	-1.140000e-01	1.138273e-02	-10.015167	1.307413e-23
X2	3.842719e-02	5.790034e-03	6.636781	3.206081e-11
X3	2.135430e-03	6.605142e-04	3.232982	1.225054e-03
X4	-5.494919e-02	1.718032e-02	-3.198380	1.382019e-03
X5	-5.207142e-02	2.824505e-03	-18.435589	6.807288e-76
X6	-3.776427e-05	1.552896e-05	-2.431861	1.502147e-02
X7	-1.723646e-01	3.820214e-02	-4.511909	6.424673e-06
X8	5.232400e-01	6.068622e-02	8.622055	6.576282e-18
X9	2.166652e-01	8.410715e-03	25.760618	2.451196e-146
X10	4.959100e-01	1.772565e-02	27.976970	3.098007e-172
X11	2.341174e-01	1.710716e-02	13.685347	1.242133e-42

Residual standard error: 0.928 on 15607 degrees of freedom

These heteroscedastic robust standard errors are used to construct the confidence intervals around the model.

Finally, an attempt was made to correct for self-reporting errors in the data. A meta-analysis of the relevant studies evaluated by Newell, Girgis, Fisher, & Savolainen (1999) provides some useful insight into self-reporting measurement errors. The studies that were relevant to this research show a consistent self-reporting bias (when compared to what the authors considered a

gold standard of objective reporting). Their data, graphed below looks very much like the inverse of the data published by Kruger & Dunning (1999).

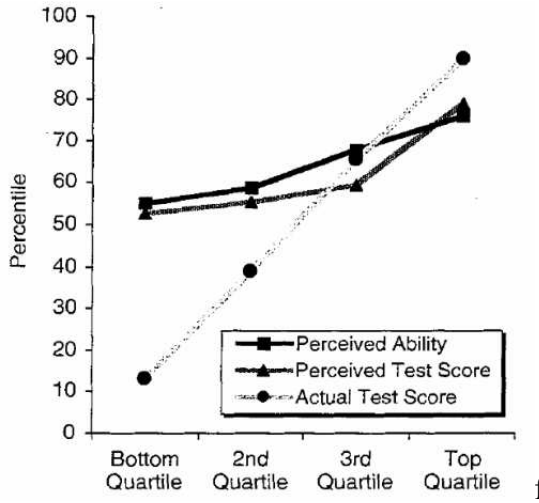


fig 1

Perceived logical reasoning ability and test performance as a function of actual test performance (Kruger & Dunning, 1999).

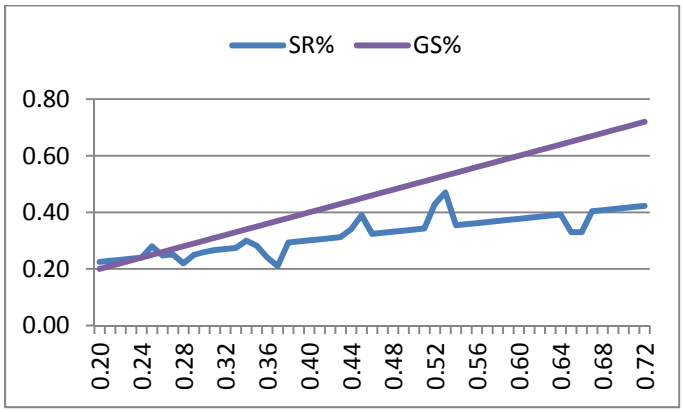


fig 2

Self-Reported versus Gold Standard (Newell, et al, 1999)

By estimating regressions for both scenarios in the data of Newel, et al (over reporting a positive attribute and under reporting a negative attribute) the following estimates were produced:

For Negative aspects (e.g. obesity)

Actual=-.21814max+2.088403Self-Reported

For Positive aspects (e.g. exercise)

Actual=-0.87026564max+2.088402505Self-Reported

The positive over-reporting model, when graphed, better illustrates the similarity with Kruger & Dunning's (1999) findings.

fig 3

These models are adjusted by multiplying the intercept by the maximum measured value to account for the fact that they are both built from percentage reporting data. They can be used for percentages by simply removing the "max" term. Had the self-reporting errors been simply a question of misremembering the information the self-reported responses would be randomly distributed about the actual values and the slope of the regression would not be significantly different from zero. Since the slopes are 2.088 (with a p-value of 7.97344E-19), this is not the case, and the data suggests that self-reported data is misreported for other than random reasons (self-reporting bias); most likely a desire (conscious or not) to under-report at risk behavior, and over-report positive traits.

These models were applied to the self-reported continuous data with the most likelihood of measurement error in the ATUS data; Earnings and Education. The new data was then evaluated via the prior methodology in R. No appreciable change to the R^2 or F stat from the base model was noted using the modified data set.

Although a useful idea and one that bears further study, it proved to be of little additional worth in this study.

Initial Findings

The regression detailed above resulted in an equation in the generic form of:

$$H_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \beta_7 X_{7i} + \beta_8 X_{8i} + \beta_9 X_{9i} + \beta_{10} X_{10i} + \beta_{11} X_{11i} + \varepsilon_i$$

And with coefficients (for the non-robust version)

Coefficients:

	Estimate
(Intercept)	3.004e+00
X1	-1.072e-01
X2	3.593e-02
X3	2.067e-03
X4	-5.720e-02
X5	-4.870e-02
X6	-4.343e-05
X7	-1.643e-01
X8	5.194e-01
X9	2.090e-01
X10	4.714e-01
X11	2.371e-01

And (for the robust version):

Coefficients:

	Value
(Intercept)	3.070993e+00
X1	-1.140000e-01
X2	3.842719e-02
X3	2.135430e-03
X4	-5.494919e-02
X5	-5.207142e-02
X6	-3.776427e-05
X7	-1.723646e-01
X8	5.232400e-01
X9	2.166652e-01
X10	4.959100e-01
X11	2.341174e-01

Where:
X1="LnFamInc" (the natural log of family income)
X2="HH Size" (the household size)
X3="Age" (the age of the respondent)
X4="Married" (whether the respondent is married {1=yes})
X5="EdYrs" (years of education for the respondent)
X6="ErnWeek" (earnings per week for the respondent only)
X7="KidUnd1" (number of children under 1 year of age in the household)

X8="PhysDiff" (whether the respondent reported physical difficulty {1=yes})
X9="Rested" (whether the respondent reported being rested {1=very, 4=not at all})
X10="HighBP" (whether the respondent reported having high blood pressure {1=yes})
X11="Painmed" (whether the respondent took pain medication {1=yes})
Y="GenHealth" (General health of the respondent {1=excellent, 5=poor})

These coefficients are all consistent with the underlying assumptions. Increased income (as measured by X1 and X6) results in better overall health; increased household size results in poorer health; health decreases with age; being married improves overall health; more education improves health; having kids under 1 in the household improves health (a weak collinearity with age {R^2~2%}); Increases in physical difficulty, consumption of pain meds, and high blood pressure all reduce health; and being more rested improves health.

All of the coefficients being the correct sign (as predicted by theory) suggests that the model is well specified and doesn't suffer from large errors associated with measurement, heteroscedasticity, outliers, or too many omitted variables (although it is recognized that additional predictive variables exist and should be pursued). This is useful in several ways; first, it indicates that the underlying theory that these types of variables can predict overall health is sound; second, it shows that even though we have recognized measurement error (in the form of self-reporting bias) that those errors are not large enough to reduce the predictive power of the model.

More predictors would act to increase R^2, but would need to be chosen carefully so as not to do so at the cost of variance.

Using these values to build a confidence interval about the mean predicted y yields

$$\hat{y} \pm t_{\frac{\alpha}{2},n-k-1} * s_{\hat{p}}$$

For the robust regression (on 15607 d.f.):

$$\bar{y} \pm 2.2416 * 0.928$$
$$\bar{y} \pm 2.0802$$

This results in a possible outcome; assuming a predicted Y of 3 (mid-range); of the actual value being between 1 and 5, 95% of the time; essentially making this model not useful for simple predictive purposes, however this model is further evaluated in the model testing phase of this analysis and found to be useful in certain specific cases of prediction.

Additional Data

State by state data are also analyzed in Appendix D (and briefly discussed here) to assist in determining where public funding could best be spent. Self-reported health level data is not available so obesity rate stands as a proxy for it.

Based on regression analysis, obesity is found to explain (R^2) 61.16% of the change in physical activity (obesity having been found to correlate with hypertension, high cholesterol, smoking, and diabetes at a p level of 0.000). Obesity was found to be normally distributed among the states (and the District of Columbia).

A cutoff of 20% was chosen and the states that fall above that level of reported obesity are included in the state level analysis section of Appendix D. These states all have an obesity rate greater than 30.8%, however do not necessarily represent the states with the largest obese population. Those states are found within the group of states with the largest overall population, and include California, Texas, Florida, and New York.

Model Testing and Conclusion

The robust OLS model was used in a Monte Carlo simulation to assess its ability to predict overall health. A random sampling of individual attributes was taken to form 20 simulated observations, that data was then used to predict General Health. The simulated observations were then compared to the actual data and where possible (where an actual observation contained the same values as the simulated observation) an actual observed value of General Health was obtained; where there was no perfect fit among the real observations, those observations that closely fit the simulated data (within 1 unit for discrete variables, and within 1 SE for continuous variables) were used to develop an estimate of General Health. Those actual and estimated General Health values were then regressed against the predicted values from the simulated data. This testing method resulted in a better than expected ability to predict overall health. Based on the quality of the underlying model an R^2 of ~17% is expected and an R^2 of ~20% is achieved. Increasing the number of iterations should result in the achieved R^2 tending toward 17%.

A small random sample of my coworkers (8 observations) resulted in an R^2 of 24.5% between predicted and actual General Health. Based on the limited sample size it appears as though this result is in line with the model developed from the ATUS data; as such no further analysis of this small sample (casually obtained) is warranted.

This model has some predictive power, but is not able to accurately predict General Health for individual observation with a narrow enough confidence interval to be useful. By loosening the requirements for the model to predict at the 95% confidence level and allowing predictions

within ±1t (one standard error) of the actual (68% confidence) the model has enough power to be able to inform policy decisions as it can predict General Health with some accuracy on average at this level. This predictive power (68% confidence) can be useful, as it still predicts that on average the outcome will be within 1 point of actual more often than not.

The intent of the model is not to predict individual outcomes, but rather to predict the effect of certain attributes on the average health of all observed. As it stands, the model does this, however, further refinement through the addition of other predictor variables and through objectively collected data can help to improve its accuracy at higher confidence levels.

Several of the predictors suffer from bi-directional causality (e.g. poor health can cause high blood pressure; high blood pressure is an indicator of poor overall health) so it is difficult to assess where money should be spent with regard to them. They are still useful in predicting outcomes and should not be discounted.

Based on the state level analysis finding that public monies would be best spent on education in the most populous states (see Appendix D), and the findings of the analysis regarding predictors and overall health, it is recommended that funding for education and specific assistance be spent on lower income individuals with less formal education and on older individuals.

References

ATUS Tables. (2014, January 1). Retrieved September 27, 2015, from http://www.bls.gov/tus/tables.htm

Abu-Omar, K., & Rütten, A. (2008). Relation of leisure time, occupational, domestic, and commuting physical activity to health indicators in Europe. *Preventive Medicine,* 319-323. Retrieved July 14, 2015, from http://www.sciencedirect.com/science/article/pii/S0091743508001539

American Time Use Survey-X extract builder. (n.d.). Retrieved September 27, 2015, from https://www.atusdata.org/

CDC Health Disparities and Inequalities Report. (2013, November 22). Retrieved July 19, 2015, from http://www.cdc.gov/mmwr/pdf/other/su6203.pdf

Caspersen, C., Powell, K., & Christenson, G. (1985, April 1). Physical activity, exercise, and physical fitness: Definitions and distinctions for health-related research. Retrieved July 19, 2015, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1424733/

Changing demographics implications for physicians, nurses, and other health workers. (2003). Rockville, Md.: U.S. Dept. of Health and Human Services, Health Resources and Services Administration, Bureau of Health Professions, National Center for Health Workforce Analysis.

Crimmins, E., & Cohen, B. (Eds.). (2011). Explaining Divergent Levels of Longevity in High-Income Countries. Retrieved July 19, 2015, from http://www.nap.edu/catalog/13089/explaining-divergent-levels-of-longevity-in-high-income-countries

Databases, Tables & Calculators by Subject. (2015, January 1). Retrieved July 17, 2015, from http://data.bls.gov/

Fatoye, F. (2013). Editorial: Understanding of health economics among healthcare professionals. *Journal of Clinical Nursing, 22*, 2979-2980. Retrieved July 14, 2015, from http://pn8vx3lh2h.search.serialssolutions.com/?ctx_ver=Z39.88-2004&ctx_enc=info:ofi/enc:UTF-8&rfr_id=info:sid/summon.serialssolutions.com&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&rft.genre=article&rft.atitle=Editorial: Understanding of health econ

Fletcher, M. (2013, March 10). Research Ties Economic Inequality to Gap in Life Expectancy. Retrieved July 19, 2015, from http://www.washingtonpost.com/business/economy/research-ties-economic-inequality-to-gap-in-life-expectancy/2013/03/10/c7a323c4-7094-11e2-8b8d-e0b59a1b8e2a_story.html

Fuchs, V. (2000). The future of health economics. *Journal of Health Economics, 19*(2), 141-157. Retrieved July 14, 2015, from http://pn8vx3lh2h.search.serialssolutions.com/?ctx_ver=Z39.88-2004&ctx_enc=info:ofi/enc:UTF-8&rfr_id=info:sid/summon.serialssolutions.com&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&rft.genre=article&rft.atitle=The future of health economics&rft.jtitle

Glanz, K. (2002). Health behavior and health education: Theory, research, and practice (3rd ed.). San Francisco: Jossey-Bass.

Health Status and Risk Factors. (2013, May 30). Retrieved July 19, 2015.

Hummer, R., & Hernandez, E. (2013, June 1). The Effect of Educational Attainment on Adult Mortality in the U.S. Retrieved July 19, 2015, from http://www.prb.org/Publications/Reports/2013/us-educational-attainment-mortality.aspx

Jiang, Y., & Hesser, J. (2006). Associations between health-related quality of life and

demographics and health risks. Results from Rhode Island's 2002 behavioral risk factor survey. *Health and Quality of Life Outcomes, 14*(4). Retrieved July 14, 2015, from http://www.hqlo.com/content/4/1/14

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, Vol. 77, No. 6., 1121-1134. Retrieved October 31, 2015, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.64.2655&rep=rep1&type=pdf

Kruschke, J. (2014, November 21). Doing Bayesian Data Analysis. Retrieved October 26, 2015.

Lleras-Muney, A. (2005). The Relationship between Education and Adult Mortality in the United States. Retrieved July 19, 2015.

Newell, S., Girgis, A., Sanson-Fisher, R., & Savolainen, N. (1999). The accuracy of self-reported health behaviors and risk factors relating to cancer and cardiovascular disease in the general population. *American Journal of Preventive Medicine*, 17(3), 211-229. Retrieved October 28, 2015, from http://www.ajpmonline.org/article/S0749-3797(99)00069-0/pdf

Marmot, M. (2002). The Influence Of Income On Health: Views Of An Epidemiologist. *Health Affairs,* 31-46. Retrieved July 14, 2015, from http://content.healthaffairs.org/content/21/2/31.full.pdf html

Menec, V., & Chipperfield, J. (1997). Remaining Active in Later Life: The Role of Locus of Control in Seniors' Leisure Activity Participation, Health, and Life Satisfaction. *Journal of Aging and Health,* 105-125. Retrieved July 14, 2015, from http://jah.sagepub.com.ezproxy.snhu.edu/content/9/1/105.full.pdf html

National Data. (2015, January 1). Retrieved July 17, 2015, from http://www.bea.gov/iTable/iTable.cfm?ReqID=9&step=1#reqid=9&step=3&isuri=1&90

4=2014&903=58&906=a&905=1929&910=x&911=0

Olshansky, S., Antonucci, T., Berkman, L., Binstock, R., Boersch-Supan, A., Cacioppo, J., . . . Rowe, J. (2012). Differences In Life Expectancy Due To Race And Educational Differences Are Widening, And Many May Not Catch Up. Health Affairs, 31(8), 1803-1813. Retrieved July 16, 2015, from http://content.healthaffairs.org/content/31/8/1803.abstract

United States Census Bureau. (2014, January 1). Retrieved July 17, 2015, from http://www.census.gov/hhes/www/cpstables/032014/health/h01_000.htm

Winkleby, M., Jatulis, D., Frank, E., & Fortmann, S. (1992). Socioeconomic Status And Health: How Education, Income, And Occupation Contribute To Risk Factors For Cardiovascular Disease. *American Journal of Public Health,* 816-820. Retrieved July 14, 2015, from http://ajph.aphapublications.org/doi/pdf/10.2105/AJPH.82.6.816

Appendix A: Descriptive Statistics

Each of the continuous variables used were evaluated for normality and central tendency. The discrete variables do not lend themselves to these evaluations so are discussed further after the continuous variables.

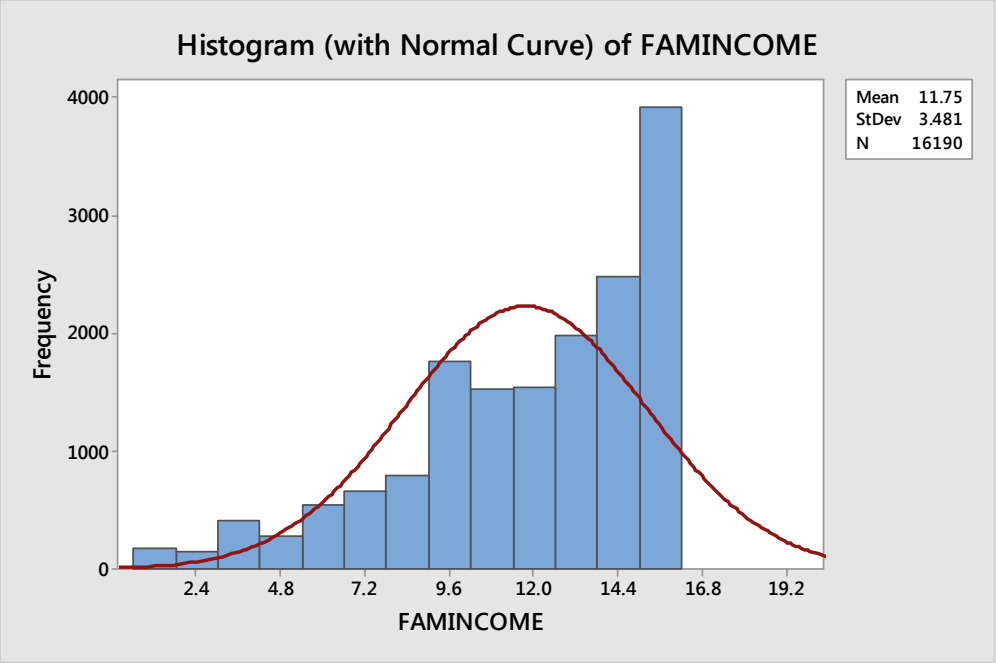
Descriptive Statistics: FAMINCOME, AGE, EDUCYRS, EARNWEEK, UHRSWORKT

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
FAMINCOME	16190	0	11.746	0.0274	3.481	1.000	10.000	13.000	14.000	16.000
AGE	16190	0	42.902	0.0982	12.500	17.000	33.000	42.000	52.000	85.000
EDUCYRS	16190	0	14.491	0.0229	2.919	0.0000	12.000	14.000	16.000	19.000
EARNWEEK	16190	0	946.08	5.22	663.72	0.00	469.20	769.23	1240.00	2884.61
UHRSWORKT	16190	0	40.873	0.0924	11.755	0.0000	40.000	40.000	45.000	110.000

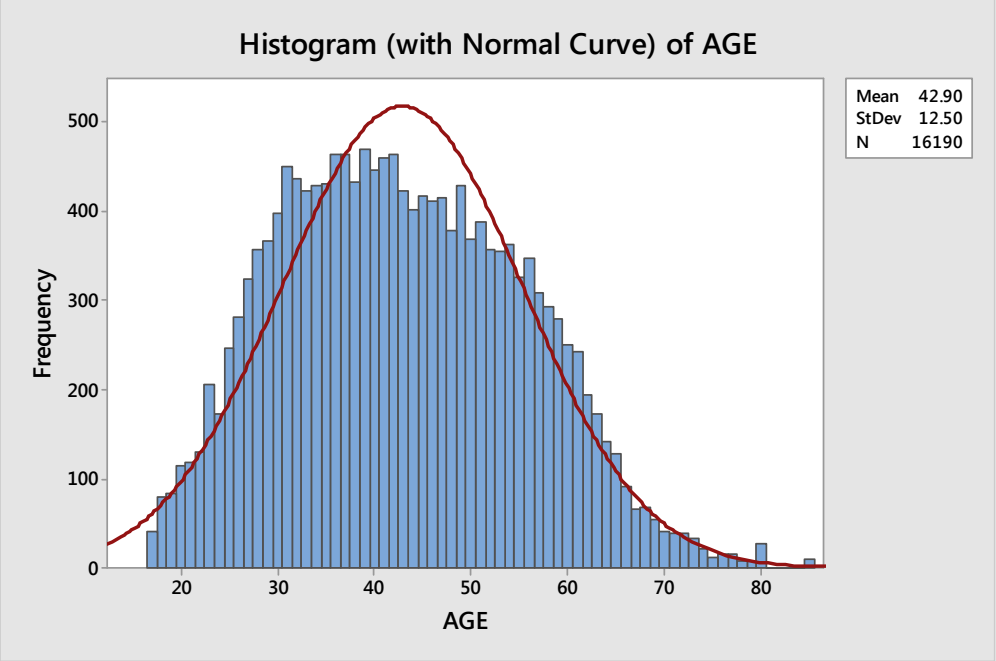
Variable	Range	Skewness	Kurtosis
FAMINCOME	15.000	-0.96	0.34
AGE	68.000	0.24	-0.54
EDUCYRS	19.000	-0.68	1.56
EARNWEEK	2884.61	1.24	1.16
UHRSWORKT	110.000	0.16	2.92

Family income ranges from 1 to 16, these act as bins as follows:

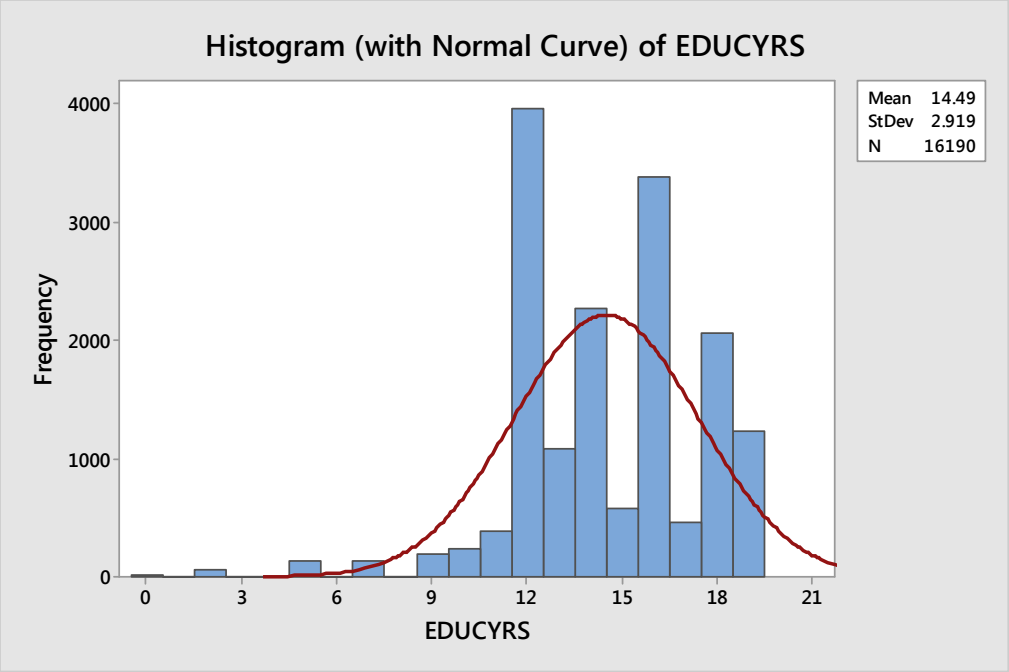
FAMINCOME	Family income
001	Less than $5,000
002	$5,000 to $7,499
003	$7,500 to $9,999
004	$10,000 to $12,499
005	$12,500 to $14,999
006	$15,000 to $19,999
007	$20,000 to $24,999
008	$25,000 to $29,999
009	$30,000 to $34,999
010	$35,000 to $39,999
011	$40,000 to $49,999
012	$50,000 to $59,999
013	$60,000 to $74,999
014	$75,000 to $99,999
015	$100,000 to $149,999
016	$150,000 and over



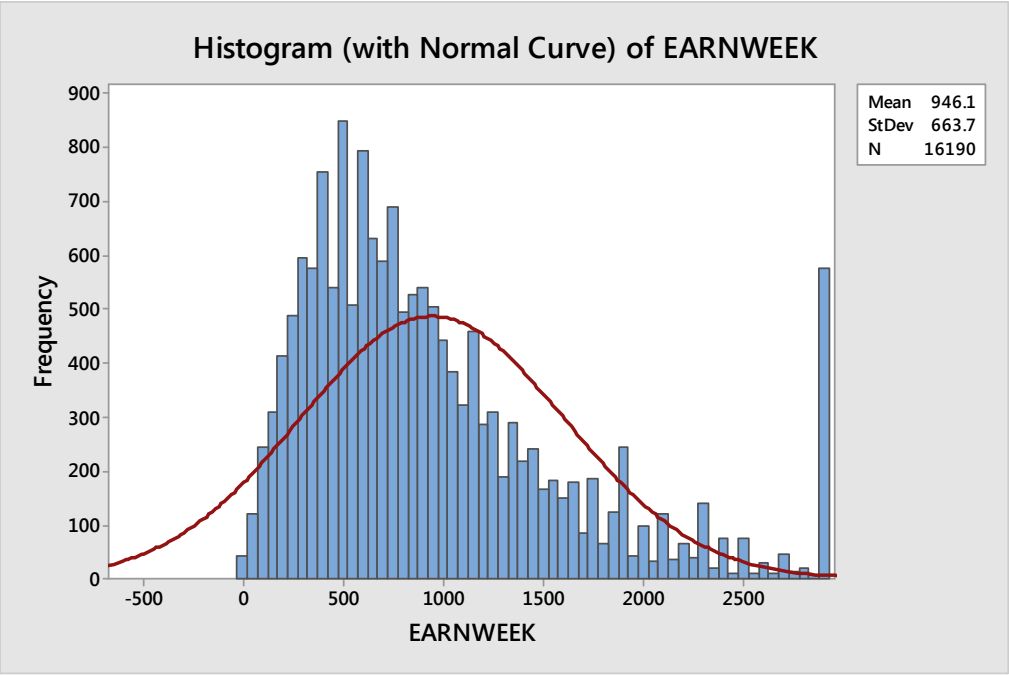
The data is left skewed for Family income. The removal of the $2884.61 values from weekly earnings reduces the size of the spike in that category (see below).



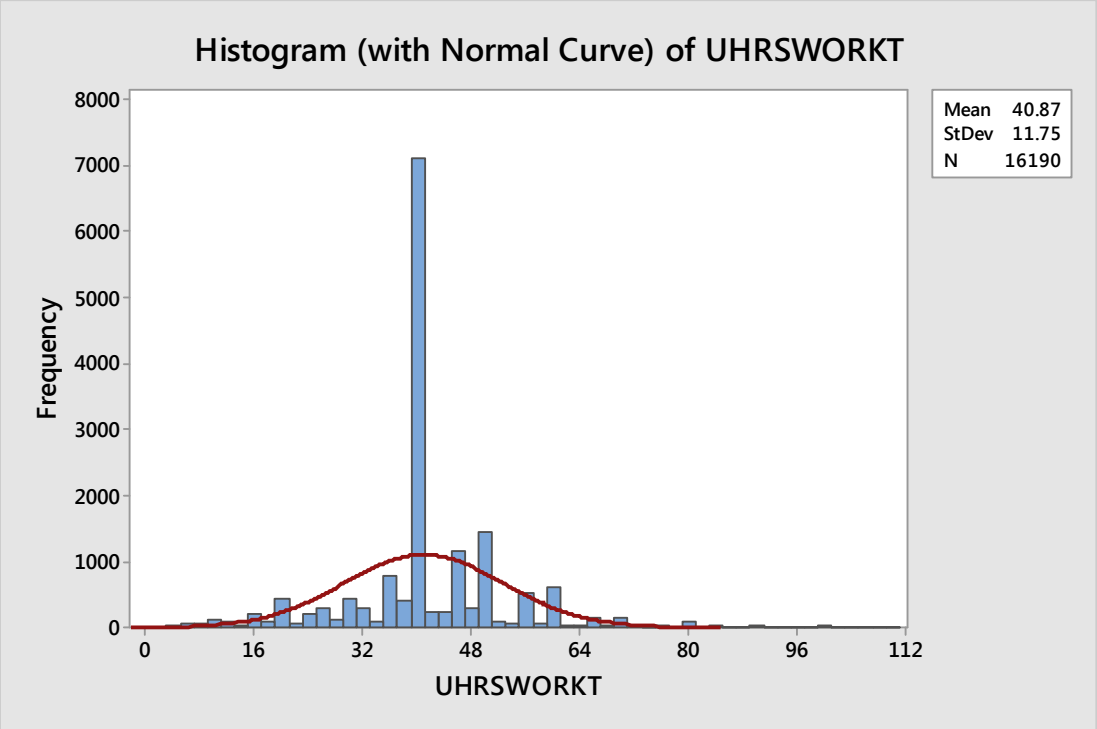
Age is slightly right skewed, but very closely follows a normal distribution.



Education has spikes at 12 years, 14 years, 16 years, and 18 years which correspond to high school, associate's degree, bachelor's degree, and master's degree. The data is close enough to normally distributed that it is treated as not violating that assumption of the OLS model.



Weekly Earnings differs from Family Income in that it is only measuring the earnings of the respondent, whereas Family Income measures the income of all householders. The data is right skewed and has a large spike at $2884.61. That category acts as a catch all for all larger values. Those values have been removed from the model to better fit the assumptions of OLS.

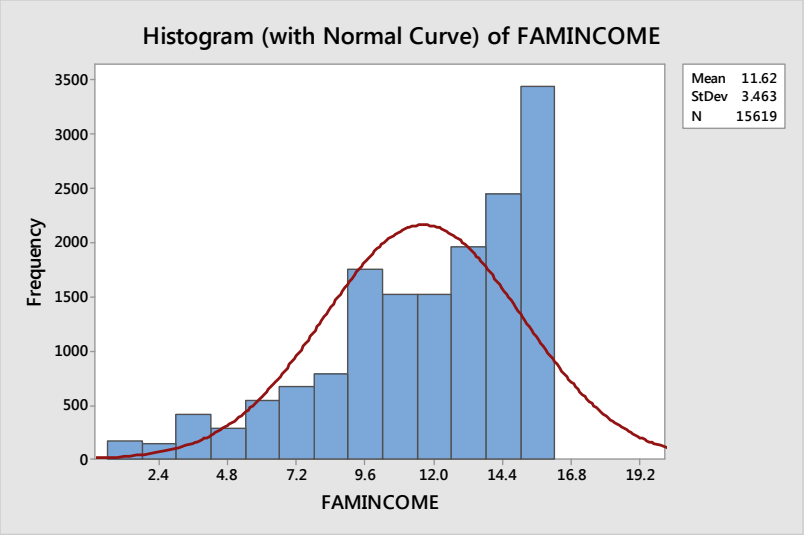


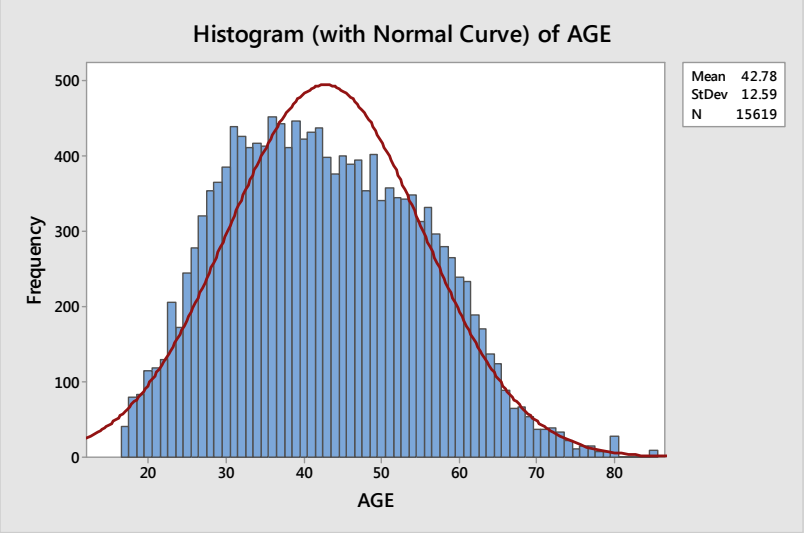
There is a large spike at 40 hours worked per week. The remainder of the data closely follows a normal distribution.

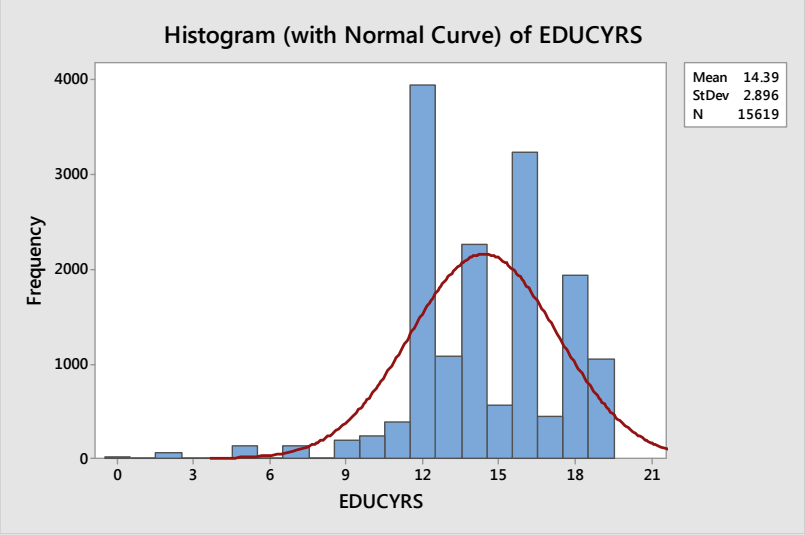
With the peak removed from Weekly Earnings the descriptive statistics and graphs are as follows. Household size was added to this group due to its nearly continuous nature.

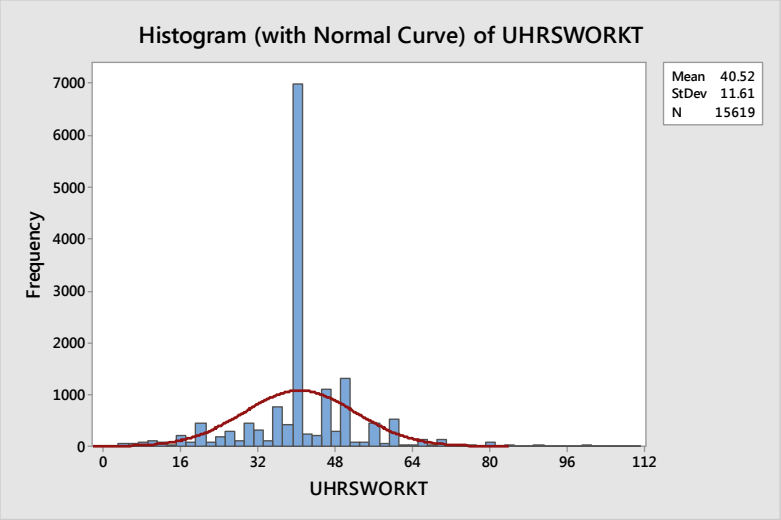
Descriptive Statistics: FAMINCOME, AGE, EDUCYRS, UHRSWORKT, EARNWEEK

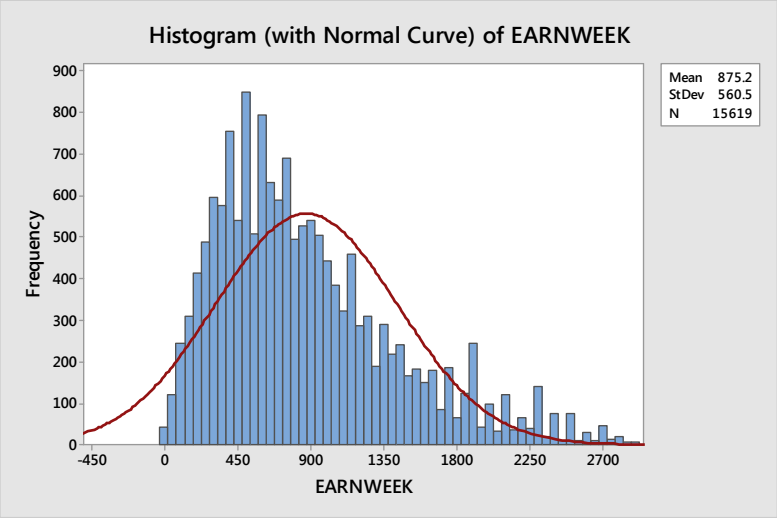
Variable	N	Mean	SE Mean	StDev	Minimum	Maximum
FAMINCOME	15619	11.618	0.0277	3.463	1.000	16.000
AGE	15619	42.782	0.101	12.588	17.000	85.000
EDUCYRS	15619	14.392	0.0232	2.896	0.0000	19.000
UHRSWORKT	15619	40.523	0.0929	11.613	0.0000	110.000
EARNWEEK	15619	875.21	4.49	560.55	0.00	2884.50
HH_SIZE	15619	2.8742	0.0120	1.4949	1.0000	15.0000

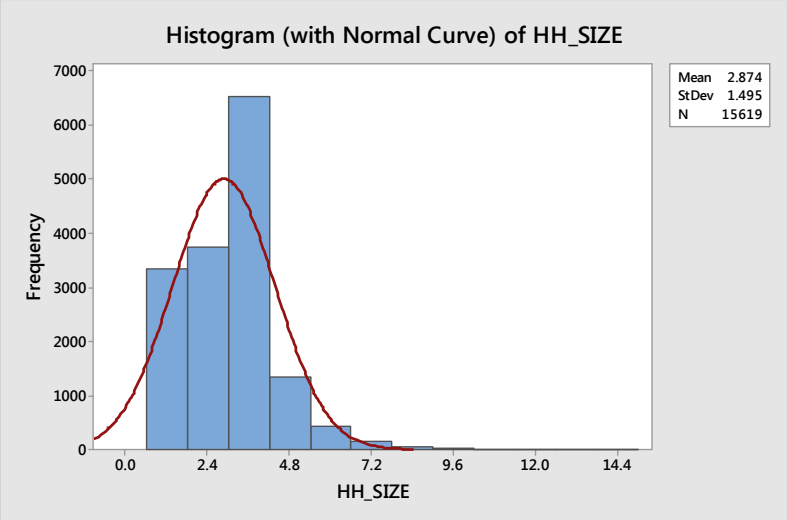

Histogram (with Normal Curve) of FAMINCOME
Mean 11.62
StDev 3.463
N 15619
Frequency
FAMINCOME
0
500
1000
1500
2000
2500
3000
3500
2.4
4.8
7.2
9.6
12.0
14.4
16.8
19.2


Histogram (with Normal Curve) of AGE
Mean 42.78
StDev 12.59
N 15619
Frequency
AGE
0
100
200
300
400
500
20
30
40
50
60
70
80


Histogram (with Normal Curve) of EDUCYRS
Mean 14.39
StDev 2.896
N 15619
Frequency
EDUCYRS
0
1000
2000
3000
4000
0
3
6
9
12
15
18
21

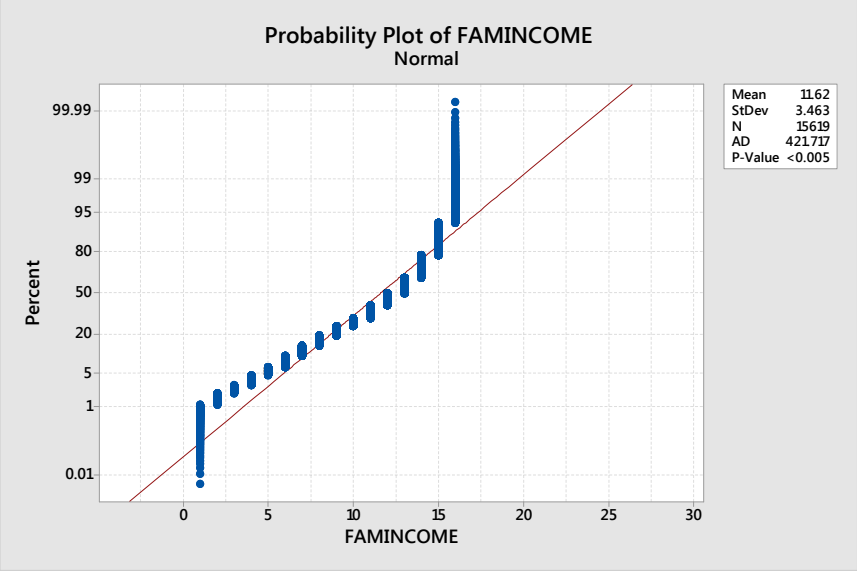

Histogram (with Normal Curve) of UHRSWORKT
Mean 40.52
StDev 11.61
N 15619
Frequency
UHRSWORKT
0
1000
2000
3000
4000
5000
6000
7000
0
16
32
48
64
80
96
112


Histogram (with Normal Curve) of EARNWEEK
Mean 875.2
StDev 560.5
N 15619
Frequency
EARNWEEK
0
100
200
300
400
500
600
700
800
900
-450
0
450
900
1350
1800
2250
2700

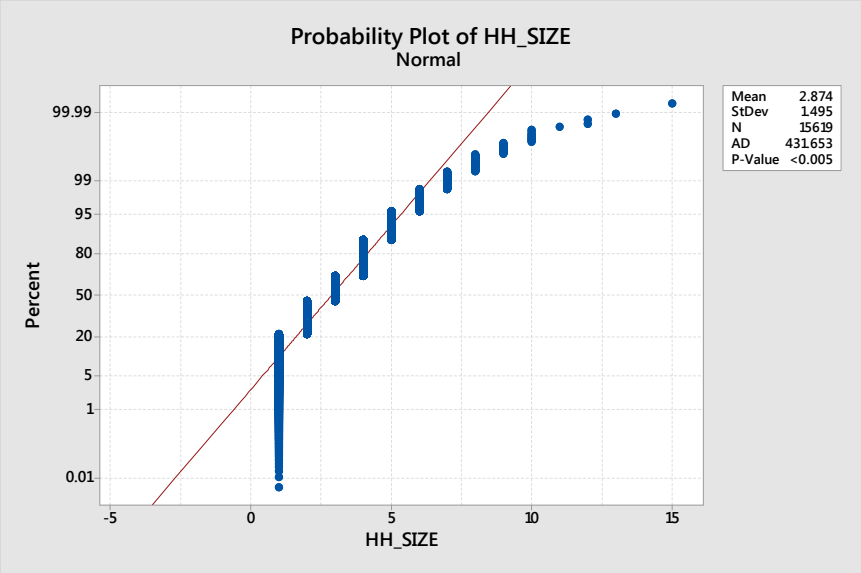

Histogram (with Normal Curve) of HH_SIZE
Mean 2.874
StDev 1.495
N 15619
Frequency
HH_SIZE
0
1000
2000
3000
4000
5000
6000
7000
0.0
2.4
4.8
7.2
9.6
12.0
14.4

After removing the non-relevant observations, each of these continuous variables was evaluated for normality using an Anderson-Darling test (previous statements on normality are based on visual fit of data with a normal distribution).

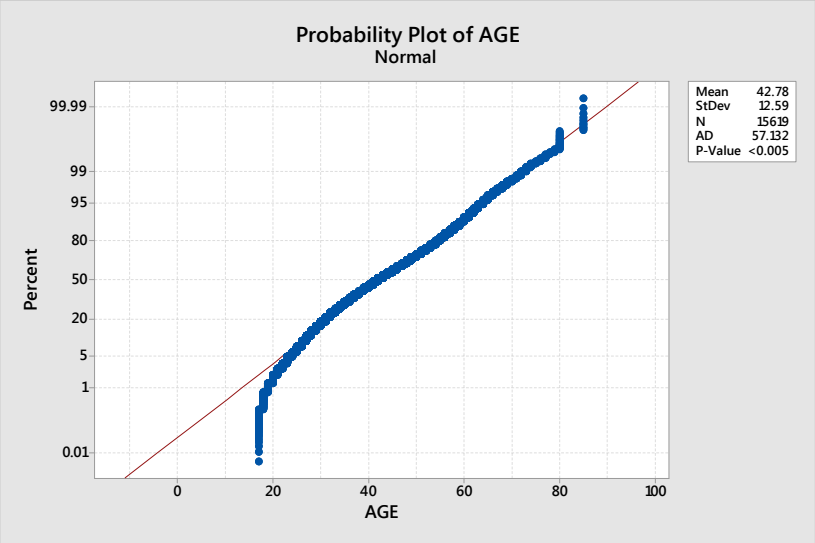
For all Anderson-Darling tests a null hypothesis of not normally distributed is used with a corresponding p-value of 0.05. For those values listed below, the null hypothesis is rejected where p-value<0.05.



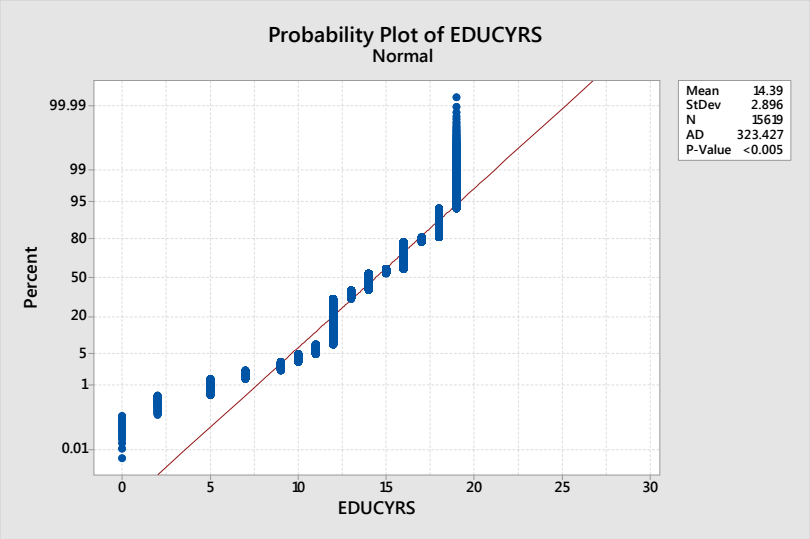
Family Income is normally distributed.



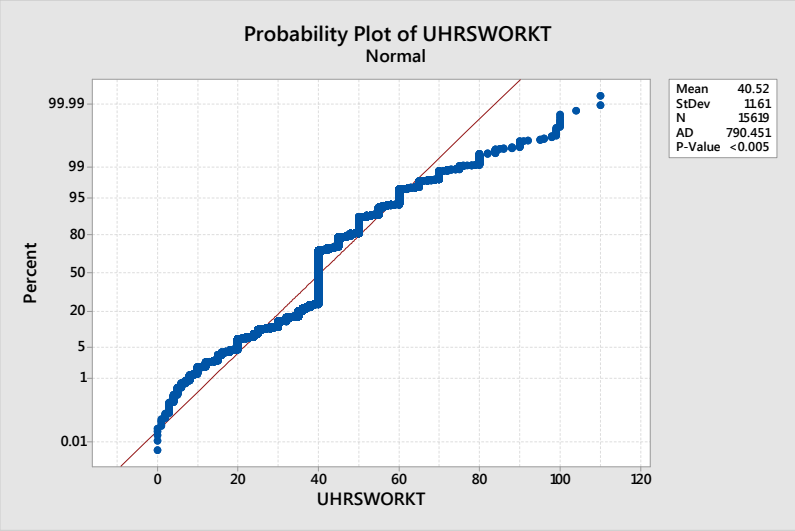
Household size is normally distributed.



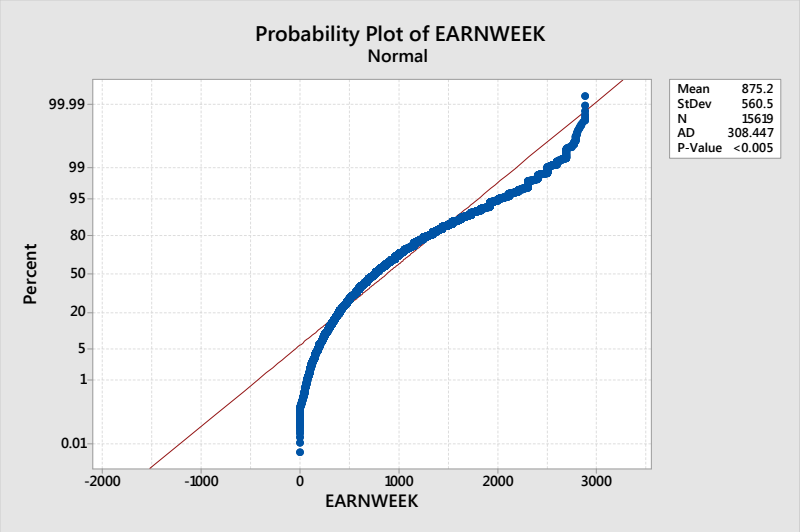
Age is normally distributed.



Years of education is normally distributed.



Usual hours worked per week is normally distributed.



Earnings per week is normally distributed.

Discrete Variables

Descriptive Statistics: KIDUND1, DIFFPHYS, HighBP_1, Painmed_1, Married

Variable	N	Minimum	Median	Maximum	IQR	Mode	N for Mode
KIDUND1	15619	0.00000	0.00000	1.00000	0.00000	0	15022
DIFFPHYS	15619	1.0000	1.0000	2.0000	0.000000	1	15394
HighBP_1	15619	0.00000	0.00000	1.00000	0.00000	0	11772
Painmed_1	15619	0.00000	0.00000	1.00000	0.00000	0	11806
Married	15619	0.00000	1.00000	1.00000	1.00000	1	8418

These statistics don't mean much for us in our evaluation, a better descriptive for those discrete variables (except Rested which is a Likert scale) would be the percentage of them that meet certain criteria.

96% have no children under 1yr old in the household

99% do not report physical difficulty in their daily activities

75% report normal blood pressure

75% report that they do not routinely take pain medication

53% are married.

The remaining two variables, Rested and General Health (dependent) are reported on a Likert scale, as such they should closely model interval data allowing us to treat them as such.

Descriptive Statistics: RESTED, GENHEALTH

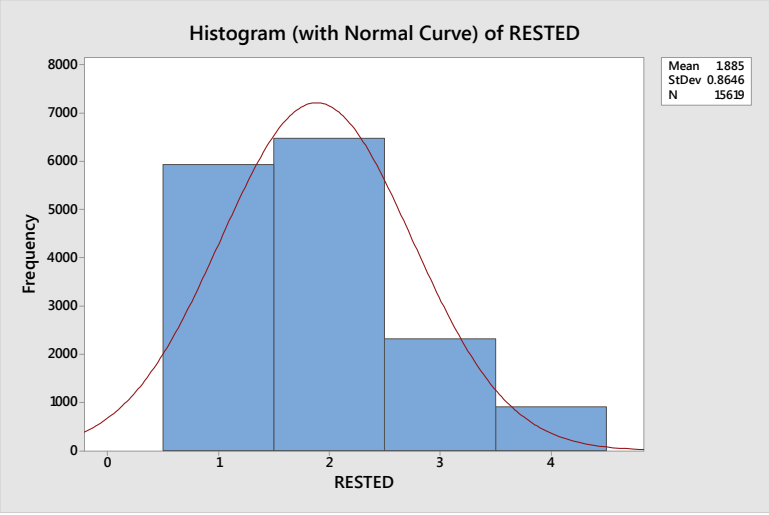
Variable	N	Mean	SE Mean	StDev	Minimum	Median	Maximum	Skewness	Kurtosis
RESTED	15619	1.8848	0.00692	0.8646	1.0000	2.0000	4.0000	0.76	-0.09
GENHEALTH	15619	2.3453	0.00755	0.9442	1.0000	2.0000	5.0000	0.30	-0.42

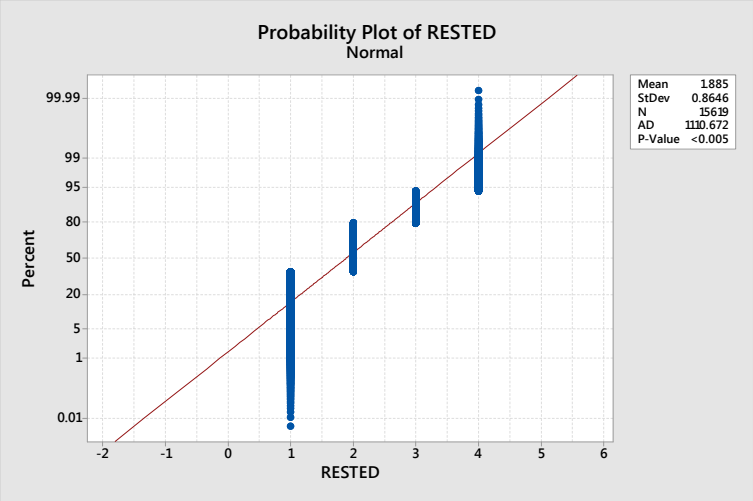
Where:

RESTED	Well-rested yesterday
01	Very
02	Somewhat
03	A Little
04	Not at all

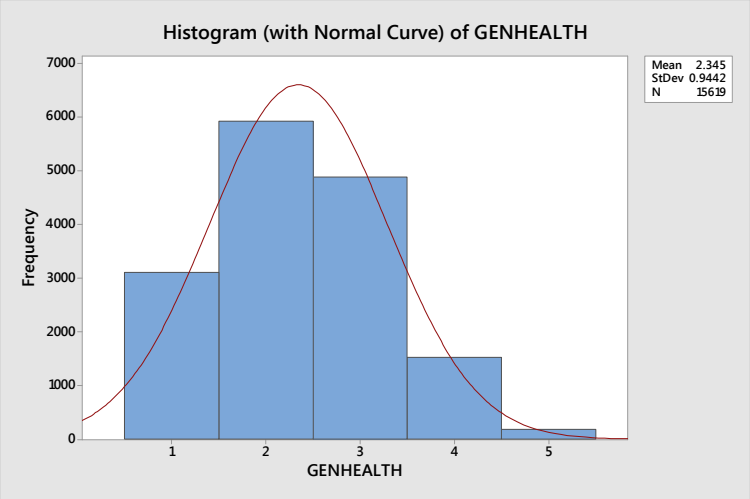
And

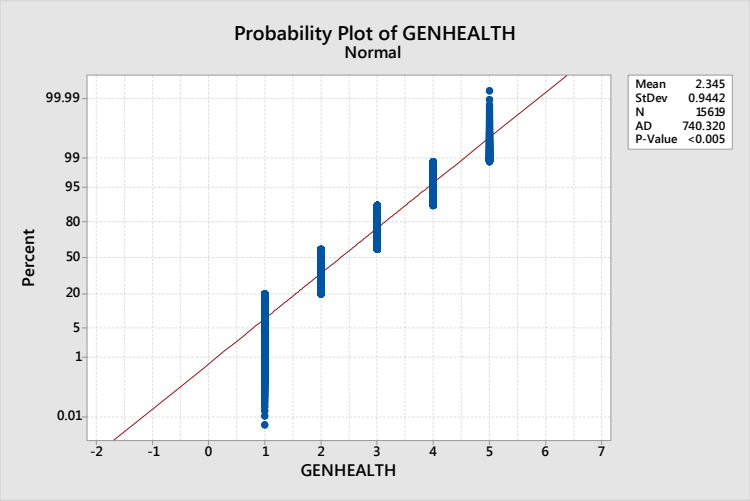
GENHEALTH	General health
01	Excellent
02	Very good
03	Good
04	Fair
05	Poor





Higher numbers represent reporting being more rested. Rested appears to closely follow a normal distribution with a right skew.





General Health, the dependent variable in our model, follows a normal distribution. The data has a slight right skew.

All of our data fits the assumptions for an ordinary least squares linear regression model. Although that is not the only model that will be pursued, the assumptions for it are closely mirrored by the assumptions in the other models. Testing of further assumptions, e.g. multicollinearity and heteroscedasticity, will be conducted in the model building portion of this analysis.

Appendix B: Model Building

The data is coded as below for the analysis performed. Most of the analysis was performed using R. Use of other software packages will be noted where appropriate.

X1="LnFamInc" (the natural log of family income)
X2="HH Size" (the household size)
X3="Age" (the age of the respondent)
X4="Married" (whether the respondent is married {1=yes})
X5="EdYrs" (years of education for the respondent)
X6="ErnWeek" (earnings per week for the respondent only)
X7="KidUnd1" (number of children under 1 year of age in the household)
X8="PhysDiff" (whether the respondent reported physical difficulty {1=yes})
X9="Rested" (whether the respondent reported being rested {1=very, 4=not at all})
X10="HighBP" (whether the respondent reported having high blood pressure {1=yes})
X11="Painmed" (whether the respondent took pain medication {1=yes})
Y="GenHealth" (General health of the respondent {1=excellent, 5=poor})

General Health is a categorical variable with the scores being as follows:

1 "Excellent"
2 "Very Good"
3 "Good"
4 "Fair"
5 "Poor"

This closely approximates a Likert scale, and as such will be treated as interval data for the purposes of this analysis. An Ordered Logistic regression was performed, and is included at the end of this appendix, to further assess the strength of the relationship. The first attempt at a model was built around a multiple linear regression of Health versus the previously discussed predictors, but with Family Income not taken as a natural logarithm. That model was thrown out with an R^2 of approximately 14% in favor of the model below; which uses the log of Family Income and has an adjusted R^2~17.3%.

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11, data = atus.data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.53392 -0.64181 -0.03292  0.61695  2.98877

Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  3.004e+00  1.284e-01   23.391   < 2e-16 ***
X1           -1.072e-01 1.098e-02   -9.759   < 2e-16 ***
X2           3.593e-02  5.585e-03   6.433    1.29e-10 ***
X3           2.067e-03  6.372e-04   3.245    0.001178 **
X4           -5.720e-02 1.657e-02   -3.451   0.000559 ***
X5           -4.870e-02 2.725e-03   -17.874  < 2e-16 ***
X6           -4.343e-05 1.498e-05   -2.899   0.003749 **
X7           -1.643e-01 3.685e-02   -4.459   8.28e-06 ***
X8           5.194e-01  5.854e-02   8.873    < 2e-16 ***
X9           2.090e-01  8.113e-03   25.766   < 2e-16 ***
X10          4.714e-01  1.710e-02   27.567   < 2e-16 ***
X11          2.371e-01  1.650e-02   14.368   < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8586 on 15607 degrees of freedom

Multiple R-squared:  0.1735
        Adjusted R-squared:  0.173

F-statistic: 297.9 on 11 and 15607 DF
        p-value: < 2.2e-16
```

This model is fairly accurate and should produce somewhat meaningful predictions. The ability to predict actual values with some probability is further addressed in the Logit Regression section below. This Ordinary Least Squares (OLS) model treats General Health as a continuous variable, although the Likert scale attempts to mimic such a scale, it is not, so further models are evaluated for a better fit.

The first advanced model evaluated treats the data as panel data with dummy variables for 2012 and 2013 (2010 remained the null value); whereas the base model combines all years as one set of data (a pooled cross-section). By evaluating for changes over years the R^2 increases to ~17.4% with no effect on the significance of the individual predictors. Further study should be

performed when the 2014 data becomes available to evaluate for changes over time that add value to the predictive power of the model. Since this increase in R^2 is not a significant one, the data will be treated as a pooled cross section for the rest of this analysis.

Multicollinearity

Multicollinearity is expected within this data. All of the predictors are likely to be collinear. The degree to which that is true is assessed further below. Each of these predictors has an effect on the outcome, some of those effects overlap and are likely caused by a non-evaluated underlying cause. The things that tend to cause people to have more education also tend to cause them to have higher income; however there are other aspects of those predictors that are independent of this underlying cause, and it is that aspect that creates the separate influence on overall health.

Each predictor was regressed against each of the other predictors using an R script. Only the significant betas are retained below:

```
Call:
lm(formula = X1 ~ X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 +
    X11, data = atus.data)

Coefficients:
      Estimate      Std. Error    t value       Pr(>|t|)
X2     6.746e-02    4.036e-03     16.716        < 2e-16 ***
X3     3.866e-03    4.634e-04     8.343         < 2e-16 ***
X4     3.266e-01    1.180e-02     27.686        < 2e-16 ***
X5     5.819e-02    1.931e-03     30.139        < 2e-16 ***
X6     5.298e-04    1.006e-05     52.648        < 2e-16 ***
X7    -8.734e-02    2.685e-02     -3.252        0.00115 **
X8    -2.159e-01    4.264e-02     -5.063        4.17e-07 ***
X10    -3.165e-02   1.246e-02     -2.540        0.01109 *

Call:
lm(formula = X2 ~ X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11,
    data = atus.data)

Coefficients:
      Estimate      Std. Error    t value       Pr(>|t|)
X3    -3.395e-02    8.781e-04     -38.658       < 2e-16 ***
X4     1.417e+00    2.046e-02     69.225        < 2e-16 ***
X5    -5.868e-02    3.801e-03     -15.440       < 2e-16 ***
X6    -4.889e-05    1.995e-05     -2.450        0.014290 *
X7     4.754e-01    5.313e-02     8.949         < 2e-16 ***
X8    -2.704e-01    8.454e-02     -3.199        0.001384 **
X9     3.950e-02    1.173e-02     3.369         0.000756 ***
```

Call:
lm(formula = X3 ~ X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11, data = atus.data)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
X4	2.2055649	0.1856793	11.878	< 2e-16 ***
X5	-0.1981480	0.0346058	-5.726	1.05e-08 ***
X6	0.0028993	0.0001804	16.072	< 2e-16 ***
X7	-9.9549104	0.4776460	-20.842	< 2e-16 ***
X8	8.2800671	0.7677434	10.785	< 2e-16 ***
X9	-1.6529563	0.1060565	-15.586	< 2e-16 ***
X10	9.0569232	0.2132871	42.464	< 2e-16 ***
X11	3.3483168	0.2157976	15.516	< 2e-16 ***

Call:
lm(formula = X4 ~ X5 + X6 + X7 + X8 + X9 + X10 + X11, data = atus.data)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
X6	1.403e-04	7.695e-06	18.228	< 2e-16 ***
X7	2.970e-01	2.045e-02	14.522	< 2e-16 ***
X8	-2.061e-01	3.305e-02	-6.236	4.61e-10 ***
X9	-1.108e-02	4.571e-03	-2.424	0.0154 * ---

Call:
lm(formula = X5 ~ X6 + X7 + X8 + X9 + X10 + X11, data = atus.data)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
X6	2.182e-03	3.741e-05	58.327	< 2e-16 ***
X7	3.487e-01	1.097e-01	3.179	0.00148 **
X9	1.312e-01	2.450e-02	5.353	8.76e-08 ***
X10	-2.972e-01	4.927e-02	-6.032	1.65e-09 ***

Call:
lm(formula = X6 ~ X7 + X8 + X9 + X10 + X11, data = atus.data)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
X8	-169.13	37.91	-4.462	8.19e-06 ***
X9	15.25	5.24	2.909	0.00363 **
X10	24.89	10.54	2.362	0.01820 *
X11	-28.60	10.67	-2.680	0.00738 **

```
Call:
lm(formula = X7 ~ X8 + X9 + X10 + X11, data = atus.data)

Coefficients:
      Estimate   Std. Error  t value    Pr(>|t|)
X8    -0.025895  0.012928    -2.003     0.04519 *
X9     0.015061  0.001783     8.446     2e-16 ***
X10   -0.021774  0.003590    -6.065     1.35e-09 ***
X11   -0.015665  0.003638    -4.305     1.68e-05 ***


Call:
lm(formula = X8 ~ X9 + X10 + X11, data = atus.data)

Coefficients:
      Estimate   Std. Error  t value    Pr(>|t|)
X9     0.003258  0.001103    2.952      0.00316 **
X10    0.018943  0.002217    8.544      < 2e-16 ***
X11    0.025645  0.002243    11.434     < 2e-16 ***

Call:
lm(formula = X9 ~ X10 + X11, data = atus.data)

Coefficients:
      Estimate   Std. Error  t value    Pr(>|t|)
X11    0.264581  0.016126    16.407     <2e-16 ***

Call:
lm(formula = X10 ~ X11, data = atus.data)

Coefficients:
      Estimate   Std. Error  t value    Pr(>|t|)
X11    0.140814  0.007947    17.72      <2e-16 ***
```

As expected there is significant multicollinearity between most of the independent variables. This relationship does not diminish the quality of the model beyond the potential to cause larger variances. As such, the multicollinearity is ignored hereafter.

Heteroscedasticity

At this point, the linear model is tested for heteroscedasticity using a Breusch-Pagen test in R. That test reveals the presence of heteroscedastic residuals:

studentized Breusch-Pagan test

data: lm.r
BP = 171.42, df = 11, p-value < 2.2e-16

At the 95% confidence level the null hypothesis that B1=B2=0 is rejected, indicating heteroscedasticity.

To attempt to correct for the heteroscedasticity White's robust standard errors were calculated for the base OLS model.

Call: rlm(formula = Y.N ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11, data = atus.data)

Residuals:

Min	1Q	Median	3Q	Max
-2.54406	-0.62323	-0.01177	0.62815	3.01040

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	3.070993e+00	1.331436e-01	23.065267	1.033741e-117
X1	-1.140000e-01	1.138273e-02	-10.015167	1.307413e-23
X2	3.842719e-02	5.790034e-03	6.636781	3.206081e-11
X3	2.135430e-03	6.605142e-04	3.232982	1.225054e-03
X4	-5.494919e-02	1.718032e-02	-3.198380	1.382019e-03
X5	-5.207142e-02	2.824505e-03	-18.435589	6.807288e-76
X6	-3.776427e-05	1.552896e-05	-2.431861	1.502147e-02
X7	-1.723646e-01	3.820214e-02	-4.511909	6.424673e-06
X8	5.232400e-01	6.068622e-02	8.622055	6.576282e-18
X9	2.166652e-01	8.410715e-03	25.760618	2.451196e-146
X10	4.959100e-01	1.772565e-02	27.976970	3.098007e-172
X11	2.341174e-01	1.710716e-02	13.685347	1.242133e-42

Residual standard error: 0.928 on 15607 degrees of freedom

This results in a confidence interval (about the mean $\hat{y}$) of

$$\hat{y} \pm t_{\frac{\alpha}{2},n-k-1} * s_{\hat{p}}$$

Or, using the robust regression values:

$$\bar{y} \pm 2.2416 * 0.928$$
$$\bar{y} \pm 2.0802$$

This results in a possible outcome; assuming a predicted y of 3 (mid-range); of the actual value being between 1 and 5, 95% of the time; making this model not useful for predicting individual values.

Logit regression

Since the underlying purpose of this analysis is to build a predictive model, and the confidence intervals resulting from OLS and Robust OLS are too wide to provide meaningful prediction the data was reevaluated using an Ordered Logistic Regression (logit).

Using R (with package "MASS" installed) to run an ordered logit regression yields:

Call:
polr(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11, data = atus.data, Hess = TRUE, method = "logistic")
Coefficients:

	Value	Std. Error	t value	p value
X1	-2.381787e-01	0.0159308533	-14.950780	1.539365e-50
X2	7.995671e-02	0.0120851509	6.616112	3.687690e-11
X3	4.449568e-03	0.0013656599	3.258182	1.121285e-03
X4	-1.161268e-01	0.0350334223	-3.314743	9.172729e-04
X5	-1.074634e-01	0.0059822574	-17.963692	3.750130e-72
X6	-8.372846e-05	0.0000332195	-2.520461	1.172011e-02
X7	-3.695395e-01	0.0802865703	-4.602757	4.169355e-06
X8	1.090727e+00	0.1084984708	10.052928	8.917730e-24
X9	4.489190e-01	0.0179374185	25.026958	3.111440e-138
X10	9.999811e-01	0.0373072697	26.803921	2.908482e-158
X11	4.804941e-01	0.0356368821	13.483056	1.967843e-41

Thresholds:

	Value	Std. Error	t value	p value
1|2	-3.150319e+00	0.0277451043	-113.545053	0.000000e+00
2|3	-1.231967e+00	0.0342099786	-36.011921	5.443983e-284
3|4	7.907682e-01	0.0410110787	19.281820	7.634237e-83
4|5	3.236943e+00	0.0800559777	40.433493	0.000000e+00

Residual Deviance: 38749.57
AIC: 38779.57

The parameters and thresholds in this model are not intuitive in their interpretation; R forces a zero intercept and fits the model based on that assumption. John K. Kruschke (2014) developed a function for use in R to transform ordered logit (or probit) regressions into a form that sets the threshold levels at intuitive values (one half the distance between the two adjacent levels) and fits a non-zero intercept and coefficients from there. Using that method yields:

Intercept:
2.979657
Coefficients:

	Value
X1	-1.118689e-01
X2	3.755445e-02
X3	2.089895e-03
X4	-5.454300e-02
X5	-5.047394e-02
X6	-3.932599e-05
X7	-1.735671e-01
X8	5.122981e-01
X9	2.108504e-01
X10	4.696759e-01
X11	2.256808e-01

Thresholds:

	Value
1|2	1.500000
2|3	2.401021
3|4	3.351069
4|5	4.500000

Using the thresholds and coefficients above, predicted values for Y (General Health) were obtained and regressed against actual values using R. The results show an R^2 of ~13% which indicates that this model has less predictive power than the base OLS model, however yields more intuitive results.

```
Call:
lm(formula = Y.N ~ Y.star.N)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4081 -0.7890 -0.1698  0.8302  2.8302

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.93146    0.02985   31.20  <2e-16 ***
Y.star.N     0.61917    0.01270   48.73  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8797 on 15617 degrees of freedom
Multiple R-squared:  0.132
        Adjusted R-squared:  0.132
F-statistic:  2375 on 1 and 15617 DF
        p-value: < 2.2e-16
```

The decrease in the standard error in this model over the OLS model causes a subsequent decrease in the confidence intervals (CI) about the predictions. This reduction isn't large enough to cause the CIs to reduce enough to make the predicted range within the total range. Since the gain in this model does not offset the loss in predictive power, this model is abandoned.

Measurement error

A meta-analysis of the relevant studies evaluated by Newell, Girgis, Fisher, & Savolainen (1999) provides some useful insight into self-reporting measurement errors. The studies that were relevant to this research show a consistent self-reporting bias (when compared to what the authors considered a gold standard of objective reporting). By estimating regressions for both scenarios in the data of Newel, et al (over reporting a positive attribute and under reporting a negative attribute) the following estimates were produced:

For Negative aspects (e.g. obesity)

Actual=-.21814max+2.088403Self-Reported

For Positive aspects (e.g. exercise)

Actual=-0.87026564max+2.088402505Self-Reported

These models are adjusted by multiplying the intercept by the maximum measured value to account for the fact that they are both built from percentage reporting data. They can be used for percentages by simply removing the "max" term.

These models are applied to the self-reported continuous data with the most likelihood of measurement error in the ATUS data; Earnings and Education. The new data is evaluated via the prior methodology in R. There is no appreciable change to the R^2 or F stat from the base model.

```
Residual standard error: 0.8586 on 15607 degrees of freedom
Multiple R-squared:  0.1735,
        Adjusted R-squared:  0.173
F-statistic: 297.9 on 11 and 15607 DF,
        p-value: < 2.2e-16
```

Although a useful idea, and one that bears further study, it proved to be of little additional value in this study.

Appendix C: Model Testing

The robust OLS model was used in a Monte Carlo simulation to assess its ability to predict overall health. A random sampling of individual attributes was taken to form 20 simulated observations, that data was then used to predict General Health. The simulated observations were then compared to the actual data and where possible (where an actual observation contained the same values as the simulated observation) an actual observed value of General Health was obtained; where there was no perfect fit among the real observations, those observations that closely fit the simulated data (within 1 unit for discrete variables, and within 1 SE for continuous variables) were used to develop an estimate of General Health. Those actual and estimated General Health values were then regressed against the predicted values from the simulated data. In Excel this produced the following in the first iteration:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.606433
R Square	0.367761
Adjusted R Square	0.332637
Standard Error	0.459658
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2.212211	2.212211	10.47027	0.004587
Residual	18	3.803132	0.211285		
Total	19	6.015343			

Predicted	*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2.724007	0.252701	10.77956	2.78E-09	2.193102	3.254912
Actual	0.206295	0.063754	3.235779	0.004587	0.072352	0.340237

Ten iterations of this simulation produced:

Iteration	Adjusted R Square
1	0.332637
2	0.118252
3	0.149233
4	0.208689
5	0.13057
6	0.266179
7	0.336294
8	0.046546
9	0.130811
10	0.287952
Mean	0.200716
SE	0.095273

This testing method resulted in a better than expected ability to predict overall health. Based on the quality of the underlying model an R^2 of ~17% is expected and an R^2 of ~20% is achieved. Increasing the number of iterations should result in the achieved R^2 tending toward 17%.

A small random sample of my coworkers (8 observations) resulted in an R^2 of 24.5% between predicted and actual General Health. Based on the limited sample size it appears as though this result is in line with the model developed from the ATUS data; as such no further analysis of this small sample (casually obtained) is warranted.

This model has some predictive power, but is not able to accurately predict General Health for individual observation with a narrow enough confidence interval to be useful. By loosening the requirements for the model to predict at the 95% confidence level and allowing predictions within ±1t (one standard error) of the actual (68% confidence) the model has enough power to be able to inform policy decisions as it can predict General Health with some accuracy on average at this level.

$$\bar{y} \pm 1 * 0.928$$
$$\bar{y} \pm 0.928$$

This predictive power (68% confidence) can be useful, as it still predicts that on average the outcome will be within 1 point of actual more often than not.

The intent of the model is not to predict individual outcomes, but rather to predict the effect of certain attributes on the average health of all observed. Further refinement through the addition of other predictor variables and through objectively collected data can help to improve the accuracy of this model at higher confidence levels.

Appendix D: State level analysis

Adult obesity, hypertension, diabetes, and physical activity were evaluated at the state level and compared with other states to attempt to find the region that would be best served by additional Federal resources.

First obesity was evaluated for its strength as a proxy for other health related characteristics. A simple Pearson correlation was used to see to what degree obesity is correlated to diabetes, hypertension, smoking, and high cholesterol. In all cases the tcalc exceeds tcrit for a two-tailed t-test at the 95% significance level. Each of these factors is strongly correlated with obesity, allowing obesity to stand in as a predictor of overall health. Data from 50 states and the District of Columbia (n=51)

Correlation: Adult Obesity, Diagnosed Diabetes, Diagnosed High Cholesterol, Diagnosed Hypertension, Adult Smoking

	Adult Obesity	tcalc	tcrit
Diabetes	0.749	10.565	2.311
	0.000		
High Cholesterol	0.670	9.429	2.311
	0.000		
Hypertension	0.760	10.771	2.311
	0.000		
Adult Smoking	0.759	10.751	2.311
	0.000		

Second a regression analysis was performed on obesity versus physical activity to ensure that the previously found relationship holds true for a state by state comparison.

Regression Analysis: Adult Obesity (2012) versus Adult Physical Activity (2011)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	0.034848	0.034848	77.14	0.000
Adult Physical Activity (2011)	1	0.034848	0.034848	77.14	0.000
Error	49	0.022135	0.000452		
Lack-of-Fit	45	0.021771	0.000484	5.33	0.056
Pure Error	4	0.000363	0.000091		
Total	50	0.056983			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0212539	61.16%	60.36%	57.51%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.5306	0.0287	18.46	0.000	
Adult Physical Activity	-0.4900	0.0558	-8.78	0.000	1.00

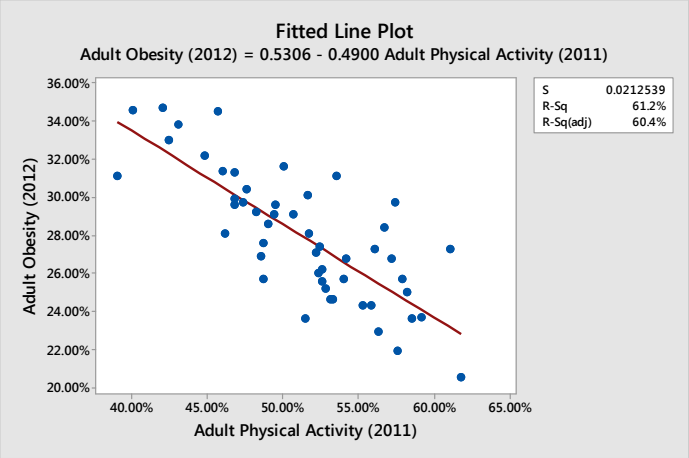
Regression Equation

Adult Obesity = 0.5306 - 0.4900 Adult Physical Activity

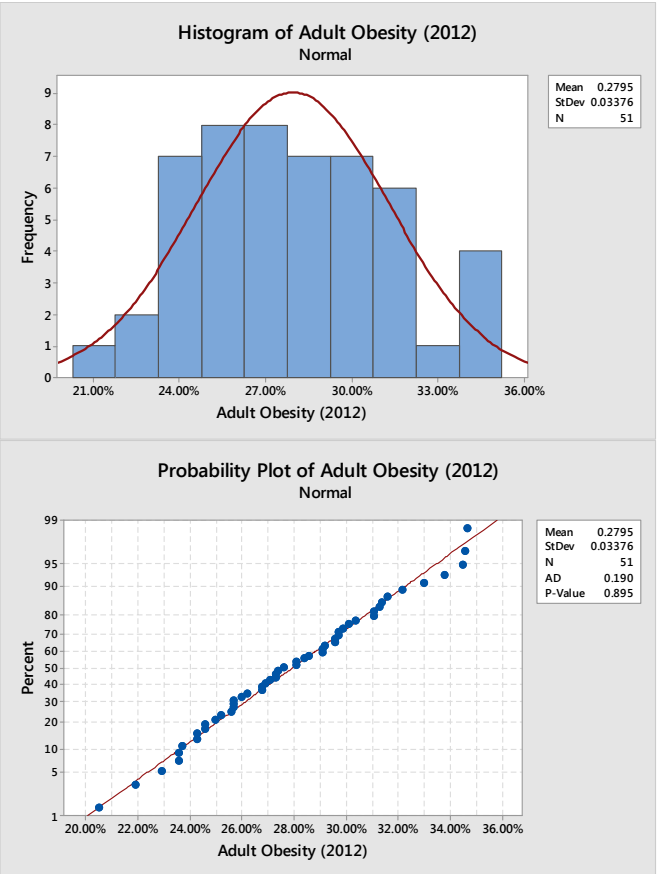
Tcrit (at the 95% level for a two tailed t test with 50 degrees of freedom)=-2.311,2.311

Tslope=-8.78 < Tcrit

Assuming a null hypothesis that states there is no correlation, reject the H_0 and conclude that Physical Activity is an effective predictor of Obesity.



Continuing with adult obesity acting as a proxy for overall health, the state by state rates were compared to attempt to find the states with the highest levels that would be best served through additional resources. Obesity rate was found to be normally distributed among the states.



Descriptive Statistics: Adult Obesity (2012)

Variable	Total Count	Mean	SE Mean	StDev
Adult Obesity	51	0.27945	0.00473	0.03376

Looking only at the high end (the right tail) of the population those states that represent to highest 20% would be our easiest targets for improvement. The method for measurement is important however, as the states with rates of obesity greater than 30.8% (top 80%) are not the same states with the highest total obese population. The states with the highest rates tend to be rural southern states with lower populations:

State	Total n	Adult Obesity (2012)	Obesity total
Louisiana	4,625,470	34.70%	1605038.09
Mississippi	2,991,207	34.60%	1034957.62
Arkansas	2,959,373	34.50%	1020983.69
West Virginia	1,854,304	33.80%	626754.75
Alabama	4,833,722	33.00%	1595128.26
Oklahoma	3,850,568	32.20%	1239882.90
South Carolina	4,774,839	31.60%	1508849.12
Indiana	6,570,902	31.40%	2063263.23
Kentucky	4,395,295	31.30%	1375727.34
Michigan	9,895,622	31.10%	3077538.44
Tennessee	6,495,978	31.10%	2020249.16

Whereas the states with the highest total obese populations, tend to have lower rates, but much larger overall populations resulting in the top 20% now being those states with obese populations greater than 3,272,908:

State	Total n	Adult Obesity (2012)	Obesity total
California	38,332,521	25.00%	9583130
Texas	26,448,193	29.20%	7722872
Florida	19,552,860	25.20%	4927321
New York	19,651,127	23.60%	4637666
Pennsylvania	12,773,801	29.10%	3717176
Illinois	12,882,135	28.10%	3619880
Ohio	11,570,808	30.10%	3482813

A smaller group of states (8 versus 11), however much larger in area and more spread out across the country. An approach that focuses on the rural southern states listed in the first table would likely result in better state wide numbers, whereas an approach that focused on the states in the second table would likely result in better national numbers. Either would be beneficial, however money spent in the poorer southern states (also those with the highest rates of obesity) may have the largest impact per dollar spent as it would amount to a larger percentage on the per capita income in those states than in the richer states from the second table.