Technical Report: County Health Calculator
(Sep. 2013)

Prepared by:
Robert E. Johnson, PhD[i]
Chunfeng Ren, MPH[ii]
Amber D. Haley, MPH[iii]
Steven H. Woolf, MD, MPH[iv]

Center on Human Needs

Virginia Commonwealth University

[i] Departments of Biostatistics and Family Medicine
[ii] Department of Biostatistics
[iii] Center on Human Needs
[iv] Center on Human Needs, Departments of Family Medicine and Epidemiology and Community Medicine

Contents

Introduction

The County Health Calculator (CHC) was developed in 2011-2012 by researchers at Virginia Commonwealth University's Center on Human Needs with support from the Robert Wood Johnson Foundation. The CHC is an online simulation tool to explore how health is associated with education, income, and other social determinants of health. By manipulating slider bars, the user can examine how mortality, diabetes prevalence, or diabetes associated medical costs—for the nation or a specific state or county—would be affected if more favorable socioeconomic conditions existed. Research and development for the tool included regression analyses of data from the U.S. Census Bureau and the U.S. Department of Health and Human Services. The simulation tool is intended for use by the public and policymakers.

The development team at Virginia Commonwealth University (VCU) in Richmond, VA was led by Steven H. Woolf, MD, MPH, director of the VCU Center on Human Needs, and by Amber Haley, MPH, Project Epidemiologist. Statistical research was performed at VCU by Robert E. Johnson, PhD and Chunfeng Ren, MPH. The project was conducted in collaboration with Michael J. O'Grady, a Senior Fellow in the Health Care Research department at the University of Chicago.

The County Health Calculator is driven by a database containing mortality, diabetes prevalence, Medicare expenditure, social-economic, and demographic data for the nation, 50 states plus the District of Columbia, and most counties. Numbers of deaths and death rates were aggregated over years 2008-2010, the most current available data. Measures of educational attainment, income, age, race, and ethnicity were obtained from the same period. Diabetes prevalence and Medicare expenditures were obtained for year 2009, the most current available data. Health measures based on the selected education or income value obtained from the database are displayed. These measures are derived from prediction (regression) equations developed by the Biostatistics team at VCU. In this document we provide details of the data sources and methods.

1. Data Sources

We obtained data from various publically available resources. These are summarized in Table 1. We obtained crude mortality rates (CMR), age-adjusted mortality rates (AMR), estimated diabetes prevalence (DP), Medicare expenditure, national diabetes per capital cost, educational attainment, median household income, and the distribution of household income relative to the federal poverty level. Additionally, we obtained age, gender, and race distributions for each county. The sources are detailed here.

Table 1. Data Sources

Variables	**Sources**	**Years**	**Notes**
Mortality	CDC Wonder: Detailed Mortality Data	2008-2010	Number of deaths, CMR, and AMR for persons less than 75 years old
Diabetes prevalence	CDC: Diabetes Data & Trends County Level Estimates of Diagnosed Diabetes	2009	Diabetes prevalence for persons 20 years and older
Medicare expenditure	Dartmouth Atlas of Health Care Medicare Reimbursement Measures	2009	Available for 3130 out of 3141 counties
National per capita diabetes cost	Diabetes Care Economic Cost of Diabetes	2007	National per capita diagnosed diabetes cost
Educational attainment	U.S. Census Bureau, American Community Survey: American FactFinder: Table B15002	2008-2010	Available for 1820 out of 3141 counties
	U.S. Census Bureau, Census 2000 Summary File 3: American FactFinder: Table P37	2000	Available for all 3141 counties and used for imputation of missing data
Income-to-poverty ratio	U.S. Census Bureau, American Community Survey: American FactFinder: Table C17002	2008-2010	Available for 1823 out of 3141 counties
	U.S. Census Bureau, Census 2000 Summary File 3: American FactFinder: Table P88	2000	Available for all 3141 counties and used for imputation of missing data
	U.S. Census Bureau, Small Area Income and Poverty Estimates, State and County Data Files	2008-2010	100% FPL: available for all 3141 counties and used for imputation of missing data
Median household income	U.S. Census Bureau, Small Area Income and Poverty Estimates, State and County Data Files	2008-2010	Available for all 3141 counties
Division of the United States	U.S. Census Bureau, Census Region and Divisions of the United States	Current since 1984	Nine groups of geographically clustered states
Gender	CDC Wonder: Detailed Mortality Data	2008-2010	Percent of population who are female
Age group distribution	CDC Wonder: Detailed Mortality Data	2008-2010	Percent in age groups: 0-9, 10-19, 20-34, 35-54, and 55+
Hispanic origin	CDC Wonder: Detailed Mortality Data	2008-2010	Percent Hispanic or Latino
Race	CDC Wonder: Detailed Mortality Data	2008-2010	Percent American Indian or Alaska Native, Asian or Pacific Island, Black or African American, and White.

1.1. Mortality

We obtained numbers of deaths, CMR, and AMR using the Center of Disease Control's tool, CDC Wonder.[1] The primary source within this tool was the *Detailed Mortality, 1999-2010 Request*. On the *request form* we grouped the results by state, county, selected age groups with age less than 75, and aggregated years 2008-2010. We requested AMR using the 2000 U.S. Standard Population.[2] The query returned data on 3141 counties. Note that for Virginia, 39 independent cities were considered county-equivalents for census purposes. Deaths and population estimates were reported separately for Baltimore City (Maryland), Carson City (Nevada), and St. Louis City (Missouri).

1.2. Diabetes prevalence

We obtained county-level DP and age-adjusted diabetes prevalence (ADP) from Diabetes Data and Trend, CDC.[3] The CDC produced these estimates by starting with data from the Behavioral Risk Factor Surveillance System (BRFSS) on the question "Have you ever been told by a doctor that you have diabetes?" They then applied a Bayesian estimation method to produce county-level data. We downloaded the most current available data (2009). ADP was generated by using indirect adjustment from the 2000 U.S. Standard Population.[2] To estimate the count of diabetes cases, we multiplied the DP by the year 2009 bridged-race population for persons age 20 years and older, available through CDC Wonder. This yielded data on 3141 counties.

1.3. Medicare Expenditures

Dartmouth Atlas of Health Care (DAHA) developed a new series of Medicare expenditure measures by hospital referral region. We obtained county-level estimates of per capita expenditures for Medicare enrollees in 2009.[4] These expenditures, generated from 20% samples of fee-for-service populations, are adjusted for age, gender, and race by way of indirect standardization using the national Medicare population as the standard. DAHA suppressed the costs for 9 counties with sample size less than 11. This yielded data on 3130 counties.

1.4. National per capital diabetes cost

In 2007, researchers in American Diabetes Association updated a previously developed diabetes economic model, which used a prevalence-based approach combining the demographics of the population in 2007 with DP, health care costs, and economic data. They concluded that "people with diagnosed diabetes incur average expenditures of $11,744 per year, of which $6,649 is attributed to diabetes."[5] We used $6,649 as the national per capita cost of diabetes.

1.5. Education

We obtained data on educational attainment for adults age 25 and older from the 2008-2010 American Community Survey 3-Year Estimates (select *All Counties*), US Census Bureau.[6] We used Table B15002 (*Sex by educational attainment for the population 25 years and over*). This yielded data on 1907 geographies, including 1842 U.S. counties and 65 municipalities in Puerto Rico. These data provided educational attainment for only 59% of the 3141 US counties. We imputed values for the remaining counties, details to follow below. We combined the reported categories into four groups:

The percentage of adults who attained

- *Less than High School Graduate:* 12th grade or less, no diploma
- *High School:* High school graduate (includes equivalency)
- *Some College:* Some college through Associate's degree (short of Bachelor's degree)
- *College:* Bachelor's degree or higher

For purposes of the slider bar, we collapsed the 3rd and 4th groups to obtain: *Percent of adults with some college education*.

1.6. Income

We measured income in two ways: median household income and percent of population with household income above 200% of the federal poverty level (FPL).

We obtained the median household income over the years 2008-2010 from *Small Area Income and Poverty Estimates* (SAIPE), U.S. Census Bureau.[7] To obtain the data, we went to the *SAIPE State and County Data Files* page, selected year 2008, and downloaded the worksheet *est07all.xls*. We repeated this process for years 2009 and 2010. We took the median of the three measures as our value for the 2008-2010 median household income. This yielded the median household income for 3140 counties. (The query returned one additional county, Kalawao County, Hawaii, with missing income information.)

We obtained data on income-to-poverty ratio (IPR) from the 2008-2010 American Community Survey 3-Year Estimates (select *All Counties*), US Census Bureau.[6] We used Table

C17002 (*Ratio of income to poverty level in the past 12 months*). This yielded data on 1907 geographies, including 1842 U.S. counties and 65 municipalities in Puerto Rico. These data provided IPR for only 59% of the 3141 US counties. We imputed values for the remaining counties, details to follow below. We combined the reported categories into four groups:

The percent of households in the past 12 months with incomes

- *Less than 50%*
- *50% – 100%:* at least 50% and less than 100%
- *100% – 200%:* at least 100% and less than 200%
- *200% +:* at least 200%

the federal poverty level.

For purposes of the slider bar, we focused on the last group, *200% +*: *Percent of adults with basic income*.

<u>**1.7.**</u> <u>Demographics</u>

We collected demographic information from CDC Wonder's *Detailed Mortality, 1999-2010 Request*.[1] Specifically, we obtained relative frequency distributions (percentages) on persons less than 75 by gender, age groups, race groups, and Hispanic origin for mortality data. We obtained relative frequency distributions (percentages) on persons aged 20 and over by gender, age groups, race groups, and Hispanic origin for diabetes model. We repeated the request for each demographic variable. This method assured that we used the same population source as was used in determining mortality rates or diabetes prevalence.

2. Education and Income Imputation

Educational attainment and IPR data were available on only 59% (1842/3141) of all counties represented in the American Community Survey (ACS) data. Since these are key measures for the tool, we chose to estimate (impute) values for the remaining counties using a linear regression equation. Our independent variables, available for all counties, were the population totals (2008-2010 ACS) and the educational attainment and IPR measures from year 2000 (2000 U.S. Census). Since several states have very few counties, *state* was not included as an independent variable in these imputations. The available 2008-2010 ACS measures were fit to the independent variables and the estimated regression equations were used to impute both educational attainment and income for counties without 2008-2010 ACS information. The details are provided here.

2.1. Counties with missing education and income information

Educational attainment and IPR were unavailable for counties with populations less than 20,000.[8] The 1,299 counties with missing information were distributed over 44 states, with North Dakota and South Dakota having the largest missing percentages, 85% and 86%, respectively.

2.2. Imputation models

The imputation model for educational attainment is given by

$$\log\left(\frac{\pi_{ik}}{\pi_{i1}}\right) = \beta_{0k} + \sum_{j=1}^{3} \beta_{jk} x_{ji} + \beta_{4k} x_{4i}; i = 1, 2, \cdots, 1842; k = 2, 3, 4 .$$

where $\pi_{ik} (k = 1, 2, 3, 4)$ are the probabilities of population with less than high school, high school, some college, and college education in 2008-2010 for the i^{th} county; $x_{ji} (j = 1, 2, 3)$ are the observed educational attainment percentages of population with less than high school, high school, some college for year 2000 in the i^{th} county; and x_{4i} is the population size in 2008-2010 for the i^{th} county. (Data source: see Table 1)

The imputation model for poverty is given by

$$\log\left(\frac{\pi_{ik}}{\pi_{i1}}\right) = \beta_{0k} + \sum_{j=1}^{3} \beta_{jk} x_{ji} + \beta_{4k} x_{4i} + \beta_{5k} x_{5i}, i = 1, 2, \cdots, 1842, k = 2, 3, 4.$$

where $\pi_{ik} (k = 1, 2, 3, 4)$ are the observed probabilities of population with household income less than 50%, 50%-100%, 100%-200%, and above 200% the FPL in 2008-2010 for the i^{th} county; $x_{ji} (j = 1, 2, 3)$ are the observed basic income percentages of population with household income less than 50%, 50%-100%, and 100%-200% for year 2000 in the i^{th} county; x_{4i} is the population size in 2008-2010 for the i^{th} county; and x_{5i} is percent of population with household income below the FPL in 2008-2010. (Data source: see Table 1)

2.3. Evaluation of imputation models

We fit a multinomial logistic regression model for the four categories of educational attainment and the four categories of IPR. The R^2 values (used to evaluate the fit of the model) were 0.88, 0.89, 0.81, and 0.94 for educational attainment, respectively, and 0.71, 0.83, 0.74, and 0.92 for IPR, respectively.

3. Building the Regression Models for CMR

One primary function of the health calculator tool is to illustrate the inverse association between health (mortality) and two social determinants (educational attainment and basic income). Counties with higher percentages of persons with some college education and basic income tend to have lower mortality rates. What if a hypothetical county with given mortality and percentage of persons with some college were to instead have a higher percentage of persons with some college? How much higher would income be? How much lower would the associated mortality be? How many fewer deaths would occur? What change would occur in that county's age-adjusted mortality rank among all counties in the state? With the health calculator tool, the user selects a county (state or nation) and views the mortality of a hypothetical county which takes on the same characteristics as the chosen county but with higher (or lower) educational attainment.

To model changes in mortality we developed a scaled logistic regression model which predicts the CMR as a function of education attainment and basic income (independent variables). We adjusted for the county's age, gender, Hispanic origin, and race distributions (covariables). We allowed coefficients of educational attainment and basic income to vary according to which division—out of 8 divisions nationally—the state of the county falls (interactions). Using this model, we predicted the CMR at the value of the observed educational attainment and basic income measures ($\widehat{CMR}(X_0)$) and at the measures chosen by the user of the tool via the slider ($\widehat{CMR}(X)$). The observed CMR ($CMR(X_0)$) is then adjusted to reflect the change in the measure. The adjusted CMR ($CMR(X)$) is related to the observed CMR by a factor equal to the ratio of the predicted CMR at the new measure to the predicted CMR at the observed measure:

$$CMR(X) = \frac{\widehat{CMR}(X)}{\widehat{CMR}(X_0)} CMR(X_0).$$

In the regression model for CMR, we omitted 70 counties with fewer than 20 deaths—aggregated over 2008-2010—since the death rates would be unreliable. The tool is available for 3072 counties, all states and D. C., and the nation.

3.1. Associations between mortality, education, and income

County CMR is correlated with both educational attainment (r = –0.63) and basic income (r = –0.62). These associations are illustrated in Figure 1A (CMR by educational attainment) and Figure 1B (CMR by basic income). The scatter plots were smoothed using a 30-point moving average after truncating 15 points on each side, left and right. The relationship was not linear over lower social-economic values. We would expect the mean CMR (approximated by the moving average) to plateau at a high CMR on the left and level out at a low CMR on the right.

Figure 1A: Crude Mortality Rate by Educational Attainment

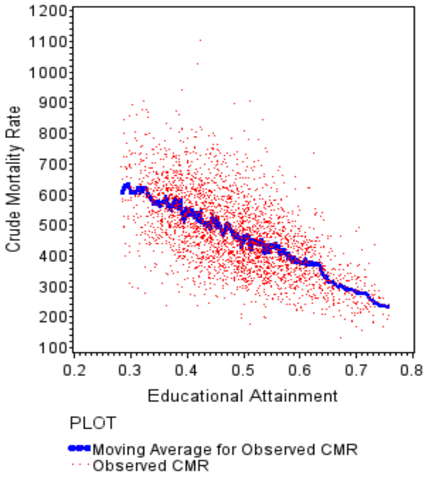
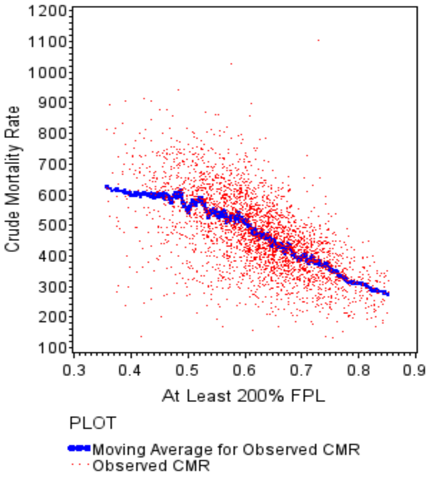


Figure 1B: Crude Mortality Rate by Basic Income



3.2. Scaled-logistic regression for CMR

CMR is proportional to the number of deaths divided by the population and may be modeled as the *probability of deaths*. A logistic regression model is suitable for such data. We also considered a Poisson model since the death counts are small relative to the population size; however, we found the logistic model to perform better for the data at hand.

The inverse canonical link, assuming binary error, is the Logistic cumulative distribution function. This allows the modeled probability to be an *S*-shaped function of the model's linear component. However, when the observed range of the estimated probability is near zero, the model may miss the high and low plateaus of the *S*-shape. We introduced an additional scale parameter to adjust the range of the inverse link. We thus performed a logistic regression with inverse link function, $g^{-1}(\eta)$ given by

$$g^{-1}(\eta) = \frac{\alpha}{1+\exp(-\eta)}$$

where $\alpha \in (0,1]$ and η is the linear component of the model. The link function is thus

$$g(\mu) = \ln\left(\frac{\mu}{\alpha-\mu}\right)$$

where μ represents the modeled probability (mean of the independent variable).

The maximum likelihood estimates of the linear component parameters and the additional scale parameter were found using SAS/STAT®, version 9.2, PROC GENMOD.

3.3. The model

We fitted a scaled logistic regression model for educational attainment and basic income to the county CMR values. We included county characteristics which may influence CMR and the social-economic measures, but would remain constant with changing the measures. These included age group distributions, race, gender, Hispanic origin, and division of the country into which the associated state falls. We omitted from the model other social-economic measures that would be influenced by changing educational attainment or basic income. Adjusting for these variables would effectively hold our measures constant and would change the nature of the association. While it is important to understand the residual effect of our measures controlling for modifiers and mediators, this is not the purpose of the tool.

3.3.1. Predicting mortality from educational attainment and basic income

The linear component of the model for a given county is

$$\eta = \beta_0 + \beta_1 X + \beta_2 Y + \sum_{j=3}^{11} \beta_j x_j + \sum_{k=1}^{8} \beta_{12k} x_{12k} + \sum_{k=1}^{8} \gamma_{1k} x_{12k} X + \sum_{k=1}^{8} \gamma_{2k} x_{12k} Y,$$

where Y is the education attainment measure;

X is the basic income measure;

x_j (j = 3,…,6) is the population relative frequency of the *jth* age group (0-9, 10-19, 20-34, and 35-54);

x_j (j = 7,8,9) is the population relative frequency of the *jth* race group (American Indian, Black, and Asian);

x_{10} is the population relative frequency of females;

x_{11} is the population relative frequency of Hispanic or Latino;

x_{12k} is a binary value which indicates the division of the country, k =1,…,8.

The parameter β_0 is an intercept term representing the mean of the reference group (overall mean + effect of division 9), $\beta_i (i = 1, 2)$ are the slopes of the educational attainment and basic income measures, $\beta_2 \dots \beta_{11}$ are slopes of the respective percentages, $\beta_{12,1} \cdots \beta_{12,8}$ are division effects which add to the intercept β_0 depending on the division, and $\gamma_{i1} \cdots \gamma_{i8} (i = 1, 2)$ are interaction terms which add to the slopes $\beta_i (i = 1, 2)$ depending on the division.

3.3.2. Updating mortality when changing education/income level

Given the value of one social-economic measure, we estimate the other one using a bisector regression (see Section 8 for details about the bisector regression). Using the fitted model, we estimate the linear component for the value of social-economic measure, *X*, as

$$\hat{\eta}(X) = \left(\hat{\beta}_0 + \sum_{j=3}^{11} \hat{\beta}_j x_j + \sum_{k=1}^{8} \hat{\beta}_{12k} x_{12k} \right) + \left(\hat{\beta}_1 + \sum_{k=1}^{8} \hat{\gamma}_{1k} x_{12k} \right) X + \left(\hat{\beta}_2 + \sum_{k=1}^{8} \hat{\gamma}_{2k} x_{12k} \right) Y ,$$

where $Y = \dfrac{1}{1 + \exp\left[-(\widehat{\phi}_0 + \widehat{\phi}_1 \log it(X) + \sum_{j=2}^{10} \widehat{\phi}_j x_j + \sum_{k=1}^{8} \widehat{\phi}_{11k} x_{11k}) \right]}$.

Similarly, using the fitted model, we also estimate the linear component for the value of social-economic measure *Y*. The predicted CMR (number of deaths per 100,000 population) is given by

$$\widehat{CMR}(X) = \frac{\hat{\alpha}}{1 + \exp\left(-\hat{\eta}(X) \right)} \times 100,000 ,$$

and the adjusted CMR is given by

$$CMR(X) = \frac{\widehat{CMR}(X)}{\widehat{CMR}(X_0)} CMR(X_0) = \frac{1 + \exp\left(-\hat{\eta}(X_0) \right)}{1 + \exp\left(-\hat{\eta}(X) \right)} CMR(X_0).$$

Using the adjusted CMR we compute the associated number of deaths and the change in the death rate, which allows us to compute the number and percentage of averted deaths. Letting P_c be the county population size, these values are as follows:

$$Deaths = CMR\left(X\right) \times \frac{P_c}{100,000}$$

$$CMR_{change}\left(X\right) = CMR\left(X_0\right) - CMR\left(X\right)$$

$$AvertedDeaths = CMR_{change}\left(X\right) \times \frac{P_c}{100,000}$$

$$PctAvDeaths = \frac{AvertedDeaths}{Deaths} \times 100 \quad .$$

3.4. Evaluation of the model

The correlation between observed crude mortality proportion and the predicated crude mortality proportion was 0.85. Efron's pseudo R-square[9] was 0.73. These values indicate a good model fit. Figure 2A shows the observed crude mortality proportion against the predicted crude mortality proportion. Figure 2B shows the standardized Pearson residuals versus the fitted values. These two figures also indicate the models fit well.

Figure 2B suggests the presence of potential outliers; however, removing them from the regression does not significantly improve the model fit. These values were retained in the model.

Figure 2A: Predicted versus Observed Crude Mortality Proportion

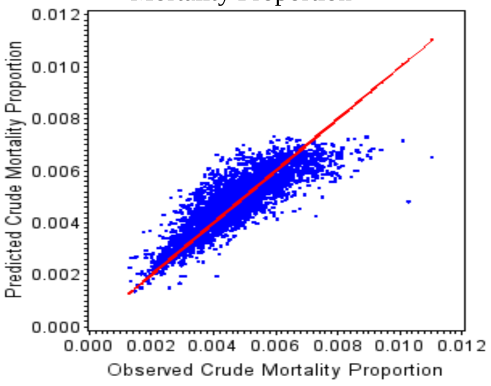
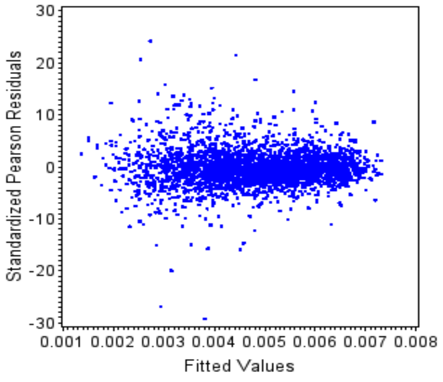


Figure 2B: Pearson Residuals versus Fitted Values



3.5. County mortality ranking

We used AMR when ranking counties within a state or ranking states across all states. Our scaled logistic regression equations predicted CMR. However, CMR and AMR are highly correlated (r = 0.99, R^2 = 0.99) when adjusting for age distribution. Therefore, we estimated AMR by way of a linear regression equation, given as

$$\widehat{AMR}\left(X\right)=\hat{\mu}+\hat{\lambda}_0\,\widehat{CMR}\left(X\right)+\sum_{i=1}^{4}\hat{\lambda}_i A_i$$

where A_i is the proportion of the county's population in the i^{th} age group (0-9, 10-19, 20- 34, 35-54). The proportion for the age group 55-74 is a function of the other four proportions so it was not included in the model. The adjusted AMR was calculated as

$$AMR\left(X\right)=\frac{\widehat{AMR}\left(X\right)}{\widehat{AMR}\left(X_0\right)}AMR\left(X_0\right)$$

where $AMR\left(X_0\right)$ was the observed AMR corresponding to the observed social-economic measure X_0.

4. State and National Level Predictions for CMR

The CHC tool was designed primarily to predict county-level values, but the user can choose any state, including the District of Columbia (DC), or the nation. We discuss here how we extended these county predictions to state and national levels.

4.1. Data sources

We obtained current values for states and the nation using the data sources as described in Table 1. The state and national level educational attainment and IPR measures supplied by ACS were based on counties with available information. Recall that fully 41% of all counties were not included in these data. We chose to use the ACS estimates to represent current values for the states and nation. However, as described in Section 2, we imputed values for these counties for purposes of this tool. As such, the average social-economic measures represented in this tool will be slightly different than those reported by ACS.

4.2. Predicting mortality at the state level

Let X and X_0 be the user-defined and observed educational attainment/basic income measures at the state level. We defined a change-factor based on these measures as follows:

$$\delta(X) = \begin{cases} \dfrac{X - X_0}{1 - X_0} & \text{for } X \geq X_0 \\ \dfrac{X}{X_0} & \text{for } X < X_0 \end{cases}$$

This change-factor was then applied to each county within the state to arrive at the adjusted educational attainment/basic income county-level measure (denoted X_c):

$$X_c = \begin{cases} X_{0c} + (1 - X_{0c})\,\delta(X) & \text{for } X \geq X_0 \\ X_{0c}\,\delta(X) & \text{for } X < X_0 \end{cases}$$

where X_{0c} is the observed county-level measure.

4.2.1. State-level adjusted CMR

To get state-level predicted CMR we used the following derivation:

$$\widehat{CMR}_s(X) = \frac{\sum_{c=1}^{n} P_c \widehat{CMR}_c(X)}{P_s},$$

where $\widehat{CMR}_s(X)$ is the predicted state-level CMR, $\widehat{CMR}_c(X)$ is the predicted county-level CMR as described in Section 3.3.2, P_c is the county-level population size, and $P_s = \sum_{c=1}^{n} P_c$ is the state-level population size (sum of the n county-level sizes).

The state-level adjusted CMR (adjusted for the user supplied social-economic measure) was derived as

$$CMR_s(X) = \frac{\widehat{CMR}_s(X)}{\widehat{CMR}_s(X_0)} CMR_s(X_0),$$

where $CMR_s(X_0)$ is the observed state-level CMR. Using the adjusted CMR we can compute the associated number of deaths and the change in the death rate which allows us to compute the number and percentage of averted deaths. These values are as follows:

$$Deaths = CMR_s\left(X\right) \times \frac{P_s}{100,000}$$
$$CMR_{s,change}\left(X\right) = CMR_s\left(X_0\right) - CMR_s\left(X\right)$$
$$AvertedDeaths = CMR_{s,change}\left(X\right) \times \frac{P_s}{100,000}.$$
$$PctAvDeaths = \frac{AvertedDeaths}{Deaths} \times 100$$

4.2.2. <u>*State mortality ranking*</u>

As described in Section 3.5, we used the state-level AMR when ranking states on their mortality rates. We obtained the state-level adjusted AMR in a manner similar to the adjusted CMR. First we estimated county-level AMRs using their adjusted CMRs and the prediction equation given in Section 3.5. We next computed the adjusted county-level AMR as follows

$$AMR_c\left(X\right) = \frac{\widehat{AMR}_c\left(X\right)}{\widehat{AMR}_c\left(X_0\right)} AMR_c\left(X_0\right).$$

To get the state-level predicted AMR we used the weighted average of the county-level AMRs:

$$\widehat{AMR}_s\left(X\right) = \frac{\sum_{c=1}^{n} P_c \widehat{AMR}_c\left(X\right)}{P_s}.$$

The state-level adjusted AMR (adjusted for the user supplied social-economic measure) was derived as

$$AMR_s\left(X\right) = \frac{\widehat{AMR}_s\left(X\right)}{\widehat{AMR}_s\left(X_0\right)} AMR_s\left(X_0\right).$$

4.3. <u>Predicting mortality at the national level</u>

As in Section 4.2, we defined the change-factor $\delta\left(X\right)$ based on the user-defined and observed educational attainment/basic income measures at the national level. We then used the formulas in Section 4.2.1 to obtain adjusted CMRs for each state plus DC. These were then combined to arrive at the national-level predicted CMR as follows:

$$\widehat{CMR}_N\left(X\right) = \frac{\sum_{s=1}^{51} P_s \widehat{CMR}_s\left(X\right)}{P_N},$$

where $\widehat{CMR}_N\left(X\right)$ is the predicted national-level CMR, $\widehat{CMR}_s\left(X\right)$ is the predicted state-level CMR, P_s is the state-level population size, and $P_N = \sum_{s=1}^{51} P_s$ is the national-level population size (sum of the 51 state-level sizes, including DC).

The national-level adjusted CMR (adjusted for the user supplied social-economic measure) was derived as

$$CMR_N\left(X\right) = \frac{\widehat{CMR}_N\left(X\right)}{\widehat{CMR}_N\left(X_0\right)} CMR_N\left(X_0\right),$$

where $CMR_N\left(X_0\right)$ is the observed national-level CMR. Using the adjusted CMR we can compute the associated number of deaths and the change in the death rate which allows us to compute the number and percentage of averted deaths. These values are as follows:

$$Deaths = CMR_N\left(X\right) \times \frac{P_N}{100{,}000}$$

$$CMR_{N,change}\left(X\right) = CMR_N\left(X_0\right) - CMR_s\left(X\right)$$

$$AvertedDeaths = CMR_{N,change}\left(X\right) \times \frac{P_N}{100{,}000}$$

$$PctAvDeaths = \frac{AvertedDeaths}{Deaths} \times 100.$$

5. Building a Regression Model for Diabetes Prevalence

The other primary function of the health calculator tool is to illustrate the inverse association between DP and the two social determinants. Counties with higher percentages of persons with some college education and basic income tend to have lower DP. What if a hypothetical county with given DP and percentage of persons with some college were to instead have a higher percentage of persons with some college? How much higher would income be? How much lower would the associated DP be? How many fewer cases of diabetes would occur? What change would occur in that county's ADP rank among all counties in the state? With the health calculator tool, the user selects a county (state or nation) and views the DP of a hypothetical county which takes on the same characteristics as the chosen county but with higher (or lower) educational attainment..

To model changes in DP, we developed a scaled logistic regression model which predicts the DP as a function of education attainment and basic income (independent variables). We adjusted for the county's age, gender, Hispanic origin, and race distributions (covariables). We allowed coefficients of educational attainment and basic income to vary according to which division—out of 8 divisions nationally—the state of the county falls (interactions). Using this model we predicted the DP at the value of the observed educational attainment and basic income measures ($\widehat{DP}\left(X_0\right)$) and at the measures chosen by the user of the tool via the slider ($\widehat{DP}\left(X\right)$). The observed DP ($DP\left(X_0\right)$) is then adjusted to reflect the change in the measure. The adjusted DP ($DP\left(X\right)$) is related to the observed DP by a factor equal to the ratio of the predicted DP at the new measure to the predicted DP at the observed measure:

$$DP\left(X\right)=\frac{\widehat{DP}\left(X\right)}{\widehat{DP}\left(X_0\right)}DP\left(X_0\right).$$

In the regression model for DP, we omitted 3 counties with missing 200% FPL income. The model includes 3138 counties. The tool is available for 3072 counties, all states and DC, and the nation since we are based on the counties included in the CMR model.

5.1. Associations between DP, education, and income

County DP is negatively correlated with both educational attainment (r = –0.63) and basic income (r = –0.52). These associations are illustrated in Figure 3A (DP by educational attainment) and Figure 3B (DP by basic income). The scatter plots were smoothed using a 30-point moving average after truncating 15 points on each side, left and right. The relationship was not linear over lower social-economic values. We would expect the mean DP (approximated by the moving average) to plateau at a high DP on the left and level out at a low DP on the right. This may be accounted when changing one social determinant and admitting the other changing.

Figure 3A: Diabetes Prevalence by Educational Attainment

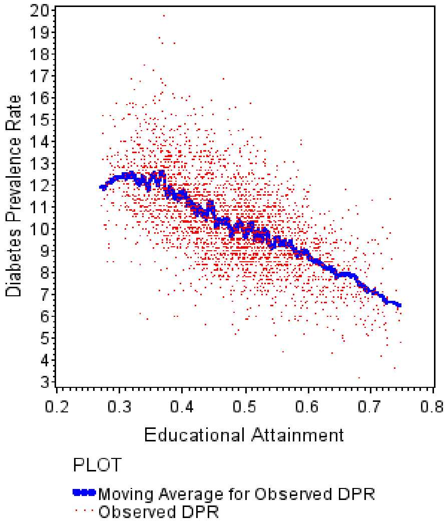
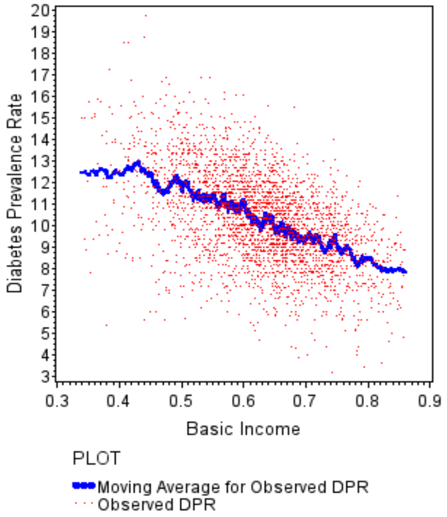


Figure 3B: Diabetes Prevalence by Basic Income



5.2. Scaled-logistic regression for DP

DP is proportional to the number of diabetes cases divided by the population and may be modeled as the *probability of diabetes*. A logistic regression model is suitable for such data.

The inverse canonical link, assuming binary error, is the Logistic cumulative distribution function. This allows the modeled probability to be an *S*-shaped function of the model's linear component. However, when the observed range of the estimated probability is near zero, the model may miss the high and low plateaus of the *S*-shape. We introduced an additional scale parameter to adjust the range of the inverse link. We thus performed a logistic regression with inverse link function, $g^{-1}(\eta)$, given by

$$g^{-1}(\eta)=\frac{\alpha}{1+\exp(-\eta)}$$

where $\alpha \in (0,1]$ and η is the linear component of the model. The link function is thus

$$g(\mu)=\ln\left(\frac{\mu}{\alpha-\mu}\right)$$

where μ represents the modeled probability (mean of the independent variable).

The maximum likelihood estimates of the linear component parameters and the additional scale parameter are found using SAS/STAT®, version 9.2, PROC GENMOD.

5.3. The model

We fitted a scaled logistic regression model for educational attainment and basic income to the county DP values. We included county characteristics which may influence DP and the social-economic measures, but would remain constant with changing the measures. These included age group distributions, race, gender, Hispanic origin, and division of the country into which the associated state falls. We omitted from the model other social-economic measures that would be influenced by changing educational attainment or basic income. Adjusting for these variables would effectively hold our measures constant and would change the nature of the association. While it is important to understand the residual effect of our measures controlling for modifiers and mediators, this is not the purpose of the tool.

5.3.1. Predicting DP from educational attainment and basic income

The linear component of the model for a given county is

$$\eta = \beta_0 + \beta_1 X + \beta_2 Y + \sum_{j=3}^{11} \beta_j x_j + \sum_{k=1}^{8} \beta_{12k} x_{12k} + \sum_{k=1}^{8} \gamma_{1k} x_{12k} X + \sum_{k=1}^{8} \gamma_{2k} x_{12k} Y,$$

where Y is the education attainment measure;

X is the basic income measure;

x_j (j = 3,…,6) is the population relative frequency of the j^{th} age group (20-34, 35-44, 45-54, and 55-64);

x_j (j = 7,8,9) is the population relative frequency of the j^{th} race group (American Indian, Black, and Asian);

x_{10} is the population relative frequency of females;

x_{11} is the population relative frequency of Hispanic or Latino;

x_{12k} is a binary value which indicates the division of the country, k =1,…,8.

The parameter β_0 is an intercept term representing the mean of the reference group (overall mean + effect of division 9), $\beta_i (i=1,2)$ are the slopes of the educational attainment and basic income measures, $\beta_2 \ldots \beta_{11}$ are slopes of the respective percentages, $\beta_{12,1} \cdots \beta_{12,8}$ are division effects which add to the intercept β_0 depending on the division, and $\gamma_{i1} \ldots \gamma_{i8} (i=1,2)$ are interaction terms which add to the slopes $\beta_i (i=1,2)$ depending on the division.

5.3.2. Updating DP when changing education/income level

Given the value of one social-economic measure, we estimate the other one using a bisector regression (see Section 8 for details about the bisector regression). Using the fitted model, we estimate the linear component for the value of social-economic measure, *X*, as

$$\hat{\eta}\left(X\right)=\left(\hat{\beta}_0+\sum_{j=3}^{11}\hat{\beta}_j x_j+\sum_{k=1}^{8}\hat{\beta}_{12k}x_{12k}\right)+\left(\hat{\beta}_1+\sum_{k=1}^{8}\hat{\gamma}_{1k}x_{12k}\right)X+\left(\hat{\beta}_2+\sum_{k=1}^{8}\hat{\gamma}_{2k}x_{12k}\right)Y,$$

where $Y=\dfrac{1}{1+\exp\left[-(\widehat{\phi}_0+\widehat{\phi}_1\log it(X)+\sum_{j=2}^{10}\widehat{\phi}_j x_j+\sum_{k=1}^{8}\widehat{\phi}_{11k}x_{11k})\right]}$.

Similarly, using the fitted model we also estimate the linear component for the value of social-economic measure *Y*. The predicted DP (diabetes cases per 100 populations) is given by

$$\widehat{DP}\left(X\right)=\frac{\hat{\alpha}}{1+\exp\left(-\hat{\eta}\left(X\right)\right)}\times 100,$$

and the adjusted DP is given by

$$DP\left(X\right)=\frac{\widehat{DP}\left(X\right)}{\widehat{DP}\left(X_0\right)}DP\left(X_0\right)=\frac{1+\exp\left(-\hat{\eta}\left(X_0\right)\right)}{1+\exp\left(-\hat{\eta}\left(X\right)\right)}DP\left(X_0\right).$$

Using the adjusted DP, we compute the associated diabetes prevalence and the change in DP which allows us to compute the number and percentage of averted diabetes prevalence. Letting P_c be the county population size, these values are as follows:

$$Cases=DP\left(X\right)\times\frac{P_c}{100}$$

$$DP_{change}\left(X\right)=DP\left(X_0\right)-DP\left(X\right)$$

$$AvertedCases=DP_{change}\left(X\right)\times\frac{P_c}{100}$$

$$PctAvCases=\frac{AvertedCases}{Cases}\times 100.$$

5.4. Evaluation of the model

The correlation between observed DP and the predicated DP was 0.86. Efron's pseudo R-square[9] was 0.73. These values indicate a good model fit. Figure 4A shows the observed DP against the predicted DP. Figure 4B shows the standardized Pearson residuals versus the fitted values. These two figures also indicate the model fits well.

Figure 4B suggests the presence of potential outliers; however, removing them from the regression does not significantly improve the model fit. These values were retained in the model.

Figure 4A: Predicted versus Observed Diabetes Prevalence

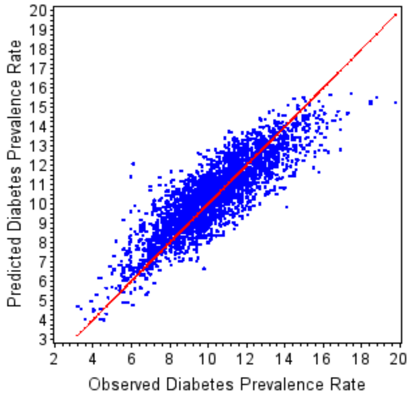
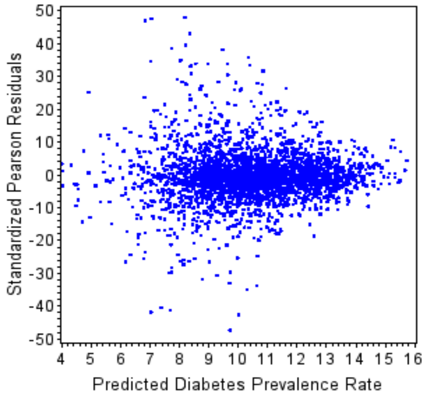


Figure 4B: Standardized Pearson Residuals versus Fitted Values



5.5. County diabetes prevalence ranking

We used ADP when ranking counties within a state or ranking states across all states. Our scaled logistic regression equations predicted DP. However, DP and ADP are highly correlated (*r* = 0.99, R^2 = 0.99) when adjusting for age distribution. Therefore, we estimated ADP by way of a linear regression equation, given as

$$\widehat{ADP}(X)=\hat{\mu}+\hat{\lambda}_0\widehat{DP}(X)+\sum_{i=1}^{4}\hat{\lambda}_i A_i$$

where A_i is the proportion of the county's population in the i^{th} age group (20- 34, 35-44,45-54, 55-64). The proportion for the age group 64+ is a function of the other four proportions so it was not included in the model. The adjusted ADP was calculated as

$$ADP(X)=\frac{\widehat{ADP}(X)}{\widehat{ADP}(X_0)}ADP(X_0),$$

where $ADP(X_0)$ was the observed ADP corresponding to the observed social-economic measures X_0 .

6. State and National Level Predictions for DP

The CHC tool was designed primarily to predict county-level values, but the user can choose any state, including the District of Columbia (DC), or the nation. We discuss here how we extended these county predictions to state and national levels.

6.1. Data sources

We obtained current values for states and the nation using the data sources as described in Table 1. The state and national level educational attainment and IPR measures supplied by ACS were based on counties with available information. Recall that fully 42% of all counties were not included in these data. We chose to use the ACS estimates to represent current values for the states and nation. However, as described in Section 2, we imputed values for these counties for purposes of this tool. As such, the average social-economic measures represented in this tool will be slightly different than those reported by ACS.

6.2. Predicting DP at the state level

Let X and X_0 be the user-defined and observed educational attainment or basic income measures at the state level. We defined a change-factor based on these measures as follows:

$$\delta(X)=\begin{cases}\dfrac{X-X_0}{1-X_0} & \text{for } X \geq X_0\\ \dfrac{X}{X_0} & \text{for } X < X_0\end{cases}$$

This change-factor was then applied to each county within the state to arrive at the adjusted educational attainment/basic income county-level measure (denoted X_c):

$$X_c=\begin{cases}X_{0c}+(1-X_{0c})\delta(X) & \text{for } X \geq X_0\\ X_{0c}\,\delta(X) & \text{for } X < X_0\end{cases}$$

where X_{0c} is the observed county-level measure.

6.2.1. State-level adjusted DP

To get state-level predicted DP we used the following derivation:

$$\widehat{DP}_s(X)=\frac{\sum_{c=1}^{n} P_c\,\widehat{DP}_c(X)}{P_s},$$

where $\widehat{DP}_s(X)$ is the predicted state-level DP, $\widehat{DP}_c(X)$ is the predicted county-level DP as described in Section 5.3.2, P_c is the county-level population size, and $P_s = \sum_{c=1}^{n} P_c$ is the state-level population size (sum of the *n* county-level sizes).

The state-level adjusted DP (adjusted for the user supplied social-economic measure) was derived as

$$DP_s(X) = \frac{\widehat{DP}_s(X)}{\widehat{DP}_s(X_0)} DP_s(X_0),$$

where $DP_s(X_0)$ is the observed state-level DP. Using the adjusted DP, we can compute the associated number of deaths and the change in the death rate, which allows us to compute the diabetes prevalence and percentage of averted cases. These values are as follows:

$$Cases = DP_s(X) \times \frac{P_s}{100}$$
$$DP_{s,change}(X) = DP_s(X_0) - DP_s(X)$$
$$AvertedCases = DP_{s,change}(X) \times \frac{P_s}{100}$$
$$PctAvCases = \frac{AvertedCases}{Cases} \times 100.$$

6.2.2. State DP ranking

As described in Section 3.5, we used the state-level ADP when ranking states on their DP. We obtained the state-level ADP in a manner similar to the adjusted DP. First, we estimated county-level ADPs using their adjusted DPs and the prediction equation given in Section 3.5. We next computed the adjusted county-level ADP as follows

$$ADP_c(X) = \frac{\widehat{ADP}_c(X)}{\widehat{ADP}_c(X_0)} ADP_c(X_0).$$

To get the state-level predicted ADP, we used the weighted average of the county-level ADPs:

$$\widehat{ADP}_s(X) = \frac{\sum_{c=1}^{n} P_c \widehat{ADP}_c(X)}{P_s}.$$

The state-level adjusted ADP (adjusted for the user supplied social-economic measure) was derived as

$$ADP_s\left(X\right)=\frac{\widehat{ADP}_s\left(X\right)}{\widehat{ADP}_s\left(X_0\right)}ADP_s\left(X_0\right).$$

6.3. Predicting DP at the national level

As in Section 4.2, we defined the change-factor $\delta\left(X\right)$ based on the user-defined and observed educational attainment/basic income measures at the national level. We then used the formulas in Section 4.2.1 to obtain adjusted DPs for each state plus DC. These were then combined to arrive at the national-level predicted DP as follows:

$$\widehat{DP}_N\left(X\right)=\frac{\sum_{s=1}^{51}P_s\widehat{DP}_s\left(X\right)}{P_N},$$

where $\widehat{DP}_N\left(X\right)$ is the predicted national-level DP, $\widehat{DP}_s\left(X\right)$ is the predicted state-level DP, P_S is the state-level population size, and $P_N=\sum_{s=1}^{51}P_s$ is the national-level population size (sum of the 51 state-level sizes, including DC).

The national-level adjusted DP (adjusted for the user supplied social-economic measure) was derived as

$$DP_N\left(X\right)=\frac{\widehat{DP}_N\left(X\right)}{\widehat{DP}_N\left(X_0\right)}DP_N\left(X_0\right),$$

where $DP_N\left(X_0\right)$ is the observed national-level DP. Using the adjusted DP we can compute the associated number of deaths and the change in the death rate which allows us to compute the number and percentage of averted deaths. These values are as follows:

$$
\begin{aligned}
Cases &= DP_N\left(X\right)\times\frac{P_N}{100}\\
DP_{N,change}\left(X\right) &= DP_N\left(X_0\right)-DP_s\left(X\right)\\
AvertedCases &= DP_{N,change}\left(X\right)\times\frac{P_N}{100}\\
PctAvCases &= \frac{AvertedCases}{Cases}\times 100.
\end{aligned}
$$

7. Diabetes Costs

We modeled the DP at county level. When changing educational attainment or income, we estimated the number of persons (adults aged 20 and older) who have diabetes at county, state, and national level. Additionally, we provided an estimate of the medical costs attributed to diabetes incurred for these diabetics. We calculated diabetes cost based on three data sources as mentioned in sections 1.2, 1.3, and 1.4. We created a cost ratio for each county defined as the county per capita Medicare expenditures divided by the national per capita Medicare expenditures. This was then multiplied by the national per capita diabetes costs to arrive at an estimate of the county per capita diabetes costs. The total cost for all diabetics in the county was estimated by multiplying the county per capita costs times the unadjusted estimated prevalence.

8. Symmetric regression between education and income

Note that educational attainment and income are independent variables with respect to the regression models, but are correlated with each other. When the percentage of educational attainment increases, the percentage of income tends to increase, and vice versa. We designed our tool so that as one social variable is altered via the slider, the other social variable changes accordingly. Therefore, we need to model the change in one social factor as a function of the other. We could use two regression equations, one regressing education on income, and the other regressing income on education, but this yields an asymmetric relationship.

It is well-known that regression model doesn't incur avertable relationship. To get an interchangeable relationship between educational attainment and basic income, we need to fit a symmetric regression. We considered (1) major axis regression (Pearson's orthogonal regression), (2) reduced major axis regression (impartial regression), and (3) bisector regression (double regression). Pearson's orthogonal regression is sensitive to the scale of the variables. The reduced major axis regression is scale-independent. By simulating runs, Isobe et al. (1990)

investigated the behavior of two asymmetric regression solutions and the solutions for the three symmetric regressions.[10] They recommended bisector regression because of the smaller standard deviation. We adopted the bisector regression in our modeling.

8.1. Association between educational attainment and basic income

Educational attainment and household income are positively associated. Measuring educational attainment as the proportion of persons with some college and basic income as the proportion of persons with household income at least twice the FPL, the correlation in the logit scale is 0.64. Logit transformation assures the estimate of one social determinant falls within the required range [0, 1] when the other changes. The associations are illustrated in Figures 5A (education against income) and 5B (income against education). The scatter plots were smoothed using a 30-point moving average after truncating 15 points on each side, left and right. Controlling the characteristics of counties, the relationship was linear in either case.

Figure 5A: Basic Income by Educational Attainment

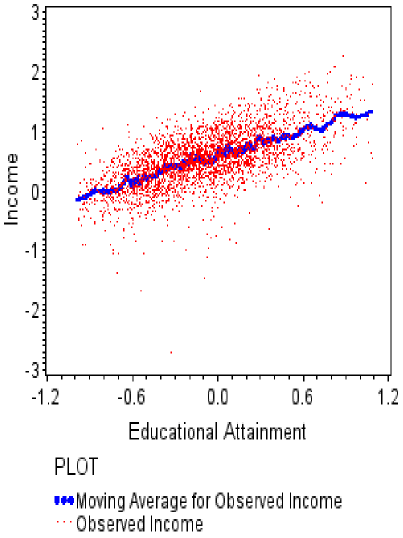
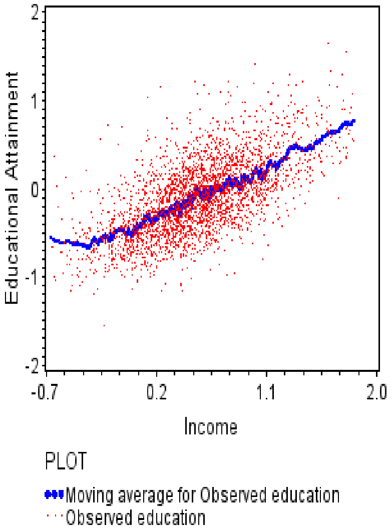


Figure 5B: Educational Attainment by Basic Income



8.2. Bisector regression model

We fitted a bisector regression model for educational attainment and income in logit. The bisector regression proceeds in three steps. First, regress y against x and x against y. Second, solve the ‘x against y’ regression for y. This provides two estimates (lines) of y as linear functions of x. Third, determine the line that bisects these two lines. Alexander V.E. and Christof

S. (1998) discussed how to fit a bisector regression model for a bivariate regression.[11] We extended it to a multiple regression.

Two asymmetric predicted models are given by

$$\widehat{\text{logit}(Y)} = \hat{\beta}_0 + \hat{\beta}_1 \text{logit}(X) + \sum_{j=2}^{10} \hat{\beta}_j x_j + \sum_{k=1}^{8} \hat{\beta}_{11k} x_{11k}$$

$$\widehat{\text{logit}(X)} = \hat{\alpha}_0 + \hat{\alpha}_1 \text{logit}(Y) + \sum_{j=2}^{10} \hat{\alpha}_j x_j + \sum_{k=1}^{8} \hat{\alpha}_{11k} x_{11k},$$

where Y/X is the education attainment/basic income measure;

x_j (j = 2,…,5) is the population relative frequency of the j^{th} age group (20-34, 35-44, 45-54, and 55-64);

x_j (j =6,7,8) is the population relative frequency of the j^{th} race group (American Indian, Black, and Asian);

x_9 is the population relative frequency of females;

x_{10} is Hispanic or Latino origin;

x_{11k} is a binary value which indicates the division of the country, k =1,…,8.

By the invariant property of maximum likelihood estimators and the definition of bisector regression, two bisector predicted regression equations are given by

$$\text{logit}(y) = \frac{1}{1-c}\left[\hat{\delta}_0 + \frac{c\hat{\theta}_0}{\hat{\theta}_1} + \left(\hat{\delta}_1 - \frac{c}{\hat{\theta}_1}\right)\text{logit}(x) + \sum_{j=2}^{10}\left(\hat{\delta}_j + \frac{c}{\hat{\theta}_1}\hat{\theta}_j\right)x_j + \sum_{k=1}^{8}\left(\hat{\delta}_{11k} + \frac{c}{\hat{\theta}_1}\hat{\theta}_{11k}\right)x_{11k}\right]$$

$$\text{logit}(y) = \frac{1}{1+c}\left[\hat{\delta}_0 - \frac{c\hat{\theta}_0}{\hat{\theta}_1} + \left(\hat{\delta}_1 + \frac{c}{\hat{\theta}_1}\right)\text{logit}(x) + \sum_{j=2}^{10}\left(\hat{\delta}_j - \frac{c}{\hat{\theta}_1}\hat{\theta}_j\right)x_j + \sum_{k=1}^{8}\left(\hat{\delta}_{11k} - \frac{c}{\hat{\theta}_1}\hat{\theta}_{11k}\right)x_{11k}\right]$$

$$\text{where } c = \frac{\sqrt{1 + \sum_{j=1}^{10}\hat{\delta}_j^2 + \sum_{k=1}^{8}\hat{\delta}_{11k}^2}}{\sqrt{1 + \frac{1}{\hat{\theta}_1^2}\left(1 + \sum_{j=2}^{10}\hat{\theta}_j^2 + \sum_{k=1}^{8}\hat{\theta}_{11k}^2\right)}}\ .$$

We chose the bisector regression model coefficients to be halfway between the coefficients of the two asymmetric regressions. Assume the chosen bisector regression is given by

$$\text{logit}(y) = \hat{\phi}_0 + \hat{\phi}_1 \text{logit}(x) + \sum_{j=2}^{10} \hat{\phi}_j x_j + \sum_{k=1}^{8} \hat{\phi}_{11k} x_{11k}\ .$$

We want this regression to be such that the predicted logit(y) is the same as the observed logit(y) when the social determinants are equal to their observed values. To do this, we adjusted the intercept $\widehat{\phi}_0$ to force the bisector regression to pass through the point (x_0, y_0), the observed values for the social determinants.

8.3. Evaluation of two asymmetric regression models

We fitted two asymmetric regression models when controlling the characteristics of counties. The R^2 values (used to evaluate the fit of the model) were 0.72 and 0.84, respectively, indicating good model fit.

9. County Health Calculator Database

The CHC MySQL database resides on a VCU Linux server and is accessed from the CHC application as needed. No sensitive person-level or person-identifiable data are stored in our database. The database consists of six tables (listed in Table 2) and twelve queries (listed in Table 3) which extract information from the tables for use in the CHC tool. The contents of the tables were populated from a series of SAS® programs by way of an open-database-connectivity (ODBC) link to the MySQL database.

Table 2: CHC Database Tables

Table Name	Notes
tblGeoInfo	Current information for all geographies
tblSlider_Educ	Counter box values for varying education levels for all geographies
tblSlider_Income	Counter box values for varying income levels for all geographies
tblCounty_MaxMin	Max/Min Educ/Income among counties by State
tblState_MaxMin	Max/Min Educ/Income among states
tblSlider_Value_Limits	Upper and lower limits for the slider by GeoType
	These values are used in the queries qrySlider_Selected_Educ_Values and qrySlider_Selected_Income_Values

Table 3: CHC Database Queries

Query Name	Notes
qrySelect_States_within_USA	Lists all states in the US
qrySelect_Counties_within_State	List all counties for the chosen state
qrySelect_USA_Current_Values	Contains current values for US
qrySelect_State_Current_Values	Contains current values for chosen state
qrySelect_County_Current_Values	Contains current values for chosen county

qrySlider_Selected_Educ_Values	Contains calculated values based on the slider selected education level
qrySlider_Selected_Income_Values	Contains calculated values based on the slider selected income level
qryState_Min-Max	Contains the maximum and minimum education and income levels among states
qryCounty_Min-Max	Contains the maximum and minimum education and income levels among counties within a chosen state
qrySelect_USA_Demog_Values	Contains demographic information for the US
qrySelect_State_Demog_Values	Contains demographic information for the chosen state
qrySelect_County_Demog_Values	Contains demographic information for the chosen county

10. Discussions

The tool provides a "ballpark estimate" of avertable deaths, diabetes prevalence, and diabetes cost with sufficient efficiency. It uses data to give the public a sense of magnitude about the importance of social determinants of health, but several factors affect its precision. Symmetric regression was used to get an avertable relationship between education and income, which provided consistency regardless of which slider the user changes. The data for CMR model are from 2008-2010 and may not reflect the current economy or current conditions in a county or state. The estimates are approximations based on a generalized linear regression equation, thus the observed values for any given county or state are different from the values predicted by the equation. In some cases, the tool relies on "imputations" of data for rural or small counties with sparse populations. The tool examines cross-sectional associations (the correspondence between education or income levels and the death rates during the same time period). Any health consequences of changing social determinants often occur many years later. The tool displays the diabetes cost generated by Medicare expenditure and national diabetes per capita cost. We made an assumption the diabetes costs for persons age less than 74 are proportional to Medicare expenditures—and older population—at county and state levels.

The most important caveat is that the tool does not imply "causality": one cannot assume that all the deaths that this tool designates as "avertable" can be prevented by changing education or income alone. Diplomas and money help, but the avertable deaths calculated by the tool reflect the total package of living conditions that exist in areas with higher education or income. People with more education or income tend to have better jobs and working conditions and greater access to medical care when they need it; they also tend to live in healthier neighborhoods with better access to nutritious foods, cleaner air and drinking water and less violent crime. Residents in such circumstances are less likely to smoke or be overweight or obese. All of these factors increase a person's chances of living a healthier, longer life.

11. References

[1] Centers for Disease Control and Prevention, National Center for Health Statistics. Underlying Cause of Death 1999-2010 on CDC WONDER Online Database, released 2012. Data for year 2009 are compiled from the Multiple Cause of Death File 2009, Series 20 No. 2O, 2012, Data for year 2008 are compiled from the Multiple Cause of Death File 2008, Series 20 No. 2N, 2011, data for year 2007 are compiled from Multiple Cause of Death File 2007, Series 20 No. 2M, 2010, data for years 2005-2006 data are compiled from Multiple Cause of Death File 2005-2006, Series 20, No. 2L, 2009, and data for years 1999-2004 are compiled from the Multiple Cause of Death File 1999-2004, Series 20, No. 2J, 2007. Accessed at http://wonder.cdc.gov/ucd-icd10.html.

[2] Centers for Disease Control and Prevention, National Center for Health Statistics. Compressed Mortality File 1999-2010. CDC WONDER On-line Database, compiled from Compressed Mortality File 1999-2010 Series 20 No. 2M, 2010. Accessed at http://wonder.cdc.gov/wonder/help/ucd.html#2000 Standard Population.

[3] Centers for Disease Control and Prevention: National Diabetes Surveillance System. Available online at: http://apps.nccd.cdc.gov/DDTSTRS/default.aspx.

[4] Dartmouth Atlas of Health Care. A New Series of Medicare Expenditure Measures by Hospital Referral Region: 2003-2009. 2012. Accessed at http://www.dartmouthatlas.org/tools/downloads.aspx#reimbursements.

[5] American Diabetes Association. Economic Costs of Diabetes in the U.S. in 2007. *Diabetes Care* 31:596–615, 2008.

[6] U.S. Census Bureau, American Community Survey 3-Year Estimates. Accessed at http://factfinder.census.gov/servlet/DCGeoSelectServlet?ds_name=ACS_2009_3YR_G00_.

[7]U.S. Census Bureau, Small Area Estimates Branch, January 2009. Accessed at http://www.census.gov/did/www/saipe/data/statecounty/data/index.html.

[8] A Compass for Understanding and Using American Community Survey data, *What General Data Users Need to Know*. October 2008. Accessed at http://www.census.gov/acs/www/Downloads/handbooks/ACSGeneralHandbook.pdf.

[9] Efron, B. (1978). Regression and ANOVA with zero-one data: Measures of residual variation. *J. Amer. Statist. Soc.* 73:113-121.

[10] Isobe T., Feigelson E. D. and et al. (1990). Linear Regression in Astronomy I. *Astrophysical Journal*, 364: 104- 113.

[11] Alexander V.E. and Christof S. (1998). *Regression Analysis for Social Science*, p222. Academic Press.