Who is Likely to be Healthy in the United States?

Presented by:
Chris Musialek and Mossaab Bagdouri
Computer Science Department
University of Maryland, College Park
October 5th, 2012

1. Introduction
In this paper we look at noticable correlations between the health of the population, the availability of medical facilities, and the prevalence of health insurance across US counties and states. Below we present the results of our data analysis. Overall, there appear to be some interesting correlations regarding the mean income of a population and the number of health businesses, the race of a population and the level of insurance, and the mean income of a population and its overall health.

2. Data Sources
For our analysis, we utilized three datasets, each containing a county FIPS code. This key helped to map the fields of each dataset to one another.

2.1. Census County Business Patterns
This 2010 US Census Bureau dataset provides granular economic data by industry and location [1]. We looked at three major fields (county FIPS code, NAICS code and Total Number of Establishments) out of 26. From an original set of 2,155,389 rows, we filtered out categories not related to health by keeping only rows with NAICS health category "Health Care and Social Assistance". Using population counts, we estimated the number of health businesses per capita for each county.

2.2. Small Area Health Insurance Estimates (SAHIE)
This 2005-2010 US Census Bureau dataset provides estimates of health care coverage for states and counties in the US [3]. Out of the original 216,725 rows and 22 columns per year, we looked at the number of insured and uninsured people in different demographic groups categorized by county, race and gender.

2.3. County Health Rankings
This dataset from The Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute measures the health of nearly all counties and ranks them within states [2]. We focused on the percentage in poor or fair health (two worst values in a 5 points scale) for 2010, which provided a coarse way to estimate the health of a county and state population.
3. Headlines

3.1. Health Businesses Prefer the Rich

The average number of health businesses per capita remains steady from low income until mean income reaches $100K, after which the number of health businesses per capita increases ~30%. As highlighted in Figure 1, the top end of the income spectrum has a higher number of health businesses per capita.

![Mean Income vs. Number of Health Businesses per Capita](image)

**Figure 1**: Counties whose mean income is less than $50K have populations with a much greater likelihood of being in poor health. Health businesses cater to counties with mean incomes greater than $100K.

Also of note, there appears to be an income threshold for reaching or surpassing the average health level of the US population. As shown by the colors in Figure 1, counties with a mean income below roughly $50,000 are the ones with the greatest likelihood of being in poor health. Figure 2 shows that poor health quality is inversely related to income. In fact, the equation of the fitting power curve is approximately of the form $y = \frac{\alpha}{x^n}$. Interestingly, the scatter plot shows a high dispersion around this curve.
3.2. White Females Most Likely To Be Insured

In Figure 3, each colored line corresponds to the level of health insurance coverage of a given race for a specific state. It is clear that Whites are more insured than Blacks, which themselves have a higher insurance rate than the Hispanics. This trend remains steady over a period of 6 years. White levels across all states almost never overlap any of the Hispanic coverage levels. That is, the worst states in health coverage for the Whites (Montana and Florida) have a higher level of health insurance than that of the Hispanics in all the states, except Massachusetts, Hawaii, and the District of Columbia. The Black race remains in between these two races, sharing comparable coverage with the worst states in health coverage for the Whites, and the best states for the Hispanics.

Females are also more likely to be insured. Using the health insurance margin of error between males and females, we present a map visualization in Figure 4. If the confidence interval for the percentage of females (males) insured is absolutely higher than that of males (females), than the corresponding county is colored pink (blue). Otherwise, the county is colored yellow. It is interesting to note that most of the counties are yellow, and none are blue. Since some counties (369 out of 3142) are pink, it may be argued that females have slightly better health coverage than males.
Figure 3: Whites have higher insurance rates than Blacks, who themselves have higher coverage than Hispanics.

Figure 4: In 12% of the counties, females are significantly more covered with health insurance than males. In no county are males significantly more covered than females.
3.3. American Health Varies Widely Per State

We used HCE to cluster the states by various characteristics. Clustering by the percentage insured and the health of the population found similar distinct groups of counties, shown in Figure 5. While some counties had high insurance rates and good health - an expected outcome - others such as those in Texas were grouped together as having good health and low insurance rates.

Hierarchical Clustering of Counties by Percentage Insured & Percentage Poor or Fair Health

![Figure 5](image)

Figure 5: Several counties in Texas have a low percentage of the population insured and also a low percentage in poor or fair health. Green denotes a low number, and red a high number.

In addition, hierarchical clustering by mean income and percentage in poor or fair health identified a small group of counties, mainly in Kentucky, whose mean income are below the average and which have the highest percentage of poor health (>35% poor/fair health). These counties are good candidates for health care assistance. Figure 6 shows the results of the clustering, which used euclidean distance to determine the closeness of the column relationships.

US Counties Requiring Health Care Assistance

![Figure 6](image)

Figure 6: A small group of counties, mainly in Kentucky, have both a low mean income and also a very high percentage of the population in poor or fair health.
4. Spotfire and HCE Critique

4.1. Spotfire

- The functionalities offered by the data import wizard are helpful. The possibilities of calculating new columns, pivoting and unpivoting tables were useful. However, there is a lack of additional data manipulation operations that facilitates the separation of data aggregated and accumulated across several columns.
- Managing data sources and relations was confusing. In fact, there was no easy way to review which data sources have already been added, or preview the data contained within those datasets. There should be a separate data sources management screen, which allows the user to view existing imported tables, view the data at will, and manage relationships all together.

4.2. HCE

- HCE had trouble importing csv files which didn’t have Windows line endings (CRLF line terminators). The tool should either add support for CR line terminators, or it should detect CR line terminators and present this as a to the user. Currently, the program crashes.
- HCE doesn’t deal well with missing values. In such cases, data had to be preprocessed before getting loaded into HCE.

References