CHAS Regression: Methods

Colorado Health Access Survey (CHAS) Data

The 2013 CHAS is a telephone survey of 10,224 randomly selected households in Colorado. It was administered between April 15 and July 27, 2013. Because more Coloradans are using cell phones, the 2013 survey includes 4,000 randomly selected households with a cell phone. Importantly, these are not cell-phone only households. Some may also have a landline phone. A member of each of the 10,224 households was randomly selected as the target for the survey questions. The information specifically relates to this targeted household member, even though the target was not necessarily the person answering the questions.

The selected households were divided among Colorado's 21 health statistics regions (HSRs) to ensure adequate representation within each of these areas and to allow most data to be analyzed at the regional level. Survey data were weighted to adjust for the fact that not all survey respondents were selected with the same probability.

The Colorado Health Institute – a nonpartisan institute focusing on data, information and analysis supporting health care policy decisions – manages the survey. It is funded by The Colorado Trust, a grant-making foundation dedicated to achieving fair and equal opportunities for Coloradans to lead healthy, productive lives regardless of race, ethnicity, income or where they live.

Limitations of Probability Regression Modeling

Regression modeling measures the statistical relationships between variables in a model. While this type of modeling helps us understand the association between variables, it does not allow us to determine causation. In other words, using regression modeling, we can conclude that certain characteristics of individuals are associated with health inequities, but we cannot conclude that they cause inequities.

Record Selection

This study focused on adults between the ages 19 to 64 years in order to include data related to employer-sponsored health insurance. Due to this focus, we analyzed only records from the CHAS for respondents between the ages of 19 and 64s. Of the 10,224 records, 6,384 records were used in this analysis.

Variable Selection

All demographic and background variables on the CHAS, such as gender, age, ethnicity, income, education attainment, geographical location, marital status, and language spoken at home, were considered for this analysis. Variables that were highly correlated with another variable were taken out of consideration. For example, income and educational attainment are highly correlated variables. We included income since it is often relevant to health policy discussions. A selection method was used during the modeling process to further eliminate some variables. (See uninsured/insured regression model section below.)

Imputation of Missing Variables

Due to the nature of survey research, there were missing values in the data set. These missing values were either populated using a leading question when appropriate or estimated through an imputation method.

The offer of employer-sponsored health insurance was populated using the question on whether a respondent was employed. If a respondent was not employed, the was set to "No".

The remaining missing values were imputed using methods from "A Multivariate Technique for Multiply Imputing Values Using a Sequence of Regression Models" published in 2001 by the Survey Methodology Program at the Institute of Social Research. Variables with missing data or "don't know" or "refused" answers that were not part of a skip pattern were imputed using a sequential regression approach. There were a number of variables that needed to be imputed. Out of the variables with missing values, the majority only had less than one percent missing values.

The general purpose of this imputation procedure is to handle relatively complex data structures where explicit full multivariate models cannot be easily formulated. The imputed values for each individual were conditional on all the values observed for that individual. The basic strategy was to create imputations through a sequence of multiple regressions varying the type of regression model by the type of variable being imputed. The variables in the data set were assumed to be one or more of the following four types: continuous, dichotomous, categorical with more than two categories, and counts.

Statistical Model

A linear probability regression model was used to conduct this analysis. The outcome for the initial model was an individual's health insurance status. The outcome variable was a dichotomous variable for insured or uninsured.

Ten predictor variables were initially included in the model:

- Geographic location: Urban/rural
- Citizenship status: U.S. citizen/non-citizen
- Income: At or below poverty/above poverty
- Gender: Male/female
- Parental status: Parent/childless adult
- Offered employer-sponsored insurance: Offered/not offered
- Self-reported health status: Excellent, very good or good/fair or poor
- Ethnicity: Hispanic/non-Hispanic
- Self-reported disability status: Disabled/not disabled
- Marital status: Married/single

A backward selection method was applied during the model building phase that excluded the parental status variable from the model. All other variables remained in the model as significant predictors.

Statistical Decomposition Process

One of the significant factors associated with the coverage gap, according to the model, is Hispanic ethnicity. It is significant even after these other factors - gender, family income, citizenship, availability of employer-sponsored insurance, self-reported health status, self-reported disability status and marital status – are held constant. If each factor were identical for each adult, Hispanics would still have a lower insurance rate than non-Hispanics, although the coverage gap would be narrowed. A statistical decomposition process was used to look at what factors are associated with the gap in health insurance status between Hispanic and non-Hispanic adults.

A linear probability model was chosen for this analysis. The regression coefficients from the model are multiplied by the difference between the means for Hispanic and non-Hispanic adults for each factor except ethnicity.

For example, consider the factor of citizenship status. From the CHAS, 77.6 percent of Hispanic adults and 97.7 percent of non-Hispanic adults are United States citizens. The regression coefficient from our original model for the citizenship variable is 0.237. In order to calculate the proportion of the 14.5 percentage point gap in health insurance coverage between Hispanic and non-Hispanic adults, we took the difference between 97.7 and 77.6, or 20.1, and multiplied it by 0.237 (the regression coefficient for citizenship status). This yielded 4.8. Therefore, citizenship status is associated with 4.8 points of the 14.5 percentage point gap in health insurance coverage between Hispanic and non-Hispanic adults. Dividing 4.8 by 14.5 yielded the 33 percent of the overall gap that is reported in the brief.