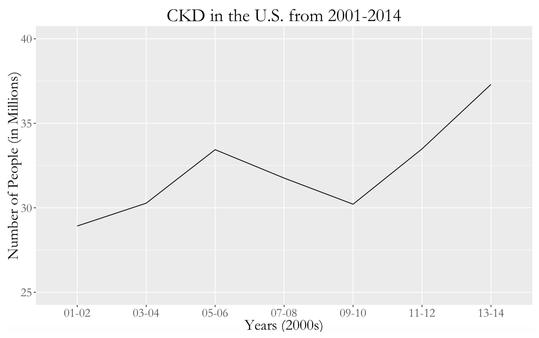
Predicting CKD and Potential Risk Factors with Multiple Linear Regression

Gabriel Goulart, Nicholas Hertle, and Lydia Lucchesi

UNIVERSITY OF MICHIGAN

CKD

Chronic kidney disease, or **CKD**, is an emergent epidemic in the United States. While ESRD (end-stage CKD) patients make up less than 1% of Medicare patients, ESRD-related expenditures total in the billions (approx. 7% of total Medicare spending). With the goal of identifying cofactors for this costly disease that greatly decreases quality of life, our research focuses on the association between CKD and other health data. Ultimately, we built a multiple linear regression model that predicts estimated glomerular filtration rate based on one's current state of health, socioeconomic status, and demographic.



NHANES

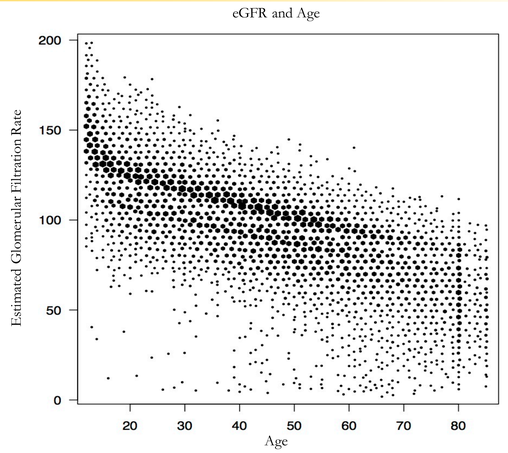
The data for this study comes from the **National Health and Nutrition Examination Survey** conducted by the CDC every year. Each survey cycle consists of two years and attempts to obtain a representative sample of the United States' population. Our dataset has 78,518 observations, contains 100 different variables, and spans 14 years (2001-2014). Using the 'Survey' package in R, we were able to account for the complex survey design used by the CDC and get a strong sense of what is happening among roughly 320 million people.

Methods

1. Narrow down variables of interest through literature review, visualizations, and chi-squared tests.

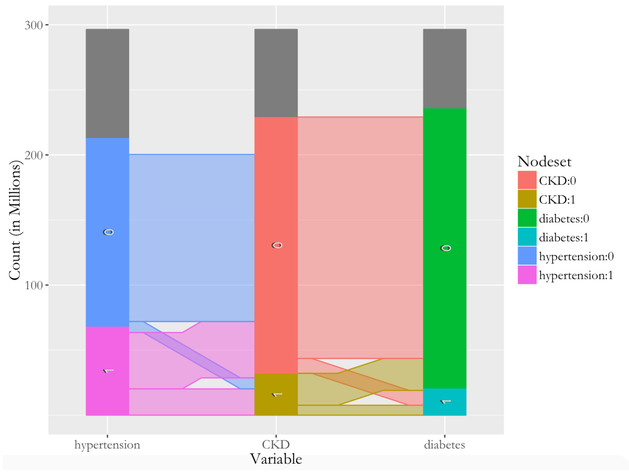
Discard some variables due to a high number of missing values and the inability to determine what the different categorical levels meant (e.g. self reported strokes).

Subset the data so that only those older than 18 are included and transform age into a categorical variable.



Using hexagonal binning, this scatter plot demonstrates strong a linear relationship between age and estimated glomerular filtration rate

Parallel sets visualization used to understand relationships between categorical variables in multivariate datasets



2. Use backward selection to get GLMs of the form:

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots \ + \beta_i x_i + \ \varepsilon_i$$

with the lowest BIC value.
Adjust the initial set of variables after finding that some were never significant.

3. Perform 10-fold cross validation and bootstrap resampling to find the most accurate multiple linear regression model for predicting estimated glomerular filtration rate (measurement for determining CKD stage).

Penalize errors on higher-weight observations by weighting errors on specific PSUs by corresponding survey weight to maintain internal consistency.

Results

Multiple Linear Regression Model:
eGFR ~ Hypertension + Gender + Diabetes + Race + Age Group

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots \ + \beta_i x_i + \ \varepsilon_i$$

Valid Observations for Analysis: **85%** (34,332)
Mean error: **12.902**
Standard deviation: **0.0771**

Coefficient		Pr(>|t|)
(Intercept)	118.5930	< 2e-16 ***
hypertension	-4.8593	< 2e-16 ***
diabetes	-0.9559	0.0275 *
age40to59	-15.9856	< 2e-16 ***
ageAbove60	-34.0484	< 2e-16 ***
male	-1.6229	1.30e-11 ***
other.hispanic	-4.4202	7.61e-12 ***
non.hispanic.white	-9.8071	< 2e-16 ***
non.hispanic.black	-0.6251	0.1360
other.and.multi.racial	-4.7807	2.62e-15 ***

Conclusion and Recommendations for Further Study

For NHANES, new individuals are surveyed each year, and geographic location is kept confidential. The inability to track the health of a single person over time prohibited us from performing causation analysis.

While the 'Survey' package in R is incredibly helpful at properly weighting each observation, many R functions are not fully supported.

Utilizing bootstrap resampling, we found that the coefficients' confidence intervals were generally clustered around the one-shot values under our Results section, although the 95% confidence interval for Non-Hispanic Black contained 0. Therefore we are unable to conclusively identify a difference in eGFR for that group.

After removing all observations that contained a missing value for one or more covariates, only 85% were left for analysis. Adding more covariates reduced this number further and we cannot conclude that the data was missing completely at random. Future studies could look at building models to impute missing values.

Acknowledgements

This project would not have been possible without the mentorship of Dr. Yanming Li, Dr. Jian Kang, and Dr. Kevin He and also without the support of the Big Data Summer Institute. We are very thankful for the opportunity to conduct research at the University of Michigan.

Citations

T. Lumley (2014) "survey: analysis of complex survey samples". R package version 3.30.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

Heike Hofmann and Marie Vendettuoli (2015). ggparallel: Variations of Parallel Coordinate Plots for Categorical Data. R package version 0.1.2.

Dan Carr, ported by Nicholas Lewin-Koh, Martin Maechler and contains copies of lattice functions written by Deepayan Sarkar (2015). hexbin: Hexagonal Binning Routines. R package version 1.27.1.

Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20.

Winston Chang, (2014). Extrafont: Tools for using fonts. R package version 0.17.

"Introduction to Volume 1: CKD in the United States." American Journal of Kidney Diseases 67.3 (2016): n. pag. Web.

"UNITED STATES RENAL DATA SYSTEM." USRDS Home Page. N.p., n.d. Web. 17 July 2016.