Comparing Principal Component Analysis to a Traditional Approach to Variable Selection: Correlates of Inpatient Hospitalization Days

Christian Dunlap[1], Anja Zgodic, ScM[1]; Jan M Eberth, PhD[1]

Rural and Minority Health Research Center, University of South Carolina

RURAL & MINORITY Health Research Center

UNIVERSITY OF SOUTH CAROLINA

BACKGROUND AND PURPOSE

- The number of inpatient hospitalization days are expected to increase 19% by 2025, creating a burden on the U.S. healthcare system.[1]

- Racial, ethnic, gender, and economic disparities have been previously found in the healthcare system[1], however, studies on disparities in inpatient hospitalization days are few.

- Statistical methodologies accounting for exposures and confounding factors can elucidate factors associated with increased inpatient hospitalization days in patient subpopulations.

Purpose:
To comparatively examine the relationship between inpatient hospitalization days and sociodemographic factors using Principal Component Analysis (PCA) and traditional Stepwise Linear Regression.

METHODS

- We used a publicly available dataset from the U.S. Department of Health and Human Services, called *Area Health Resources Files*, which includes data on health facilities, population characteristics, and hospital utilization, at the county level.[2]

- Factors considered in this study included but were not limited to race/ethnicity, insurance coverage, age, income, population density, education, death rate, and average household demographics.

- We performed stepwise linear regression where a final model was selected based on the Akaike Information Criterion. We then performed Principal Component Analysis (PCA).

- We compared variables selected by the final stepwise regression model and the PCA to examine overlap and method concordance.

RESULTS

Table 1: Regression Coefficients Significance Level and Principal Component Loadings of Selected Variables

Final Stepwise Regression Model - Selected Predictors	Stepwise Regression Significance	Principal Components 1-4 Correlation
Percent White Male	p < 0.001	> |0.20|
Percent White Female	p < 0.001	> |0.20|
Percent Hispanic Male	p < 0.001	> |0.28|
Percent Hispanic Female	p < 0.001	> |0.28|
Percent Asian Female		> |0.20|
Percent Asian Male		> |0.28|
Percent Hawaiian/Pacific Islander Male		> |0.20|
Percent Hawaiian/Pacific Islander Female	p < 0.001	> |0.20|
Percent Black Male	p < 0.001	> |0.20|
Percent Black Female	p < 0.001	> |0.20|
Percent White with no High School Diploma	p < 0.05	> |0.28|
Percent Black with no High School Diploma		
Percent White with High School Diploma or more		> |0.28|
Percent Black with High School Diploma or more	p < 0.05	
Percent Multiracial with no High School Diploma	p < 0.05	> |0.20|
Percent Asian with High School Diploma or more		
Median Household Income	p < 0.001	> |0.28|
Percent Uninsured less than 65 years old		> |0.28|

- All variables selected by the stepwise regression model (p-value < 0.001 , $R^2 = 0.87$) with significant p-values were also found to show higher correlations to the first four principal components, which accounted for 53% of total variance in the data.

- Conversely, **not** all variables with stronger component loadings were found to be statistically significant by Stepwise Regression.

- Accounting for other sociodemographic characteristics, race/ethnicity were highlighted by both methods as important a correlate of the number of inpatient hospitalization days. In PCA, race/ethnicity appeared in selected components. In regression, race/ethnicity was highly significant.

- Other important county-level factors according to both approaches were: median household income, % of population over 65, hospital admissions per 1000 residents, population density, and total standardized Medicare costs, with p-values < 0.001 and component loadings > \|0.28\|.

Key Findings:
- The traditional approach (stepwise regression) and PCA mostly concorded with one another for all but 8 out of 37 variables.

- Socioeconomic factors are valuable predictors of inpatient hospitalization days according to both models.

DISCUSSION

- Both economic factors such as median household income and social factors such as race and education play a large role in predicting inpatient hospitalization days.

- Next steps include comparing a linear regression model using the first four principal components as predictors to the stepwise regression model.

- Further dimensionality reduction (factor analysis) and variable selection techniques (regularization) could be employed to examine variable selection concordance across approaches.

REFERENCES

1. Dall, T. M. et al. An aging population and growing disease burden will require a large and specialized health care workforce by 2025. Health Affairs, 32(11), 2013–2020. https://doi.org/10.1377/hlthaff.2013.0714
2. Health Resources and Service Administration. Area Health Resources File. 2013-2019. https://data.hrsa.gov/topics/health-workforce/ahrf