Socioeconomic Status and Racial Demographics as Determinants of Healthcare Coverage

Abstract

Even after the passage of the Affordable Care Act (ACA) in 2010, Americans without proper healthcare coverage have hovered at around 10%. In this project, we aimed to investigate whether racial and socioeconomic factors influence a given community's healthcare coverage. We used data from the Seattle Census tracking healthcare coverage by employment and racial demographics and socioeconomic factors of 180 neighborhoods in the city of Seattle. We constructed an initial multiple linear regression model with 8 initial predictors and 1 response variable. Our final model highlighted 3 variables, employment status, poverty rate, and education level, as statistically significant predictors of a given community's healthcare coverage in Seattle. Our research revealed that socioeconomic and racial demographic factors of a community were key underlying contributors to healthcare inequality in the United States, providing useful insights to possible future directions towards improving the healthcare system.

Background and Introduction

The United States is one of the richest countries that does not choose to provide universal healthcare. Instead, a robust private health insurance market exists in the United States, with many Americans getting their healthcare through their employer. Since the passage of the Affordable Care Act (ACA) in 2010, uninsurance rates have dropped significantly, from roughly 17% to 10%. Since 2017, however, uninsured rates have stagnated, revealing the stagnating process in reducing the number of Americans without health insurance.[5] Today, the number of uninsured Americans hovers at around ten percent. Prior research reveals that most uninsured people cite the high cost of coverage as the reason for not having healthcare. According to a Kaiser Family Foundation research in 2021, 64% of uninsured adults said that they were uninsured because the cost of coverage was too high.

In the United States, most people's reliance on their employers adds to the inaccessibility of healthcare.[5] The ACA attempted to supplement healthcare coverage for those without insurance through their employers, but the act has not done a comprehensive job. The persistent problem of lack of health insurance in the United States led us to look into what factors still contribute to a given population's level of healthcare coverage. We were especially interested in discovering whether employment status is associated with healthcare coverage since providing unemployed people with healthcare was a main goal of the ACA. Therefore, we built a multiple linear regression model to predict healthcare coverage according to a variety of variables measuring racial demographics and socioeconomic status.

Data and Exploratory Analysis

a. *Data and variables*

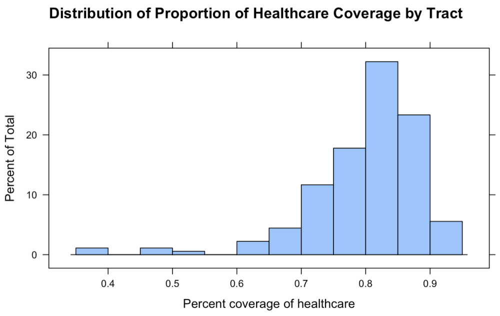
Because specific healthcare policy varies across the country, we wanted to narrow our focus to a specific geographic area, with diverse socioeconomic and racial demographics. We decided to focus on a geographic area where the state government chose to expand Medicaid, giving the option of government-subsidized health care to all citizens who fall below 138% of the poverty level.[1] This policy decision reflects the majority of the United States, as only Florida, Georgia, Kansas, Mississippi, South Carolina, Wisconsin, and Wyoming have chosen not to expand Medicaid.[4]

We identified two quality datasets focusing on the city of Seattle: one detailed healthcare coverage by employment while the other tracked racial, demographics, and socioeconomic factors.[2, 3] We combined them based on unique GEO IDs from the same census tracts used in both to distinguish different neighborhoods (see map in **Appendix A**). We only selected data from 2021 and removed any neighborhoods with missing values, which gave us complete data for 180 neighborhoods in Seattle to analyze. The data was originally collected as counts, which we normalized into percentages to account for the different population size of each neighborhood.

Our response variable is the percentage of a given neighborhood population with health insurance (% Healthcare Coverage). Our initial model included 6 quantitative predictor variables: employment rate (% Employed), percent neighborhood population whose income falls below 200% of the poverty level (% Under Poverty), percent neighborhood population identifying as a non-white race, multiracial, or Hispanic/Latino ethnicity (% People of Color), percent neighborhood population with less than a college degree (% Less than a Bachelor's Degree), % English Language Learners, and Life Expectancy at Birth. We also included 1 categorical predictor variable, Health Disadvantage Quintile, which was calculated from multiple health condition metrics such as the percent population in a given neighborhood with asthma or obesity, and split into 5 equal-sized categories.

b. *Exploratory Data Analysis*

We found that the distribution of the response variable (% Healthcare Coverage) is fairly left skewed, unimodal, and has a center at 0.8. We looked at the relationship between each predictor variable with the response variable and found that % Employed and % Under Poverty had the strongest correlations with % Healthcare Coverage, and these scatterplots did not reveal any outliers that dramatically impacted the correlation coefficients (**Appendix B).** Different categories of Health Disadvantage Quintile do not display significant difference in % Healthcare Coverage.



Predictor Variable	Correlation with Response Variable (R^2)
% Employed	0.88
% Under Poverty	-0.744
% People of Color	-0.454
% Less than a Bachelor's Degree	-0.452
% English Language Learners	-0.411
Life Expectancy at Birth	0.185

Model and Results

a. *Analytic Methods*

We created a multiple linear regression model (MLR) from this dataset to answer our research question. We selected MLR because our response variable (% healthcare coverage) and 7 independent predictors are all continuous quantitative variables. We also added 4 indicator variables to our model to account for 1 categorical predictor variable with 5 categories (Health Disadvantage Quintile). We did not use any interaction terms because it would only further complicate the initial model which already has 8 predictors. Hence, we made a reasonable assumption that the effect of our quantitative variables does not differ by different health quintile categories.

Among the initial model diagnostic plots (scatterplots of every pair of predictor variables, residual versus predicted values, residual versus each predictor variable, histogram of residuals, and QQPlot), we observed heteroskedasticity in the residual plots. Therefore, we performed a Boxcox Transformation of the response variable (% healthcare coverage) to the power of 1.5 (**Appendix D**). All model assumptions of MLR – linearity, equal variance, mean zero, independence, and normality – were satisfied after transformation (**Appendix E**). No outliers were identified.

We observed minor multicollinearity (VIF < 5) between 4 of our predictors, % People of Color with % Under Poverty, % Less than a Bachelor's Degree, and % English Language Learners (**Appendix F**). We then performed an F-test to find that our full model is a more effective predictor than an intercept-only model, and we used both-direction stepwise selection to remove 5 out of 8 predictors from the initial model based on AIC score minimization. The same result was confirmed by forward, backward, and best subset selections (**Appendix G**). The 3 predictors left were % Employed, % Under Poverty, and % Less than a Bachelor's Degree. We then performed t-tests and calculated confidence intervals to confirm the significance of the slope of each predictor.

b. Final Model and Results

$$\% \ \widehat{Healthcare\ Coverage}^{1/3} = 0.034 + 0.94(\text{\% Employed}) - 0.92(\text{\% Under Poverty}) - 0.16(\text{\% Less than a Bachelor's Degree})$$

Our final model, as described by the equation above, highlighted % Employed, % Under Poverty, and % Less than a Bachelor's Degree as the 3 key predictors of % Healthcare Coverage of a neighborhood in the city of Seattle. The F-test we conducted yielded F(3, 176) = 332.8 with p-value = $2.2\text{x}10^{-16}$. Since p-value < 0.05 (α=0.05), we concluded that our final model was more effective at estimating % healthcare coverage than an intercept-only model. The model's R^2_{adjusted} = 0.833, which suggested that 83.34% of the variance in % Healthcare Coverage can be explained by our MLR model. The small residual standard error of 0.04645 indicated narrow prediction and confidence intervals which would provide better estimations.

In addition to assessing the overall model effectiveness, we also evaluated the individual slopes of each predictor using the t-test. All slopes yielded p-values < 0.05, which allowed us to conclude that holding the other 2 variables constant, each of the three predictor variables in our final model are significant in predicting the % healthcare coverage of a Seattle neighborhood. We constructed 95% confidence intervals for all slopes and no interval overlapped with zero, which further supported the significance of our predictors and the effectiveness of our model.

	MLR slope estimates	p-value from t-test	95% confidence interval
% Employed	0.94	$2.2*10^{-16}$	[0.832,1.047]
% Under Poverty	-0.92	0.018	[-0.167,-0.0159]
% Less than a Bachelor's Degree	-0.16	$1.04*10^{-9}$	[-0.212,-0.113]
Intercept	0.034	0.490	N/A

Conclusion and Discussions

In this project, we aimed to investigate racial and socioeconomic factors that could influence a given community's healthcare coverage. We identified three predictor variables, % Employed, % Under Poverty, and % Less than a Bachelor's Degree, that have statistically significant correlations with % healthcare coverage of neighborhoods in the city of Seattle.

The presence of employment status in our final model suggests that many people still rely primarily upon their employer for healthcare coverage. We also suspect that cost is a limiting factor in people's access to health insurance because communities where a greater proportion of the population are below 200% of the poverty level tend to have less health insurance coverage. Interestingly, our model also shows that communities with a higher average education level are more likely to have better healthcare coverage, which might be due to better awareness regarding healthcare knowledge. However, since we have observed moderate multicollinearity of education level with % people of color and % English language learners, the significance of education level as a predictor may be the result of it overlapping the influence (over the response variable) of other predictor variables not included in the final model.

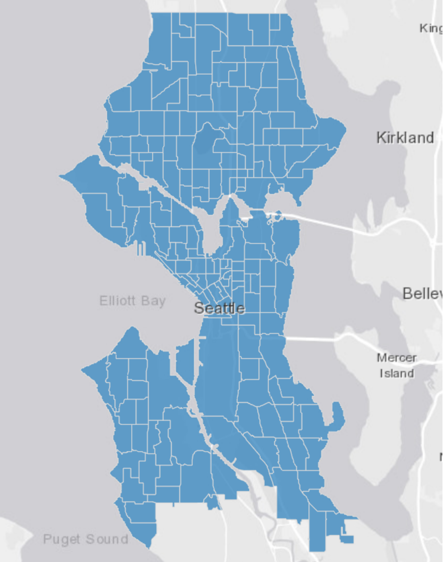
Our model is also limited in that it can only predict healthcare coverage on the population level such as neighborhoods but has no power in predicting the likelihood of an individual having healthcare coverage or not. Additionally, the scope of our model only pertains to the city of Seattle and thus, its prediction powers cannot expand to include other parts of the country. For future directions, we would like to further investigate factors that affect healthcare coverage on an individual level, while also expanding our dataset to include other parts of the country for a more comprehensive model.

References

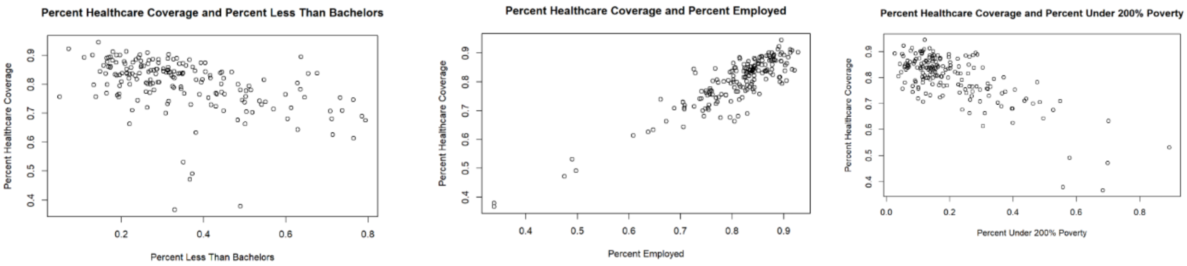
1. “Federal Poverty Level (FPL) - Glossary.” *HealthCare.Gov*, https://www.healthcare.gov/glossary/federal-poverty-level-fpl. Accessed 10 Dec. 2023.
2. *HEALTH INSURANCE BY EMPLOYMENT STATUS (B27011)*. https://data-seattlecitygis.opendata.arcgis.com/datasets/SeattleCityGIS::health-insurance-by-employment-status-b27011/about. Accessed 8 Dec. 2023.
3. *Racial and Social Equity Composite Index Current*. https://data-seattlecitygis.opendata.arcgis.com/datasets/3a6bcc7fa4c14c4daabdb1cd8f329758_0/about. Accessed 8 Dec. 2023.
4. “Status of State Medicaid Expansion Decisions: Interactive Map.” *KFF*, 1 Dec. 2023, https://www.kff.org/medicaid/issue-brief/status-of-state-medicaid-expansion-decisions-interactive-map/.
5. Tolbert, Jennifer, et al. “Key Facts about the Uninsured Population.” *KFF*, 19 Dec. 2022, https://www.kff.org/uninsured/issue-brief/key-facts-about-the-uninsured-population/.

Appendix

A. Seattle neighborhoods denoted by census tract GEO IDs



B. Scatterplot relationship of three predictor variables (% Less than a Batchelor's Degree, % Employed, and % Under Poverty)

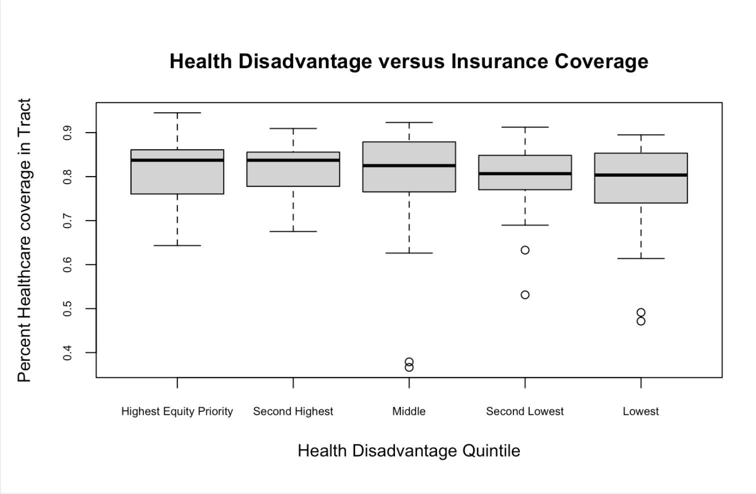


R = -0.45

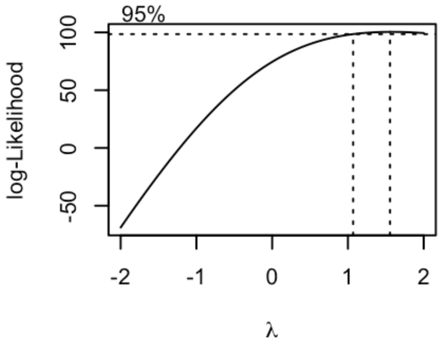
R = 0.88

R = -0.744

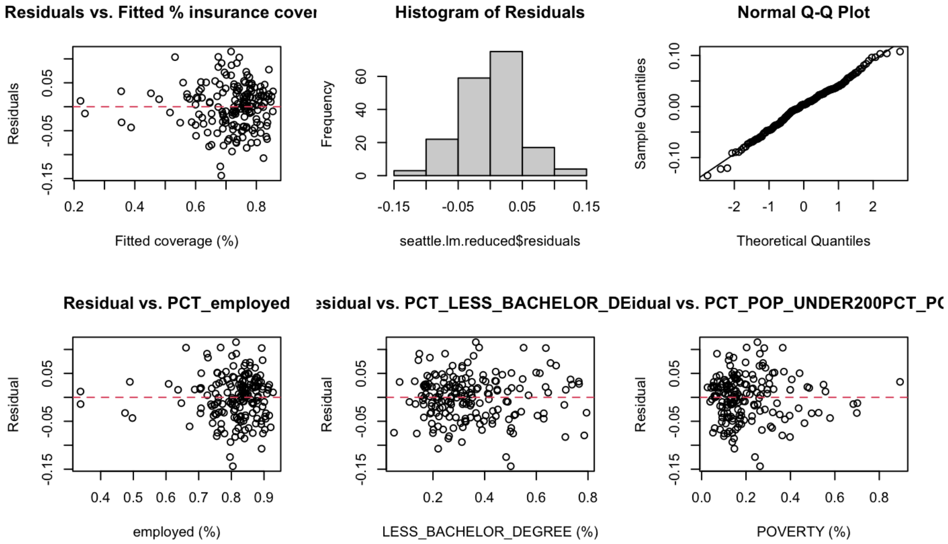
C. Side by side boxplot of % Healthcare Coverage by Health Disadvantage Quintile



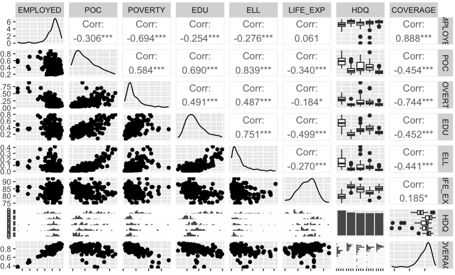
D. BoxCox transformation

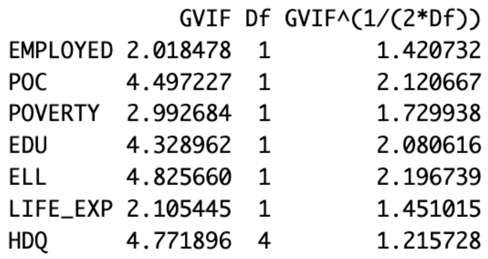


E. Final model diagnostic plots (see scatterplots of every pair of predictor variables in E.)



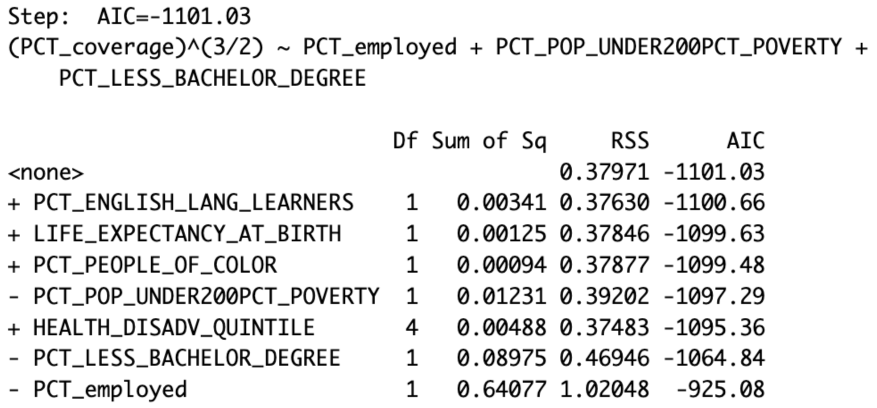
F. Multicollinearity plots and VIF values of all preliminary predictor variables





```
              GVIF Df GVIF^(1/(2*Df))
EMPLOYED 2.018478  1        1.420732
POC      4.497227  1        2.120667
POVERTY  2.992684  1        1.729938
EDU      4.328962  1        2.080616
ELL      4.825660  1        2.196739
LIFE_EXP 2.105445  1        1.451015
HDQ      4.771896  4        1.215728
```

G. both-direction stepwise selection process



```
Step:  AIC=-1101.03
(PCT_coverage)^(3/2) ~ PCT_employed + PCT_POP_UNDER200PCT_POVERTY +
    PCT_LESS_BACHELOR_DEGREE

                              Df Sum of Sq     RSS      AIC
<none>                                     0.37971 -1101.03
+ PCT_ENGLISH_LANG_LEARNERS    1   0.00341 0.37630 -1100.66
+ LIFE_EXPECTANCY_AT_BIRTH     1   0.00125 0.37846 -1099.63
+ PCT_PEOPLE_OF_COLOR          1   0.00094 0.37877 -1099.48
- PCT_POP_UNDER200PCT_POVERTY  1   0.01231 0.39202 -1097.29
+ HEALTH_DISADV_QUINTILE       4   0.00488 0.37483 -1095.36
- PCT_LESS_BACHELOR_DEGREE     1   0.08975 0.46946 -1064.84
- PCT_employed                 1   0.64077 1.02048  -925.08
```