# What Impacts the Cost of Healthcare?

Autumn Smith
Wellesley College 2022
Data Science Major Capstone

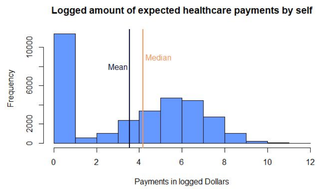
## Background

It is no secret that healthcare costs in the United States tend to be high. The aim of this project is to consider how much an individual may pay for healthcare, given certain characteristics.

To investigate my question, I used a dataset consisting of 81 columns and 31,880 rows from IPUMS Health Surveys. Specifically, I used the the IPUMS MEDICAL EXPENDITURE PANEL SURVEY, which is a set of large-scale surveys of families and individuals, their medical providers, and employers across the United States that is implemented through the Department of Health and Human Services.

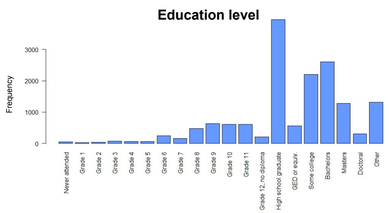
## Research Question

What factors influence how much an individual is expected to pay for healthcare services?

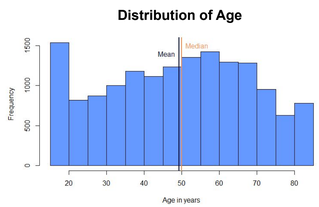



The response of interest is EXPSELFPAY in the year 2017, which captures the sum of direct payments made by the person and the person's family for care provided during the year.

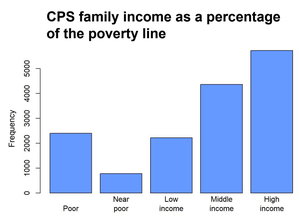
## Exploratory Data Analysis




Highest level of education completed by the year 2017.



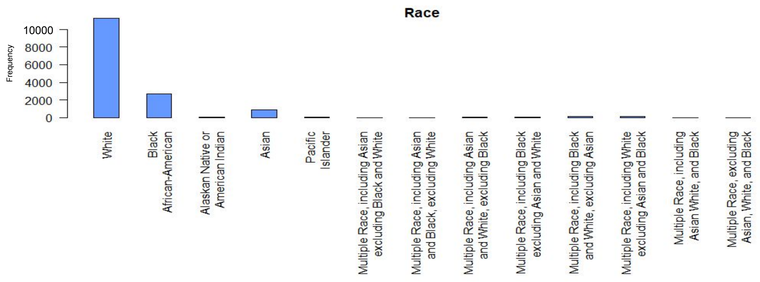

The correlation between age and logged expected payments is about 0.38 which suggests a light, positive correlation between age and expected healthcare payments.



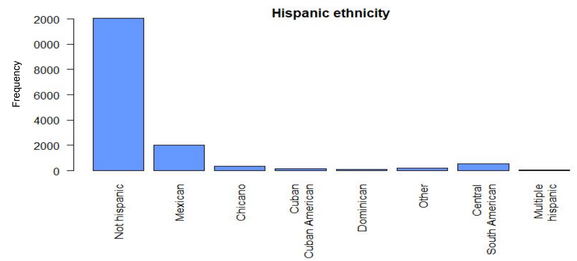

In 2017, the federal poverty line was defined to be $12,060 for a one-person household, or $24,600 for a four-person household.Over half of all survey respondents, 65.1%, are in middle or high income brackets

Data Source:
Lynn A. Blewett, Julia A. Rivera Drew, Risa Griffin, Kari C.W. Williams, and Daniel Backman. IPUMS Health Surveys: Medical Expenditure Panel Survey, Version 1.0 [dataset]. Minneapolis: University of Minnesota, 2018. http://doi.org/10.18128/D071.V1.0

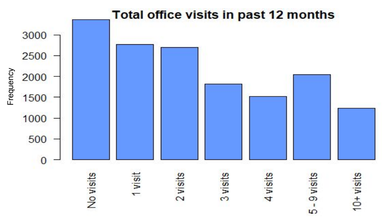
## Exploratory Data Analysis Continued




Out of all surveyed individuals, 72.8% identify as white, 17.2% identify as Black or African-American, and 6% identify as Asian. All other race breakdowns amount to 3.9%.




Out of the survey, 21.9% of individuals identify as Hispanic, with the majority of those identifying as Mexican.




21.7% of surveyed people had no medical office visits in the past 12 months. Out of those who did have an office visit, the majority had between one to three visits.

## Model Building Process

The final list of features used in the model is:

1. Age
2. Legal marital status
3. Education level
4. Sex
5. Race
6. Hispanic ethnicity
7. Number of healthcare visits in the past year
8. CPS family income as a percentage of the poverty line
9. Insurance status
10. Private insurance status

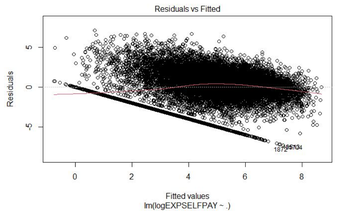
I built a linear regression model that first ran on 12 features, resulting in 62 coefficients.

A VIF threshold of 5 was used to examine the presence of multicollinearity.

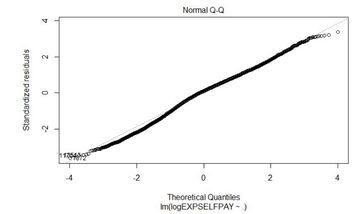
Stepwise regression was performed to determine the best first order model, resulting in the drop of variables related to marriage status and usual travel time to healthcare services.

The final model ran on 10 features and has an adjusted R-squared value of 0.3794.

## Model Evaluation




The residuals do not appear to be random, indicating a non-linear model may be appropriate.




VIF Scores, AIC Criterion

```
##                GVIF Df
## AGE        1.310574  1
## EDUC       1.787484 19
## SEX        1.035348  1
## RACEA      1.255611 11
## HISPETH    1.460583  7
## VISITYRNO  1.343787  6
## POVCAT     1.651248  4
## HINOTCOV   1.194788  1
## HIPRIVATE  1.599098  1
## ADILCR     1.147709  1
```

The VIF scores, calculated with the AIC criterion, are all under 5, indicating the absence of a severe multicollinearity issue.

## Key Findings & Conclusions

The model suggests that the following features are significant:

- Age
- Race
- Sex
- Family income
- Insurance coverage status
- Whether an individual needed immediate medical attention in the past 12 months.

Per the model, those who identify as Black or Asian, on average, are expected to pay less than White individuals for healthcare, and those who identify as some form of Hispanic are also expected to pay less for healthcare. This could be due to a number of factors, but one route particularly worth considering is that the number of individuals on private insurance, broken down by race, is not consistent. For example, White and Asian individuals are more likely to have private insurance than those who are Black or Hispanic. Costs covered by private insurance vary, but the quality of doctors who accept non-private forms of insurance vary. Thus, the costs paid may vary by race, but the amount paid cannot speak to the quality of care received.

Sex also was deemed to be a significant predictor, with the model suggesting that those who are female are, on average, expected to pay more for healthcare. This could be due to several reasons, one of which being that there are many specialty healthcare services for people who are female that are not considered to be a part of general healthcare practice. In these cases, seeing a specialist for what are colloquially known as “woman’s health services” would incur more costs.

In conclusion, the model suggests that factors such as race and sex can explain the cost of healthcare for an individual. Further work exploring the relationships between the acquisition of private insurance, race & sex, and the quality of healthcare may shed more light on how much individuals are expected to spend on healthcare.