Analyzing Socioeconomic Effects on Health of Workers

Kyle Yuen

Georgia Institute of Technology
Professor Huo

Summary

This report focuses on various socioeconomic factors that have the largest impact on the health of workers taken from the R dataset “Wage”, modelled using Logistic Regression, Smoothing Splines, and the Generalized Additive Model (GAM).

Introduction

The Generalized Additive Model combines several nonlinear variables while keeping the addictive structure of the linear model. In this report, various factors that could affect a person’s health are analyzed. To figure out the relation between them, the GAM model is used since many factors may not be linear and the GAM model offers a lot of flexibility for multiple nonlinear factors.

Current Events

Given the current state of Covid-19, and the drastic effects it has had on workers in the lower working class, we believe that it is necessary to look into the impact wages, job classification, education, and race have on the health of individuals.

Wages impact health in a variety of ways. Higher wages enable people to obtain better healthcare services, and higher wages increase job satisfaction, which in turn improves mental health. As will be shown in the analysis, wages have the most significant impact on predicting the health outcomes of an individual. With the lockdowns that accompanied the Covid-19 pandemic, many people lost their jobs, and therefore wages, causing depression and suicide to rise. Many studies on the relationship between health and wages focus on low paying jobs, leading to the second key parameter, education.

Education is another key variable that impacts health, because obtaining a higher education tends to lead to a better paying job, which improves access to healthcare. Education is tightly linked to wage and employment when it comes to its impact on health, both of which were discussed in the preceding paragraphs. Education is a strong predictor of good health in an individual. This is most likely to do with the link between education and wages, which is the strongest predictor of health.

The impact of race on health is widely discussed by the Center for Disease Control (CDC). The CDC identifies four major areas that reduce the health of minority communities, particularly black communities. These areas include access to healthcare, occupation, education level, and housing. With the onset of the Covid-19 pandemic, these four areas exacerbated the discrepancies in hospitalizations and deaths in areas where racial and ethnic minorities live and work. Strategies meant to stop/slow the spread of Covid-19 particularly hurt these communities, as the strategies directly led to a decrease in wages because their jobs were most impacted with lay-offs. As shown in the analysis, wages are the best predictor of health, so a loss in wages from the pandemic has a greater harm minority groups even more.

Model Fitting Procedure in R:

```
rm(list=ls())
require(ISLR)

attach(Wage)
require(gam)
```

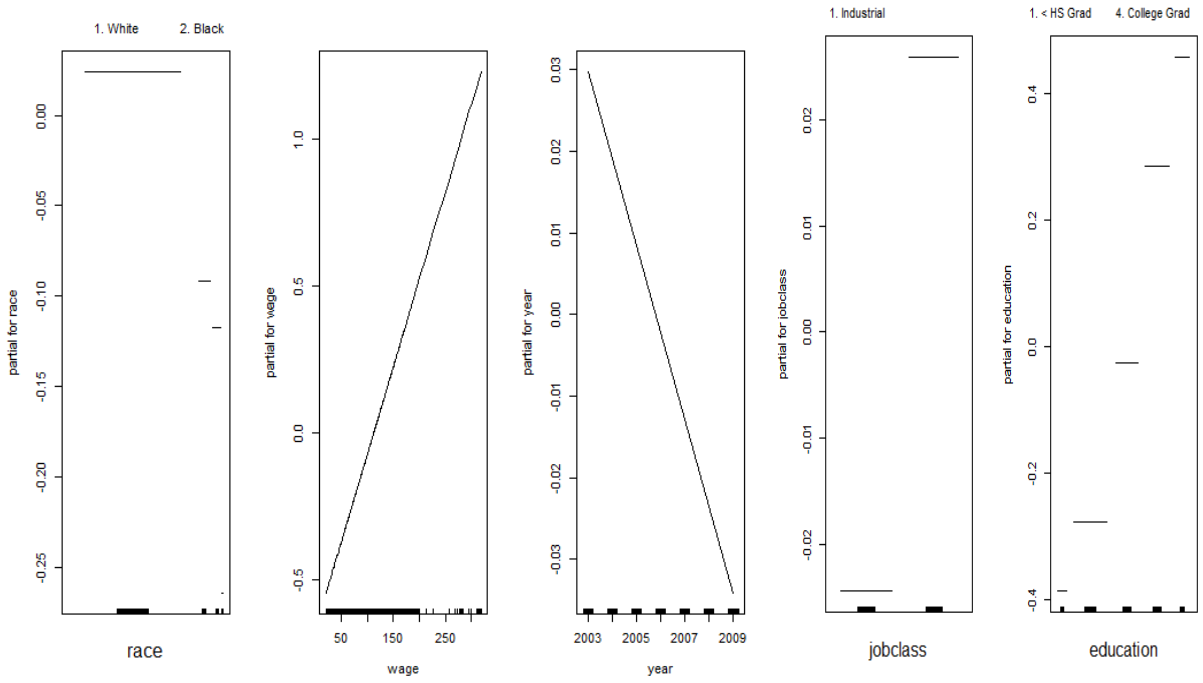
Basic GAM Model:

We started out with a basic GAM model, used to answer the research question:

"What socioeconomic factors have a significant impact on health?"

```
gam2<-gam(health~race+wage+year+jobclass+education,data=Wage,family=binomial)

par(mfrow=c(1,3))
plot(gam2)
```



The summary output shows that there is significance given their low p-values among wage, jobclass, and education in their effects on health.

```
summary(gam2)

## Anova for Parametric Effects
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## race         3    4.69   1.562  1.5533   0.19871
## wage         1   64.79  64.790 64.4362 1.419e-15 ***
## year         1    0.49   0.491  0.4883   0.48475
## jobclass     1    3.93   3.931  3.9095   0.04811 *
```

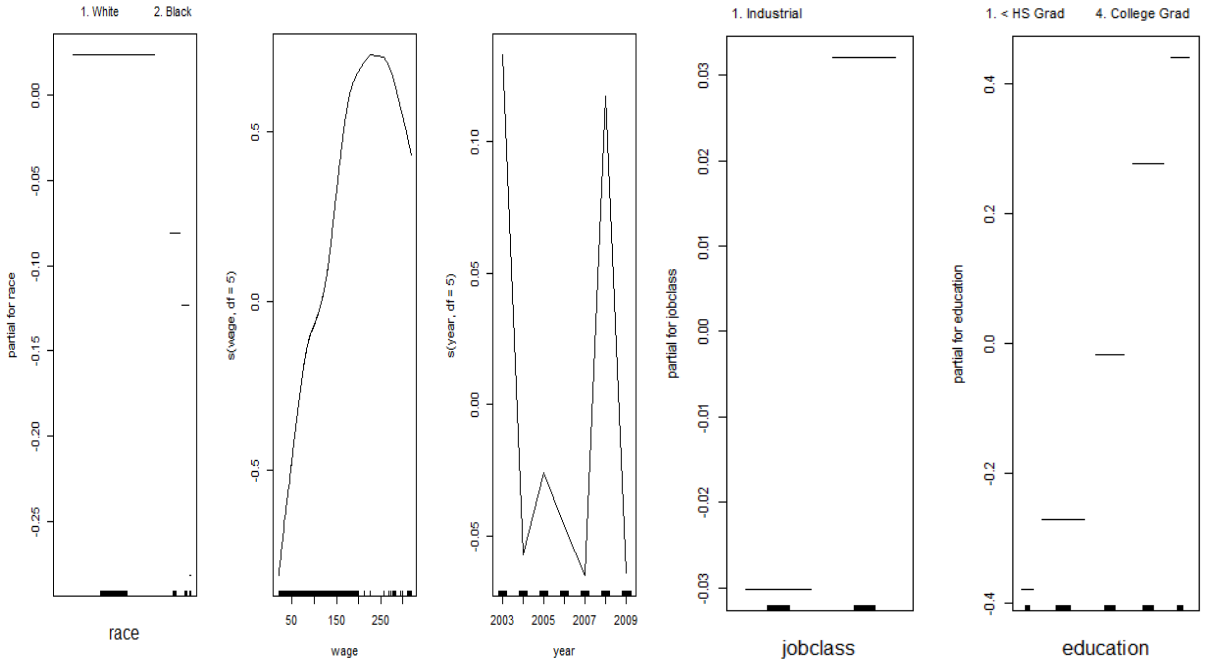
```
## education    4   34.79   8.698  8.6506 6.124e-07 ***
## Residuals 2989 3005.41   1.005                     ## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Categorical variables such as race, jobclass, and education, were added onto the GAM model using the binomial family. From there, we changed various factors such as degrees of freedom, or adding polynomial and logistic regressions, to attempt to better fit the model.

```
#smooth out wage and year
gam3<-gam(health~race+s(wage,df=5)+s(year,df=5)+jobclass+education,data=Wage,
family=binomial)
```

Smoothing splines, denoted by s() around certain factors, were also used to prevent large tail movement behavior in the case of polynomial regression.

```
par(mfrow=c(1,3))
plot(gam3)
```



Parametric vs Non-Parametric Terms:

When printing the summary, there are two categories: parametric vs non-parametric terms. Parametric terms typically have some sort of interval between data that has meaning, such as age and income(wage). Non-parametric terms on the other hand are nominal or categorical.

ANOVA can only be run on the nonparametric terms, and smoothing wage and year allows this:

The test returns a p-value for a Chi-squared independence test, indicating whether each individual factor is related to, or has an impact on health. Setting an alpha level (that p must

be less than) of 5% or 0.05, we see that no amount of changing the degrees of freedom (Df) improves the test, as the P(Chi) remains above 0.05 by a significant amount.

```
summary(gam3)
```

```
## Anova for Parametric Effects
##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## race               3    4.31   1.437  1.4288   0.23237
## s(wage, df = 5)    1   66.01  66.015 65.6326 7.841e-16 ***
## s(year, df = 5)    1    0.57   0.572  0.5690   0.45071
## jobclass           1    4.41   4.412  4.3863   0.03631 *
## education          4   33.21   8.303  8.2549 1.283e-06 ***
## Residuals       2981 2998.36   1.006
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                 Npar Df Npar Chisq P(Chi)
## (Intercept)
## race
## s(wage, df = 5)       4     4.9626 0.2912
## s(year, df = 5)       4     5.2878 0.2590
## jobclass
## education
```

```
#changes to df have only a worse effect on p(Chi)
gam5<-gam(health~race+s(wage,df=20)+s(year,df=20)+jobclass+education,data=Wag
e,family=binomial)
summary(gam5)
```

```
## Anova for Nonparametric Effects
##                  Npar Df Npar Chisq P(Chi)
## (Intercept)
## race
## s(wage, df = 20)      19    20.1515 0.3855
## s(year, df = 20)       5     5.7219 0.3342
## jobclass
## education
```

As a result, we decide to only smooth the year variable, and not the wage variable due to its p-value significance in the first output:

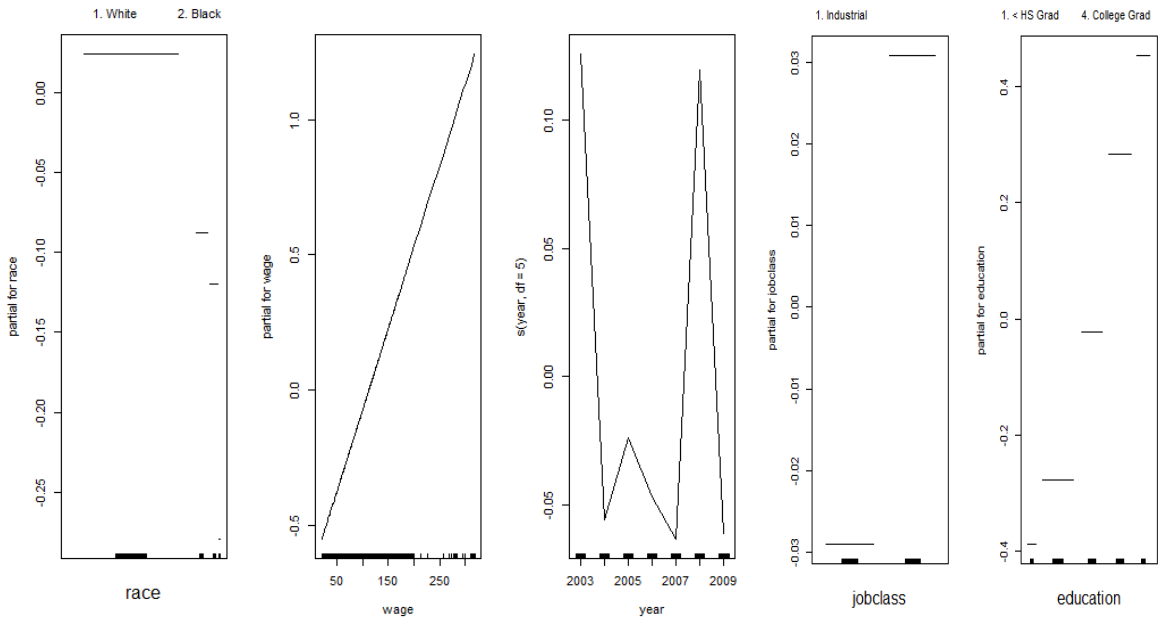
```
gam4<-gam(health~race+wage+s(year,df=5)+jobclass+education,data=Wage,family=b
inomial)
summary(gam4)
```

```
## Anova for Parametric Effects
##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## race               3    4.68   1.559  1.5484   0.19994
## wage               1   65.68  65.680 65.2261 9.590e-16 ***
## s(year, df = 5)    1    0.50   0.496  0.4927   0.48278
```

```
## jobclass            1    4.38   4.381  4.3510   0.03707 *
## education           4   34.48   8.620  8.5604 7.251e-07 ***
## Residuals        2985 3005.76   1.007
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                  Npar Df Npar Chisq P(Chi)
## (Intercept)
## race
## wage
## s(year, df = 5)       4     5.1366 0.2736
## jobclass
## education
```

And the final plots:

```
par(mfrow=c(1,3))
plot(gam4)
```



Conclusion

As seen from the ANOVA, wage, jobclass, and education are all significant predictors of a worker's health, as their p-values all fall below the 5% alpha levels. Surprisingly, the race variable is insignificant, with p-values remaining around 0.2. This is particularly intriguing given the studies performed and social expectations that race would have a large effect on healthcare access.