BST 222 Survival Analysis, Fall 2023

Final Project Assignment
Department of Public Health Sciences
RESEARCH

The association between history of low hourly wage and all-cause mortality among middle-aged workers

Xiaowei Zeng[2,*], Kuang Li[1,*] and Mingqian Zhang[1,*]

[1]University of California, Davis and [2]Fudan University

[*]xwzeng@ucdavis.edu; xxxx@ucdavis.edu; xxxx@ucdavis.edu

Abstract

Background: Low-wage earning constitutes an important and growing public health concern. However, wages are relatively underexamined as a social determinant of health. The purpose of this study is to examine whether a history of sustained low-wage earning is associated with all-cause mortality among middle-aged (35-50) workers.

Methods: We conducted an analysis based on the data from Health and Retirement Study (HRS). 3460 US participants aged 35-50 are included in the study. Initially, nonparametric Kaplan-Meier estimator was employed to estimate survival curves for exploratory analysis. The main statistical methods used in this study were Cox proportional hazards regression model, accelerated failure time model and Aalen additive hazards regression model, and thus we conducted the analysis on both multiplicative and additive scale. For multiplicative models, model selection techniques such as Lasso and stepwise were employed on the full model. For additive models, we fit a semiparametric model to examine both the constant effects and time-variant effects. Lastly, a series of model diagnostics were performed and some interesting findings were revealed.

Conclusions: Sustained low hourly wage earning is significantly associated with elevated mortality risk and excess deaths. If causal, our findings suggest that social and economic policies that improve the financial standing of low-wage workers (eg, minimum wage laws) could improve mortality outcomes.

Key words: all-cause mortality; low wage; cox proportional hazards; accelerated failure time; aalen additive hazards

Introduction

Prior research has established a robust link between socioeconomic status (SES) and mortality, showing that as socioeconomic indicators such as income and wealth decrease, mortality risk increases and life expectancy shortens. However, wages are relatively underexamined as a social determinant of health despite being an important component of SES [1].

Sustained low wages can be intricately linked to a range of health problems and even increased mortality rates due to the multifaceted impact of financial insecurity on various aspects of individuals' lives, including limited access to healthcare, increased stress levels, limited educational opportunities, occupational hazards, reduced access to social support and housing insecurity. These interconnected factors can result in the progression of preventable illnesses, contributing to poorer health outcomes. Therefore, a quantitative analysis for examining whether low wage exactly affect the risk of all-cause mortality is crucial for developing effective interventions and policies aimed at breaking the cycle of poverty and improving overall health outcomes.

The reason why we mainly focused on wages in this study is that it extends the literature on SES and mortality in 2 important ways. First, wages possess unique characteristics as a social determinant of health because they capture aspects of both income and occupation that may impact health in distinct ways. Second, wages can be directedly controlled by the existed policy and thus is modifiable and actionable for the govenment to potentially improve people's health and life. The low-wage workers are more likely to enjoy the positive health outcomes brought by policies on wages.

Hence, this article aims to examine whether the sustained low hourly wage history is associated with all-cause mortality among middle-aged (30-50) workers. This article proceeds as follows. We first introduce the dataset and encoding methods used in this study and conduct some exploratory analysis using Kaplan-Meier (KM) estimator in Section 2. Section 3 delineates the procedure of fitting Cox proportional hazards (PH) models, along with the interpretation of the results. The outcomes of accelerated failure time (AFT) model are outlined in Section 4. Details of the construction and outcomes of Aalen additive hazards models is presented in Section 5. We conclude the findings, limitations and suggestions in Section 6. The significance level in this paper is set to 0.05.

Dataset

The Health and Retirement Study (HRS) is a longitudinal household survey conducted by the Institute for Social Research at the University of Michigan with funding from the National Institute

Compiled on: December 11, 2023.

Table 1. Recoded HRS (1992-2020) dataset.

Var.	Definition
hhidpn	participant id
event	death time
status	1=death, 0=censored
wghist	low wage history (sustained, intermittent, never)
wave_num	wave number
prop_pt	Proportion of time in part-time employment
employ	employment stability
age	age (in 2020)
race	Hispanic, NH white, NH black, NH other
gender	male, female
edu	personal education years
pedu	maximum parental education years
religion	belief (none, catholic, jewish, protestant, other)
live	living area (northeast, midwest, west, south, move)
marriage	married or not married
wealth	household wealth / savings
insur_gov	whether covered by government insurance
insur_com	with or without employer's insurance
bmi	BMI score
self_health	self-reported health
smoke	never or ever smoke

on Aging and the Social Security Administration. It contains variables on demographics, health, health insurance, Social Security, pensions, family structure, retirement plans, expectations, and employment history, as well as imputations for income, assets, and medical expenditures developed at RAND. The RAND HRS Longitudinal File 2020 (V1) includes 15 waves of Core Interview data (biennially 1992-2020). All variables are named and derived consistently across survey years, and any cross-wave differences are documented.

From 17,013 variables contained in the HRS (1992-2020V1) data, we selected 192 variables that are potentially related to all-cause mortality. Some of them were recorded each wave and had 15 different values in total, while the others were demographic covariates that would not change over time. We encoded them into 21 variables for later data analysis (Table 1). Then we selected 3460 US participants aged 30-50 reporting wage larger than 6 waves in 15 waves.

There are three common definition of low wage earning. The first computes the hourly wage rate for full-time, full-year work (ie, 2087 hours) for annual earnings, and low wage is defined as lower than the federal poverty line. The second uses the approach above but defined fulltime, full-year work as 35 hours per week for 50 weeks annually (ie, 1750 hours). The third was defined as an hourly wage lower than two-thirds of the federal median wage for the corresponding year, which was employed in this study. Besides, we standardized the hourly wage reported at each survey year to 2020 dolloars using the consumer price index (CPI) inflation adjustments. Workers' low-wage earning histories were categorized as never earning a low wage or as having intermittent or sustained low-wage earning.

Stratified by the history of low-wage earning, the data is summarized in Table 10. Sustained low-wage history is associated with more fluctuated employment (sometimes employed and sometimes unemployed), more females, fewer personal and parental education years, more people living in south areas, less wealth, less exposure to employer's insurance, higher BMI and lower self-reported health status.

For exploratory data analysis, we employed nonparametric Kaplan-Meier method to estimate survival curves (Figure 1), observing the association between death and each covariate. The key point is that KM is nonparametric estimators, and thus it only takes into account the event we are interested in, say all-cause death, and the covariate we stratify on, without regard to any other variable. Hence, the results interpreted here may be partial, but they provide some useful insights for the modeling in the next section.

Low-wage history. According to the result of the log-rank test (p-value < 0.001), the low-wage history exhibited a significant difference on their impacts on death hazards. Those never have low wage has the highest survival probability, intermittent next, and lastly those with sustained low-wage earning. Interestingly, the sustained level curve slightly crosses the other two curves between 17-22 years. This may be attributed to the small sample size of sustained low-wage group. There's almost no intersection of three curves, so we can assume that it satisfies the proportionality of hazards assumption.

Employment stability. Workers with stable employment had a significantly higher probability of survival, namely a lower risk of mortality than workers with fluctuated employment (log-rank test p-value < 0.001). However, there's an intersection between the two curves in the 22nd year, which means the employment stability variable may not satisfy the PH assumption.

Race. The p-value of the log-rank test is 0.002 < 0.05, indicating that there's a significant difference in some categories of race. From the KM plot, we can find that at nearly all time points, hispanic participants have the highest survival probability among all the other nonhispanic people.

Personal and parental education years. The p-values of the log-rank tests are both less than 0.001, which means higher education level (no matter it is personal or parental) will lead to lower death risk.

Smoke history. According to the result of the log-rank test (p-value < 0.001), those who never smoked had a significantly higher probability of survival, namely a lower risk of death for all time points than those who ever smoked.

Gender, marriage and self-reported health. The p-values of the log-rank tests are 0.8, 0.6, and 0.1 respectively for gender, marriage and self-reported health status, all larger than 0.05, indicating that no significant difference exists in each group.

Cox Proportional Hazards Model

The Cox proportional hazards regression model, namely the multiplicative model, is a semiparametric statistical model used for analyzing the relationship between the survival time of patients and several predictor variables [2].

There are two methods we select variable that are lasso and two way stepwise regression. Lasso regression could reduce the effect of multicollinearity and also prevent over-fitting; we also build two way stepwise regression as a baseline model for the comparison based on metrics of AIC.

Model Fitting

LASSO. Based on the LASSO selection, sustained low wage history, and long smoke history have positive association increasing the risk of hazard. Greater education experience, good self health condition, have insurance covered by government sometime and always have significant negative association. It also leave co-variate live in the model but none of the areas provide statistical contribution.

Two Way Stepwise Regression. The model also indicates sustained low wage history, and long smoke history have statistical meaning to increase the probability of hazard, but the model actually remove

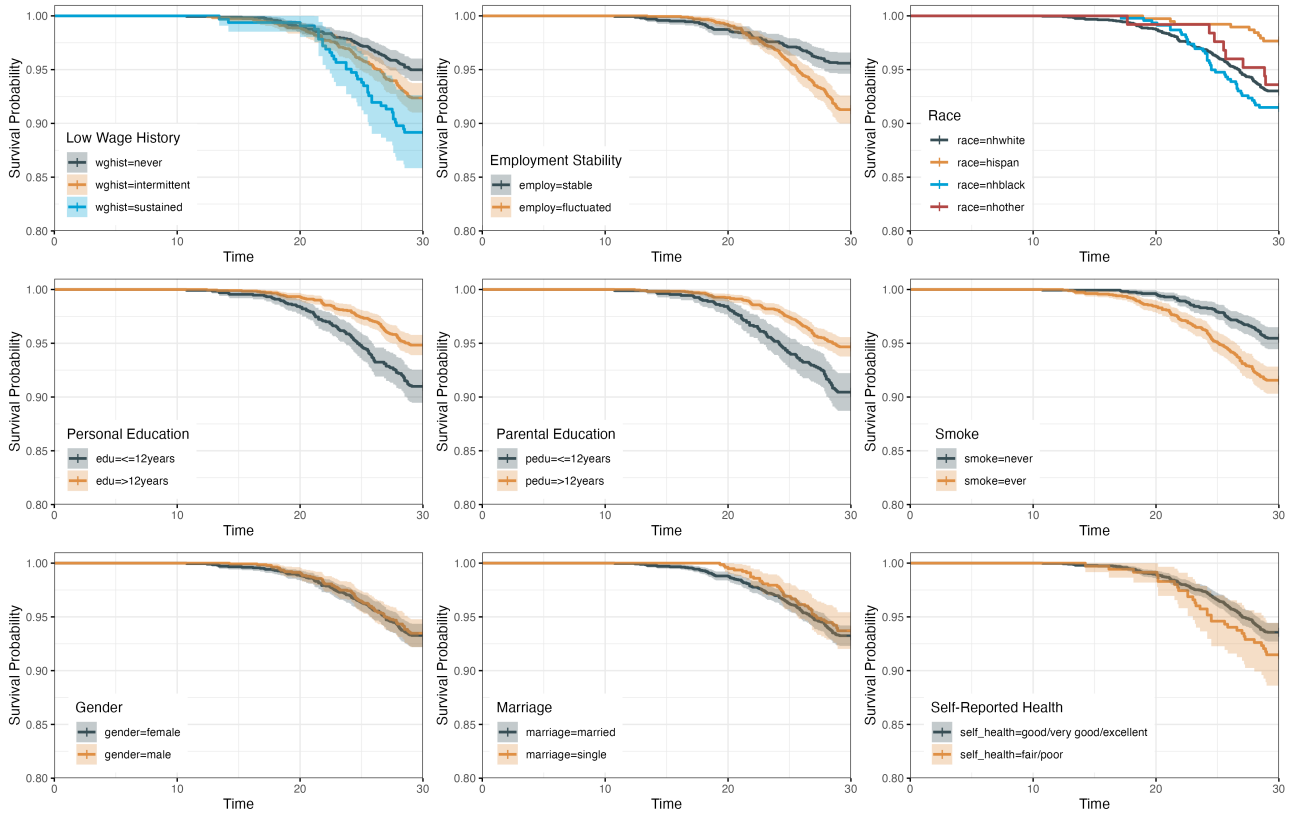

Figure 1. KM curves for the death event, considering low wage history, employment stability, race, gender, personal education, parental education, marriage, self-reported health and smoke history.

Table 2. Cox Proportional Hazard Regression using LASSO as feature selection.

Covariate	Estimate	SE	*p*
wghist: intermittent	1.19 (0.89, 1.60)	0.15	0.24
wghist: sustained	1.78 (1.15, 2.76)	0.22	0.00
age	1.23 (1.19, 1.28)	0.02	0.00
race: hispan	0.32 (0.16, 0.66)	0.36	0.00
race: nhblack	1.12 (0.77, 1.62)	0.19	0.55
race: nhother	1.02 (0.49, 2.14)	0.38	0.95
pedu: >12years	0.63 (0.48, 0.84)	0.14	0.00
live: midwest	1.04 (0.67, 1.60)	0.22	0.88
live: move	0.69 (0.36, 1.29)	0.32	0.24
live: south	1.05 (0.70, 1.57)	0.21	0.82
live: west	0.73 (0.43, 1.22)	0.26	0.23
insurance gov: always	0.09 (0.02, 0.37)	0.73	0.00
insurance gov: sometimes	0.20 (0.14, 0.28)	0.18	0.00
self_health: fair/poor	1.63 (1.09, 2.46)	0.21	0.02
smoke: ever	1.73 (1.31, 2.29)	0.14	0.00

Table 3. Cox Proportional Hazard Regression using stepwise as feature selection.

Covariate	Estimate	SE	*p*
wghist: intermittent	1.21 (0.90, 1.62)	0.15	0.20
wghist: sustained	1.86 (1.21, 2.87)	0.22	0.00
age	1.23 (1.19, 1.28)	0.02	0.00
race: hispan	0.29 (0.14, 0.58)	0.36	0.00
race: nhblack	1.16 (0.80, 1.68)	0.19	0.43
race: nhother	0.86 (0.42, 1.76)	0.37	0.68
pedu: >12years	0.62 (0.47, 0.82)	0.14	0.00
insurance gov: always	0.09 (0.02, 0.37)	0.73	0.00
insurance gov: sometimes	0.20 (0.14, 0.27)	0.18	0.00
self_health: fair/poor	1.62 (1.08, 2.43)	0.21	0.02
smoke: ever	1.70 (1.28, 2.24)	0.14	0.00

live covariate compared to LASSO, so LASSO selection is not strict compared to AIC selection in our case.

Model Diagnosis

The test assumption of stepwise cox model indicates that "GLOBAL" has p-value less than 0.05 which means at least one of the variables does violate the PH assumption; insurance coverage by government has a p-value < 0.001, indicating that it does not meet the PH assumption. It also reflects on Schoenfeld residual plot (Figure 6) that insurance coverage does have a clear downtrend and the KM curves. However, we did not choose to fit a stratified model because multiple level categorical is one of the consideration that could potentially causing complexity to train the cox regression model. Besides, KM estimator (Figure 6) shows that the intersection of the three curves occurs in the last two years, which is not a serious problem.

When it comes to Cox-Snell residual (Figure 2), it does maintain 45 degree line most of the time, but the straight does not exceed 95% confidence interval thought the time. For deviance and Martingale residual, the visualization becomes more likely a classification, the lower line moving into a down trend and couple outliers on the top and it might be multiple categorical data problem that resulted in complexity to the modeling.

The Dfbeta value indicates that there are several outliers for each variable but there is not a mutual outliers that all variables share in the Cox model (Figure 7).

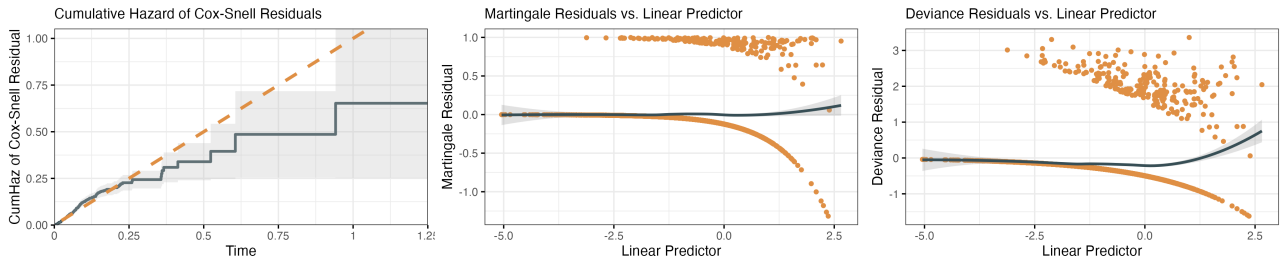

Figure 2. Cox-snell, martingale, deviance residuals for stepwise regression.

Table 4. AIC for Parametric Models

Name	AIC
Weibull	2669.582
Log-logistic	2658.936
Log-normal	2635.860
Exponential	2999.149

Table 5. Results of Weibull regression model.

Covariate	Value	SE	p
wghist: intermittent	-0.044	0.035	0.200
wghist: sustained	-0.145	0.052	0.005
age	-0.048	0.005	0.000
race: hispan	0.289	0.084	0.001
race: nhblack	-0.036	0.043	0.407
race: nhother	0.037	0.085	0.665
pedu: >12years	0.111	0.034	0.001
insurance gov: always	0.562	0.172	0.001
insurance gov: sometimes	0.383	0.047	0.000
self_health: fair/poor	-0.112	0.048	0.020
smoke: ever	-0.123	0.034	0.000

Parametric Model

Model Selection

We first compute AIC for 4 commonly used parametric models: Weibull, Log-logistic, Log-normal, Exponential (Table 4).

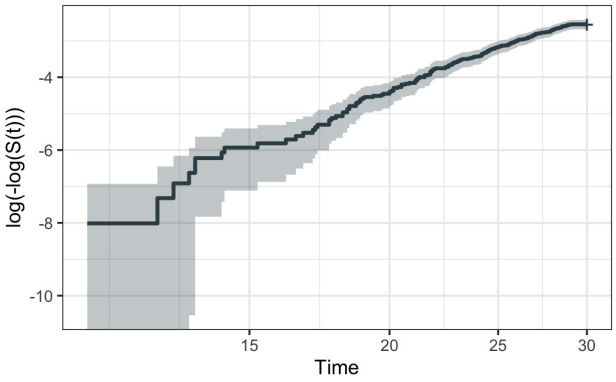

Figure 3. Complementary log-log survival plot.

The lowest AIC is for the log normal and log logistic, but the weibull is also viable. In this case, the choice probably does not matter much and might be left to other considerations. Note that it may be lower than the former models, but their scale can be different. So they shouldn't be compared.

As in the Complementary Log-Log Survival plot (Figure 3), the trend is linear, then a Weibull model is a good choice.

Weibull Model

The Weibull distribution is within the framework of the Accelerated Failure Time (AFT) model in survival analysis. The AFT model is an alternative to the Proportional Hazards (PH) model, and it directly models the effect of covariates on the survival time.

In an AFT model, the survival time of an individual is assumed to be accelerated (or decelerated) by a factor due to the covariates. The Weibull AFT model, in particular, assumes that the log-transformed survival times follow a Weibull distribution, which is equivalent to saying that the survival times themselves follow a log-Weibull distribution.

The general form of the Weibull AFT model can be written as:

$$\log(T) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \sigma\epsilon,$$

where T is the survival time, $X_1, X_2, \ldots, X_p$ are covariates, $\beta_0, \beta_1, \ldots, \beta_p$ are the regression coefficients, σ is the scale parameter, and ϵ is an error term following the standard Weibull distribution.

In this AFT model, the coefficients represent the log of the acceleration factor, where a positive coefficient indicates a decrease in survival time (acceleration of the event), and a negative coefficient indicates an increase in survival time (deceleration of the event). The Weibull distribution is particularly convenient for AFT models because it has a closed-form expression for the survival function and allows for varying hazard rates over time.

The Weibull regression model identified several significant predictors of survival time, we can see from the result below.

- Age (age): The coefficient suggests a negative relationship with the survival time, indicating that the risk of the event increases with age.
- Weight History (wghistintermittent, wghistsustained): Sustained low wage history was found to have a significant negative effect on survival time.
- Race/Ethnicity (racehispan): Being Hispanic showed a significant positive effect on survival time compared to the baseline race.
- Education (parent education>=12years): Higher parent education levels were associated with increased survival time.
- Insurance Status : Having government insurance always or sometimes was associated with increased survival time.
- Self-assessed Health (self health: fair/poor): A fair or poor self-assessed health status was significantly associated with decreased survival time.
- Smoking History (smoke ever): A history of smoking was associated with decreased survival time.

Table 6. Weibull convert results with 95% CI.

Covariate	Hazard ratio	LB	UB
wghist: intermittent	1.21	0.90	1.62
wghist: sustained	1.87	1.21	2.88
age	1.23	1.19	1.28
race: hispan	0.29	0.14	0.57
race: nhblack	1.17	0.81	1.69
race: nhother	0.85	0.42	1.75
pedu: >12years	0.62	0.47	0.82
insurance gov: always	0.09	0.02	0.37
insurance gov: sometimes	0.19	0.13	0.27
self_health: fair/poor	1.63	1.08	2.44
smoke: ever	1.70	1.28	2.25

The log-likelihood of the full model was notably higher than that of the intercept-only model, indicating a good fit of the predictors to the data.

The Estimated coefficients above in the original Weibull model are not clinically meaningful. That is why Weibull regression model is not widely used in medical literature. Since Weibull regression model allows for simultaneous description of treatment effect in terms of hazard ratio, is a measure of how the hazard of one group compares to the hazard of another group. and relative change in survival time. we can convert output from original models to more clinically relevant parameters.

Each value represents the hazard ratio for the corresponding variable or category (Table 6). A hazard ratio greater than 1 indicates an increased hazard (or risk) compared to the reference category, while a hazard ratio less than 1 indicates a decreased hazard.

- Age (age): An HR of 1.23 means that each additional unit increase in age is associated with a 23.1% increase in the hazard, assuming other variables are held constant.
- Race/Ethnicity (racehispan):Different race categories compared to a reference (not shown here) have different hazard ratios. For example, Hispanics have an HR of 0.29, indicating a substantially lower hazard compared to the reference race category.
- Insurance Status : people always get government insurance and sometimes have HRs significantly less than 1, indicating a lower hazard for these groups compared to the reference insurance category.

Residual Analysis for Weibull Model

Residuals concentration. The majority of the working residuals (Figure 4) are clustered around the horizontal axis near zero, which suggests that for many observations, the model's predictions are close to the observed values.

Outliers or leverage points. There are several points with large negative residuals, especially one very pronounced outlier near the bottom of the plot with a residual close to -80. This point, along with others that significantly deviate from zero, may be outliers or leverage points that could potentially influence the model fit.

Outliers in survival analysis (Table 7) can have various sources, such as data entry errors, exceptional cases that do not fit the model assumptions, or true individual variation. The presence of outliers can affect the model's predictive accuracy and the estimated effects of the covariates. In this context, outliers can indicate:

- Censoring Issues: Incorrectly recorded events can appear as outliers if they deviate greatly from the expected survival time.
- Influential Observations: These outliers may disproportionately influence the regression coefficients and the model's overall fit.
- Data Quality: They can point to potential issues with data quality that require verification or additional context.

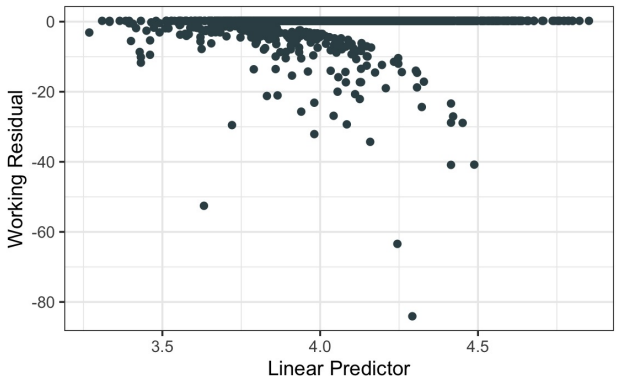

Figure 4. Working residuals vs. linear predictor in Weibull model.

Table 7. Charateristics of outliers.

Case	483	36
Event Time:	18.67 months	10.75 months
Status:	Event occurred	Event occurred
Low Wage History:	Never	Never
Age	68 years	74 years
Race	White	White
Gender	Male	Female
Parent Education	$\geq$ 12 years	$\leq$12 years
Government Insurance	Sometimes	Never
Self-assessed Health	Good to excellent	Good to excellent
Smoking History	Former smoker	Never smoked

Both identified outliers are event occurrences with relatively short event times and distinct characteristics in terms of demographics and health behaviors. Case 483 is a relatively older male with a good health rating and some government insurance coverage, while Case 36 is an older female without government insurance and no history of smoking. Their survival times are shorter than what might be expected given their relatively favorable health statuses and behaviors, suggesting that other unmeasured factors like accident or occurance of other deadly disease could be influencing their survival times.

All in all the Weibull parametric model provided a statistically significant fit for the survival data. The factors identified should be considered in policy-making and health interventions aimed at prolonging survival time in the studied population.

Aalen Additive Hazards Model

The Cox PH regression model assumes that the effects of the covariates are to act multiplicatively on an unknown baseline hazard function, and the risk coefficients are unknown constants whose value do not change over time. By contrast, Aalen additive hazards regression model assumes that the covariates act in an additive manner, and allows the unknown risk coefficients to be functions of time so that the effect of a covariate may vary over time [3]. One of the greastest advantages of this additive model is that it does not requires PH assumption, which is unlikely to be satisfied (especially for `insur_gov` variable).

We did not conduct variable selection in this section and simply used the previously selected variables in multiplicative models to fit the following additive models. The seven covariates include history of low hourly wage, age, race, parental education level, government

Table 8. Nonparametric (Additive Model 1) and semiparametric Aalen additive models (Additive Model 2): 300 simulation-based tests for non-significant effects and for time-invariant effects. The test for non-significant effects is the supremum test, and the two tests for time-invariant effects are Kolmogorov-Smirnov test and Cramer von Mises test.

	Additive Model 1						Additive Model 2					
	Test for insignificance		Tests for time-invariant effects				Test for insignificance		Tests for time-invariant effects			
Covariate	T_{Sup}	p	T_{KS}	p	T_{CvM}	p	T_{Sup}	p	T_{KS}	p	T_{CvM}	p
(Intercept)	9.06	0.00	0.43	0.00	2.97	0.00	9.12	0.00	0.45	0.00	3.30	0.00
wghist: intermittent	1.73	0.61	0.01	0.31	0.00	0.16	2.20	0.26	0.01	0.20	0.00	0.08
wghist: sustained	3.23	0.01	0.05	0.00	0.02	0.01	3.10	0.02	0.05	0.01	0.02	0.01
age	9.58	0.00	0.01	0.00	0.00	0.00	9.61	0.00	0.01	0.00	0.00	0.00
race: hispan	5.30	0.00	0.03	0.00	0.01	0.00						
race: nhblack	2.92	0.06	0.02	0.14	0.00	0.20						
race: nhother	3.12	0.04	0.03	0.30	0.00	0.54						
pedu: >12years	3.84	0.01	0.02	0.07	0.01	0.02						
insur_gov: always	5.72	0.00	0.06	0.00	0.05	0.00	5.85	0.00	0.05	0.00	0.05	0.00
insur_gov: sometimes	6.80	0.00	0.04	0.00	0.03	0.00	6.82	0.00	0.04	0.00	0.03	0.00
self_health: fair/poor	2.29	0.22	0.02	0.23	0.01	0.12						
smoke: ever	4.84	0.00	0.02	0.00	0.01	0.00	4.85	0.00	0.02	0.00	0.01	0.00

insurance, self-reported health status and smoke history.

Additive Model 1. We first constructed a nonparametric additive model assuming that all of them have time-varying effects. The race, parental education and self-reported health variable shows no significant time-varying effects (Table 8), so we can simplify the model by reducing the number of nonparametric component, say by wrapping both variables with `const` in the function `aalen`. The other covariates show strong evidence that they have time-variant effects in both Kolmogorov-Smirnov (KS) test and Cramer von Mises (CvM) test though some of them do not pass the supremum test. The significance of time-varying effects differ dramatically in different models, so the results of supremum test are not of great importance until we reach the final model.

Additive Model 2. A semiparametric Aalen additive hazards model is applied to the HRS dataset, where effects of race, parental education and self-reported health are assumed to be constant and the remaining covariate effects are allowed to be time varying. Results of this model are presented in Table 8 and the last 5 rows of Table 9.

This reduced semiparametric model gives a better fit to the dataset, as it is simpler in the interpretation and able to discriminate between constant and time-varying effects. Moreover, going from the nonparametric to the semiparametric additive model, comparison of Additive Model 1 and Additive Model 2 in Table 8 reveals that test statistics of the supremum, KS and CvM tests are almost unchanged except the intercept term.

Among the constant effects, only the hispanic level of race and parental education is significant (p-value < 0.05), indicating that compared to non-hispanic whites, hispanic workers are associated with 19.9 less deaths per 10,000 person-years; compared to workers whose parental education is more than 12 years, those with less than 12-year parental education have 13.0 excess deaths per 10,000 person-years. Though not significant, the results show that workers who reported a fair or poor health would have 11.1 excess death per 10,000 person-years.

For those variables with significant time-varying effects, it is clear to observe the patterns with time in Figure 5. The cumulative regression coefficients curve of sustained low-wage is nearly flat in the first 22 years, and then dramatically increase in the next 8 years. So we can conclude that the history of low hourly wage has less impact on death risks in the first 22 years because the curve is nearly flat, but then increases sharply. For the other covariates, on the contrary, the time-varying effects keep a continuous upward or downward trend. However, the intermittent low-wage history is not significant for time-varying effect, so I tried to adjust the low-wage history variable from time-varying effect to constant effect in Additive Model 3.

Table 9. Additive Model 3: 300 simulation-based estimates of constant effects.

Constant effects	Excess deaths[a]	Rb. se	p
wghist: intermittent	5.29 (-1.01, 11.60)	3.28	0.11
wghist: sustained	18.60 (5.02, 32.20)	6.66	0.01
race: hispan	-19.80 (-27.80, -11.80)	4.21	0.00
race: nhblack	5.53 (-4.60, 15.70)	5.68	0.33
race: nhother	-0.41 (-15.80, 15.00)	8.01	0.96
pedu: >12years	-13.00 (-20.90, -5.06)	4.20	0.00
self_health: fair/poor	11.10 (-0.58, 22.80)	5.99	0.06

[a] Excess infections per 10000 person-years.

Additive Model 3. A new semiparametric Aalen additive hazards model is applied to the HRS dataset, where effects of low-wage history, race, parental education and self-reported health are assumed to be constant and the remaining covariate effects are allowed to be time varying. Results of this model are presented in 9.

For the constant effects, the sustained level of low-wage history has a p-value of 0.005 < 0.05, indicating that workers with sustained low hourly wage have 0.19% higher estimated excess death rate than those always with high wage, which means they experienced 18.6 excess deaths per 10,000 person-years. Workers with intermittent low hourly wage is associated with 5.3 excess deaths per 10,000 person-years (though not significant, p-value = 0.1). The constant effects of race and parental education remain nearly the same with Additive Model 2.

Conclusions

All the models showed that sustained low-wage earning is significantly associated with elevated mortality risk and excess deaths. If it is causal, policies such as minimum wage laws could directly impact hourly wage, which could potentially improve health and health inequality according to the research.

However, there're some limitations of this research. First, the dataset do not include all the confounders that may have negative impact on the estimates, but we have done our best to control the potential confounders. Second, in the data cleaning process, we selected participants of 35-50 years old as middle-aged workers without theoretical support. This may lead to selection bias in the

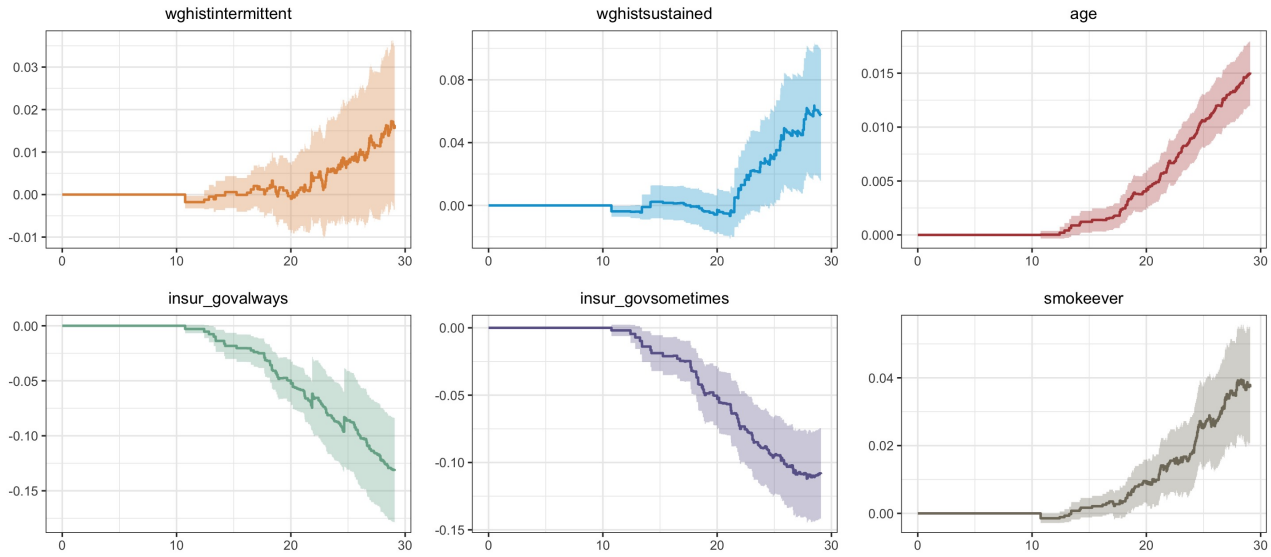

Figure 5. Estimated cumulative regression coefficients for variables with time-varying effects in Additive Model 2, together with 95% confidence intervals.

analysis. In further work, we can conduct sensitivity analysis by selecting participants with different age ranges. Another way to deal with the selection bias is to implement inverse probability weighting (IPW) strategy. IPW can adjust the sample weight and make the expectation of sample distribution the same as that of exact distribution.

Availability of Source Code and Requirements

- Project name: BST 222 Survival Analysis Final Project
- Project home page: https://github.com/xw-zeng/Survival-Analysis-2023Fall
- Programming language: Python 3.10, R 4.2.1
- License: The MIT License (MIT)

References

1. Kezios KL, Lu P, Calonico S, Al Hazzouri AZ. History of Low Hourly Wage and All-Cause Mortality Among Middle-aged Workers. JAMA 2023 02;329(7):561–573.
2. Cox DR. Regression Models and Life Tables. Journal of the Royal Statistic Society 1972;B(34):187–202. https://www.jstor.org/stable/2985181.
3. Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data. Statistics for Biology and Health, Springer New York; 2005. https://books.google.com/books?id=jS2Cy0lezJIC.

Table 10. Characteristics of participants by history of low hourly wage.

Characteristic	never, N = 1,695[a]	intermittent, N = 1,442[a]	sustained, N = 323[a]
wave_num	7 (6, 8)	7 (6, 9)	7 (6, 8)
prop_pt	0.00 (0.00, 0.08)	0.00 (0.00, 0.17)	0.13 (0.00, 0.33)
employ: stable	929 (55%)	610 (42%)	120 (37%)
employ: fluctuated	766 (45%)	832 (58%)	203 (63%)
age	68.0 (64.0, 73.0)	71.0 (67.0, 75.0)	70.0 (66.0, 74.5)
race: nhwhite	1,321 (78%)	1,015 (70%)	156 (48%)
race: hispan	126 (7.4%)	170 (12%)	89 (28%)
race: nhblack	179 (11%)	213 (15%)	66 (20%)
race: nhother	69 (4.1%)	44 (3.1%)	12 (3.7%)
gender: female	887 (52%)	929 (64%)	249 (77%)
gender: male	808 (48%)	513 (36%)	74 (23%)
edu: <=12years	452 (27%)	632 (44%)	248 (77%)
edu: >12years	1,243 (73%)	810 (56%)	75 (23%)
pedu: <=12years	376 (22%)	513 (36%)	190 (59%)
pedu: >12years	1,319 (78%)	929 (64%)	133 (41%)
religion: none	221 (13%)	132 (9.2%)	31 (9.6%)
religion: catholic	462 (27%)	388 (27%)	100 (31%)
religion: jewish	46 (2.7%)	15 (1.0%)	4 (1.2%)
religion: other	43 (2.5%)	30 (2.1%)	6 (1.9%)
religion: protestant	923 (54%)	877 (61%)	182 (56%)
live: northeast	301 (18%)	159 (11%)	22 (6.8%)
live: midwest	382 (23%)	347 (24%)	63 (20%)
live: move	119 (7.0%)	122 (8.5%)	19 (5.9%)
live: south	495 (29%)	544 (38%)	171 (53%)
live: west	398 (23%)	270 (19%)	48 (15%)
marriage: married	1,332 (79%)	1,121 (78%)	228 (71%)
marriage: single	363 (21%)	321 (22%)	95 (29%)
wealth	399,921 (162,188, 843,061)	215,019 (87,260, 508,485)	76,668 (28,928, 181,698)
insur_gov: never	610 (36%)	298 (21%)	61 (19%)
insur_gov: always	31 (1.8%)	28 (1.9%)	13 (4.0%)
insur_gov: sometimes	1,054 (62%)	1,116 (77%)	249 (77%)
insur_com: never	134 (7.9%)	263 (18%)	133 (41%)
insur_com: always	470 (28%)	107 (7.4%)	10 (3.1%)
insur_com: sometimes	1,091 (64%)	1,072 (74%)	180 (56%)
bmi	27.8 (24.7, 31.7)	28.1 (24.9, 31.9)	28.3 (25.2, 31.8)
self_health: good/very good/excellent	1,569 (93%)	1,298 (90%)	241 (75%)
self_health: fair/poor	126 (7.4%)	144 (10.0%)	82 (25%)
smoke: never	806 (48%)	635 (44%)	147 (46%)
smoke: ever	889 (52%)	807 (56%)	176 (54%)

[a] Median (IQR); n (%).

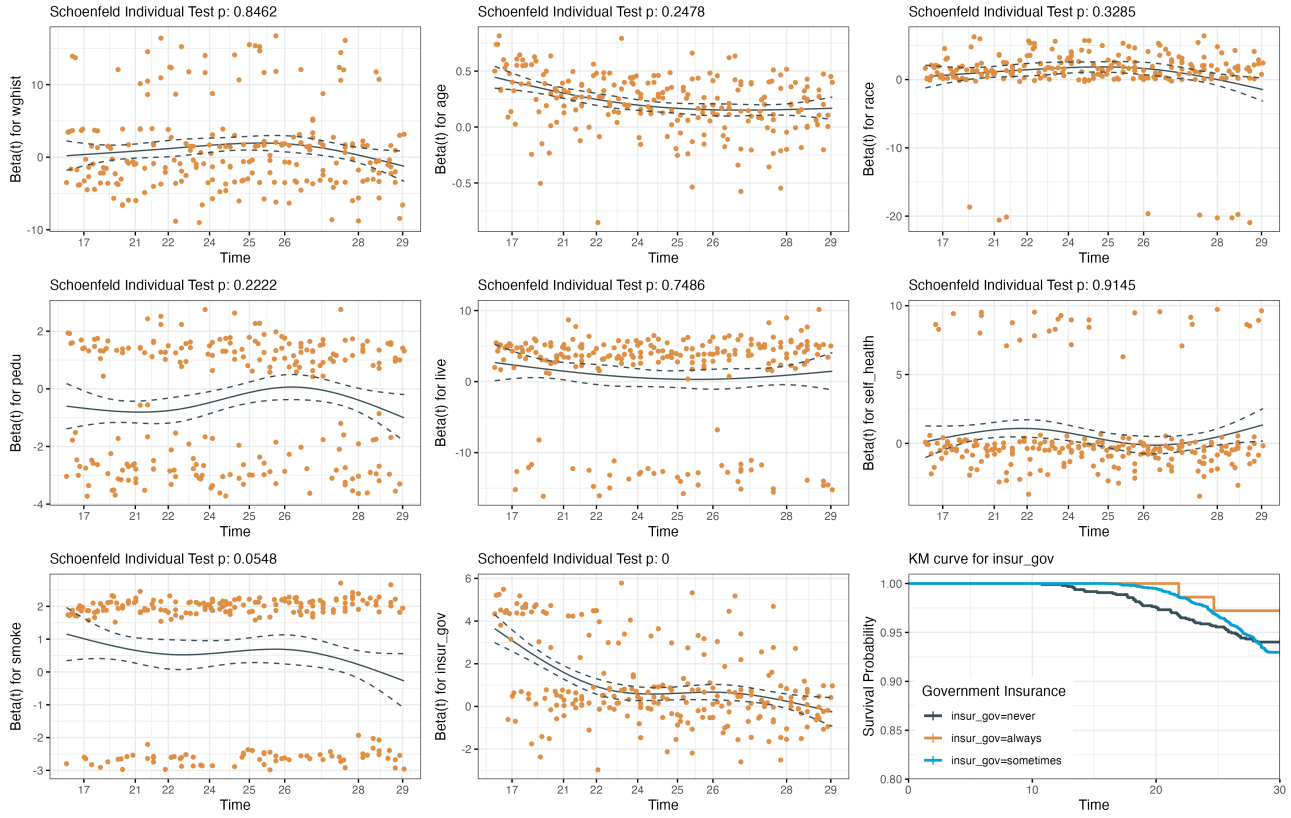

Figure 6. Schoenfled residuals for stepwise regression.

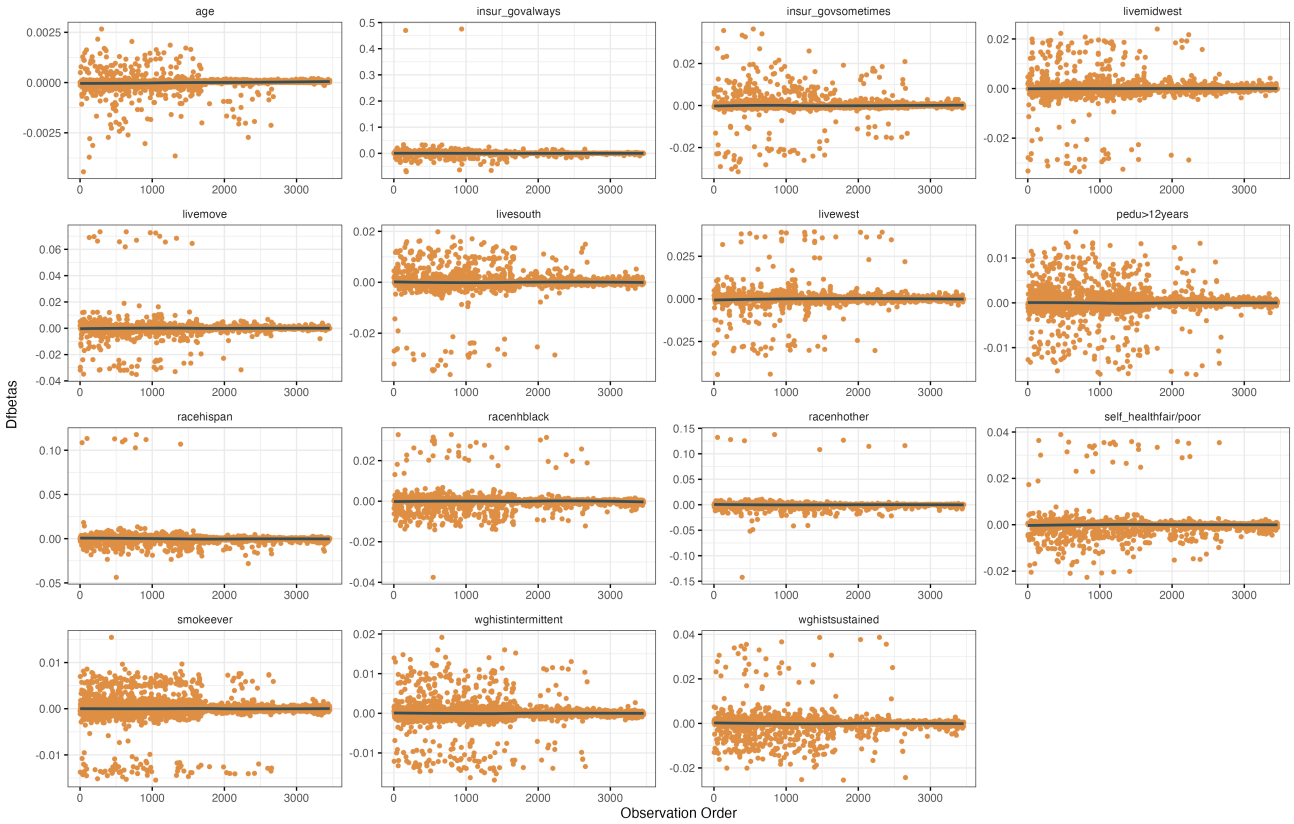

Figure 7. Dfbeta Value for Stepwise Regression.