이 학 석 사 학 위 논 문

A Cox Proportional Hazard Model with Time-Varying Coefficients Using Revised Survival Time to Investigate Stroke Risk Factors

뇌졸중 위험요인 분석을 위하여 개선된 생존시간을 이용하는 시간변동계수를 포함한 Cox 비례위험모형

2014년 2월

서울대학교 대학원

통계학과

장 지 영

A Cox Proportional Hazard Model with Time-Varying Coefficients Using Revised Survival Time to Investigate Stroke Risk Factors

뇌졸중 위험요인 분석을 위하여 개선된 생존시간을 이용하는 시간변동계수를 포함한 Cox 비례위험모형

지도교수 Myunghee Cho Paik

이 논문을 이학석사 학위논문으로 제출함
2013년 12월

서울대학교 대학원
통계학과
장 지 영

장지영의 이학석사 학위논문을 인준함
2013년 12월

위 원 장	이 영 조	(인)
부위원장	Myunghee Cho Paik	(인)
위 원	이 재 용	(인)

Abstract

A Cox Proportional Hazard Model with Time-Varying Coefficients Using Revised Survival Time to Investigate Stroke Risk Factors

Jiyeong Jang
The Department of Statistics
The Graduate School
Seoul National University

According to estimates by World Health Organization (WHO), globally, cardiovascular disease is the first leading cause of death. Among various cardiovascular diseases, the risk of stroke is constantly increasing even though the mortality rate for it is decreasing. Including representative studies, Framingham study and Northern Manhattan stroke study (NOMAS), many of studies relating to the risk factors of stroke have been performed. In this study, using the dataset obtained from NOMAS, the Cox proportional hazard model is used in order to investigate stroke risk factors. In particular, the NOMAS used follow-up time to onset of stroke, whereas, in this study, the age of having a stroke is treated as survival time for fitting the models. The proportional hazard assumption is evaluated through graphs, and some risk factors whose risk effects vary according to the variation of age are detected. The pattern of change and the significance of these risk effects are evaluated by constructing a Cox's model with time-varying coefficients. In addition to the study concerning occurrence of all types of strokes, the study concentrating only on the ischemic stroke occurrence is also performed and these two are compared.

Keywords: *stroke, Northern Manhattan stroke study, Cox proportional hazard model, survival time, proportional hazard assumption, time-varying coefficient*

Student Number: 2012-20231

Contents

List of Tables

List of Figures

Chapter 1

Introduction

A stroke occurs when blood supply to the brain is rapidly interrupted. This can be due to a hemorrhage or ischemia (i.e., lack of blood flow) caused by blockage such as thrombosis and arterial embolism (Sims and Muyderman, 2009). Stroke is one of the major threats to the public health in the world since it is a leading cause of disability and death. Appropriate approaches to predict individuals with high risk of vascular disease can help health care professionals to detect the people at risk people (Goldstein et al., 2006). Various risk factors for stroke have known such as old age, high blood pressure, diabetes, high cholesterol, smoking, etc.

Many tools have been devised to predict the risk of cardiovascular disease. The Framingham-based models have been used widely as a tool of risk prediction of stroke (D'Agostino et al., 1994), coronary heart disease (Wilson, 1998), and cardiovascular risk (D'Agostino, 2008). However, these models have limitation; they do not contains behavioral or anthropometric factors, and they constructed from a white ethnicity only (D'Agostino et al., 2001). In order to supplement these limitations, the Northern Manhattan Study (NOMAS) was performed. NOMAS is a research study of stroke and stroke risk factors in the Northern Manhattan community conducted at the

Neurological Institute, Columbia University, Division of Stroke and Critical Care. NOMAS is the first study of its kind to focus on stroke risk factors in whites, blacks, and Hispanics living in the same community. It is helping to fill gaps in our knowledge of stroke epidemiology in minority populations. This study includes population of Washington Heights in Northern Manhattan. The ongoing study, which began in 1990, is based in the Neurological Institute of Columbia Presbyterian Hospital, located in Washington Heights. The goal of NOMAS is to improve currently available global cardiovascular disease risk prediction tools by additionally using behavioral risk factors and racially diverse cohort for cardiovascular risk prediction of African-American and Hispanic people (Sacco et al., 2009). A survival model was built to predict combined cardiovascular outcomes. There are three main outcomes of interest, stroke (ischemic or intracerebral hemorrhage), myocardial infarction (MI), and vascular death. When constructing a survival model, the NOMAS used survival time, follow-up time to stroke onset. However, this approach is less meaningful since the ages when subjects enrolled were different one another. Therefore, regardless of subjects' ages when cardiovascular diseases occur, follow-up time indicates just the duration from subjects' enrollment to occurrence of diseases.

In this thesis, we try to replace the survival time based on follow-up time with the age when a stroke occurs. This approach can provide estimates indicating how risk factors affect the age of having a stroke. The age group can be also detected in which the risk effect of each covariate is dominant and we try to find time-varying effect of risk factors through the log-log Kaplan-Meier, and cumulative coefficients plots from extended version of Cox's model. In addition, these plots are also drawn using the models with ischemic stroke as outcome of interest.

Chapter 2

Statistical Methods

The NOMAS trials is a prospective, population-based cohort of 3,298 subjects recruited between 1993 and 2001. The detail explanation of NOMAS cohort selection and follow-up process was described in Sacco et al. (1997).

Censoring and Truncation

Standard statistical methods are not appropriate because survival time data is typically incompletely observed. Most common type of incompleteness is right censoring. Right censoring occurs when a subject leaves the study prior to an event occurrence, or an individual do not experience the event of interest before the end of study. Truncation is the other main cause to make incompleteness of survival time data. Truncation is occurred due to study design. Left truncation occurs when a subject has been at risk before enrollment in study. Under the left truncated right censored data, for each i-th individual in study, $i = 1,2, \dots, n$, X_i denotes the time between an origin and event, and C_r indicates the fixed censoring time. Each of X_i is assumed to be identically distributed with probability density function $f(x)$ and survival function $S(x)$. The data $\{(T_i, \delta_i)\}_{i=1,2,\dots,n}$ are observed only when $T_i > L_i$ where L_i

is left truncation value, $T_i = \min(X_i, C_r)$ and δ_i is *failure* or *uncensoring* indicator, $I(X_i \leq C_r)$. For the inference with left-truncated data, T_i and L_i are assumed to be independent only in the region in the data observed. This assumption is weaker than the independent assumption and called quasi-independent truncation (Tsai, 1990).

In the NOMAS data, right censoring occurs. In addition, In this study, when survival time calculated based on a subject's follow-up time is replaced by a subject's age of having a stroke occur, left truncation occurs. This is because only a subject whose stroke event time exceeds the age at enrollment in study can be observable. Thus, when age of having a stroke is used as survival time, each subject's age at enrollment in study is regarded as a truncation value, L_i and the risk set is corrected at each point where event occurs. The quasi-independent truncation is assumed to be satisfied in this study.

Kaplan-Meier curves

The Kaplan-Meier estimator (Kaplan and Meier, 1958), also known as the product limit estimator is used to estimate the survival function from survival data. If the uncensored observation time points are distinct and $t_{(1)}$, $t_{(2)}$, …, $t_{(D)}$ denote the ordered failure times, the Kaplan-Meier estimator and its estimated variance based on Greenwood formula are as follows.

$$\hat{S}(t) = \prod_{i:t_{(i)} \leq t} \left\{1 - \frac{d_i}{Y_i}\right\}$$

$$\widehat{Var}\left[\hat{S}(t)\right] = \left[\hat{S}(t)\right]^2 \hat{\sigma}_S{}^2(t) = \left[\hat{S}(t)\right]^2 \prod_{i:t_{(i)} \leq t} \frac{d_i}{Y_i\{Y_i - d_i\}}$$

where Y_i and d_i denote the number of risks and the number of deaths at time $t_{(i)}$, respectively.

The values of confidence interval were derived from the estimator $\hat{S}(t)exp\left\{\pm z_{\alpha/2}\hat{\sigma}_s(t)\right\}$, the default confidence interval values of `survfit` function in R statistical software. Kaplan-Meier curves for survival free of stoke are plotted for three cases when: 1) survival time based on follow-up time is used; 2) survival time based on age of having a stroke is used without correction for left truncation; 3) survival time based on age of having a stroke with correction for left truncation.

Cox proportional hazard model

In order to investigate the effect of explanatory variables on survival time X, a Cox proportional hazard (PH) model (Cox, 1972) is used. For each i-th individual, explanatory variables $\boldsymbol{z_i}$ is observed, where $\boldsymbol{z_i} = (z_{i1}, z_{i2}, \dots, z_{ip})^T$ is a p-dimensional fixed covariate. The Cox proportional hazard model is written as

$$h_i(t; \boldsymbol{z_i}) = Y_i(t)h_0(t)\exp(\beta^T \boldsymbol{z_i})$$

where $Y_i(t)$ is the at risk indicator represented as $I(T_i > t)$, $h_i(t; \boldsymbol{z_i})$ is hazard function of observation i at time t and $h_0(t)$ is base line hazard function (the hazard function for an individual with $\boldsymbol{z_i}$ is zero).

Most of statistical software packages estimate β by maximizing the log partial likelihood function (Cox, 1975).

$$LL(\beta) = \sum_{i=1}^{D} \delta_i \left[\beta^T \boldsymbol{z_{(i)}} - \boldsymbol{bg} \left(\sum_{I \in R_i} \exp(\beta^T \boldsymbol{z_{(i)}}) \right) \right]$$

where $\boldsymbol{z_{(i)}}$ denotes the covariate vector of i-th observation at ordered failure time $t_{(i)}$ and R_i represents the risk set at time $t_{(i)}$.

The base line cumulative hazard is $H_0(t) = \int_0^t h_0(u)du$ and the Breslow (1972) estimator

$$\widehat{H}_0(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n Y_i(u)\exp(\widehat{\beta}^T \boldsymbol{z}_i)}$$

is generally used in many cases in practice.

The Cox PH model with survival time based on subjects' age of having a stroke and follow-up time for a stroke are constructed using the variables 'STROKE_TIME' (calculated follow-up time for all stroke as an outcome) and 'INSTORKE' (censoring indicator for all stroke). These models are also built regarding ischemic stroke using 'ISCHEMIC_TIME' (calculated follow-up time for ischemic as an outcome) and 'ISHCEMIC' (censoring indicator for ischemic stroke).

As explanatory variables, the risk factors of the final model in NOMAS are considered. The variables in NOMAS data set, 'WAIST (waist circumference)', 'etmod' (moderate alcohol consumption), 'actmodheavy' (moderate-to-heavy physical activity), 'BLACK' (black race), 'HISP' (Hispanic race), 'SYSOTLIC' (systolic blood pressure), 'SMOKER1' (former smoking), 'SMOKER2' (current smoking), 'LLDL' (LDL cholesterol level) and 'LHDL' (HDL cholesterol level) are used as the same ones. 'Peripheral vascular disease' and 'Fasting blood sugar' variables are replaced by 'NCAD' (any cardiac disease) and 'DIABETES' (diabetes) in the given data set, respectively. 'Diastolic blood pressure' and 'Antihypertensive medication' variables are excluded since they are not available. A 'systolic' indicator variable is made using the criteria of metabolic syndrome for certifying the relationship of two risk factors, systolic blood pressure and Hispanic race. If systolic blood pressure is higher than 130 mmHg, 'systolic' indicator is 1 and 0 for otherwise. Using the likelihood ratio test criteria, variables which contribute to the fit are kept in the model.

Evaluating proportionality assumption

One of the assumptions of Cox proportional hazard model is proportional hazard assumption. The survival function can be represented as

$$\ln[-\ln S(t;\boldsymbol{z})] = \beta^T \boldsymbol{z} + \ln[-\ln S_0(t)]$$

where $S(t;\boldsymbol{z})$ is survival function for X and $S_0(t)$ is survival function when $\boldsymbol{z}$ is zero.

Thus, the logarithm of the negative logarithm of the survival functions are parallel, given different covariates $\boldsymbol{z_i}$. In order to evaluate the proportional hazard assumption of Cox proportional hazard model, log-log Kaplan-Meier curves are plotted for each binary covariate.

Cox's models with time-varying coefficients

The proportional hazard assumption of Cox's model is not satisfied when the relative risk of a covariate is not constant with time. In this case, the extension of Cox proportional hazard model is needed to reflect the time-varying effects of some of the covariates. The basic extension of Cox's model to accommodate time-varying coefficients is

$$h_i(t;\boldsymbol{z_i}) = Y_i(t) h_0(t) \exp(\beta^T(t)\boldsymbol{z_i}(t))$$

where $\beta^T(t) = (\beta_1(t), \beta_2(t), \dots, \beta_p(t))$ is a $p \times 1$ vector of time-varying regression coefficients.

This model has been studied by many authors, e.g., Zucker & Karr (1990), Murphy & Sen (1991), Grambsch & Therneau (1994), Pons (2000), Martinussen et al. (2002), Cai & Sun (2003) and Sinnett & Sasieni (2003) (Martinussen and Scheike, 2006). This fully nonparametric version of the extended Cox's model is not easy to fit due to the model high flexibility, for the small to medium sized data. The time-varying

assumption for all covariates may not be needed in some situation. In this case, the semiparametric version of extension model

$$h_i(t;\boldsymbol{z_i}) = Y_i(t)h_0(t)\exp(\beta^T(t)\boldsymbol{z_i}(t) + \gamma^T\boldsymbol{w_i}(t))$$

where $(\boldsymbol{z_i}(t), \boldsymbol{w_i}(t))$ is a $(p+q)$-dimensional covariate and $\beta(t)$ is nonparametric p-dimensional parameter and γ is q-dimensional regression parameter of the model.

This model is suggested to compromise between model flexibility and size of the data. This model has been studied by Martinussen et al. (2002) and Scheike & Martinussen (2004) (Martinussen and Scheike, 2006). For the covariates whose log-log Kaplan-Meier curves are not parallel, the semiparametric version of the model is considered to investigate time-varying effects of covariates. In order to fit the model, the `timecox` function of a `timereg` package in R statistical software is used. This turns out the cumulative regression function

$$B(t) = \int_0^t \beta(s)ds$$

of each covariate in the semiparametric version of Cox's models (Martinussen and Scheike, 2006).

Chapter 3

Results

3.1. All Types of Strokes

The survival function of follow-up time to all types of strokes onset is drawn by the Kaplan-Meier estimate (Fig. 3.1). The survival probability corresponds to the last stroke occurrence is 0.845. The Kaplan-Meier curves are plotted when follow-up time is replaced by the age of having all kinds of strokes (Fig. 3.2). The red line represents the estimates of survival function without the correction of left truncation and the blue one is drawn under the correction of left truncation. When left truncation is corrected, the Kaplan-Meier estimates are lower compared to the case when there is no correction of left truncation. This is natural since the number of people at risk at each stroke event time is reduced under the left truncation. The last stroke occurs when the subject's age is 103.5. At this age, the survival probabilities are 0.615 and 0.50 in two Kaplan-Meier curves, respectively. The first stroke occurs when the subject's age is 42.16. While drawing a Kaplan Meier curve with the correction of left truncation, this subject having earliest stroke is excluded. At this time of stroke occurrence, the number of people at

risk is 16. This small risk set makes the standard error of Kaplan-Meier estimator is significantly large. After excluding this individual, remain subjects experienced strokes after age 50 years.

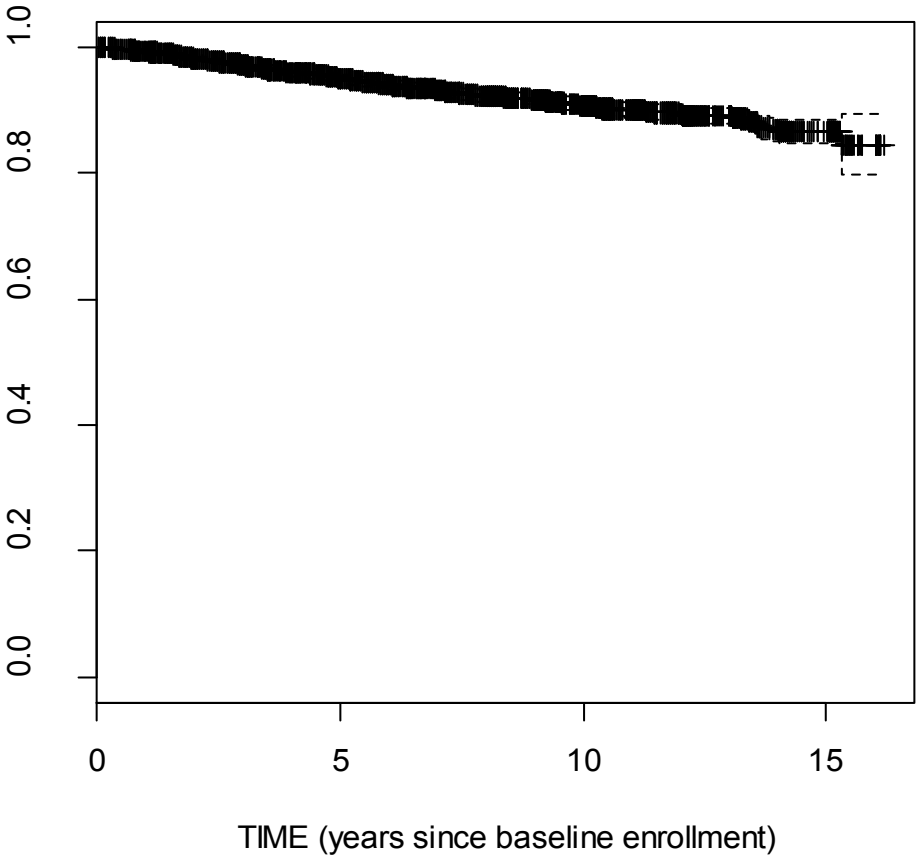


Figure 3.1 Kaplan-Meier estimate of the survival distribution function of follow-up time to all types of strokes occurrence

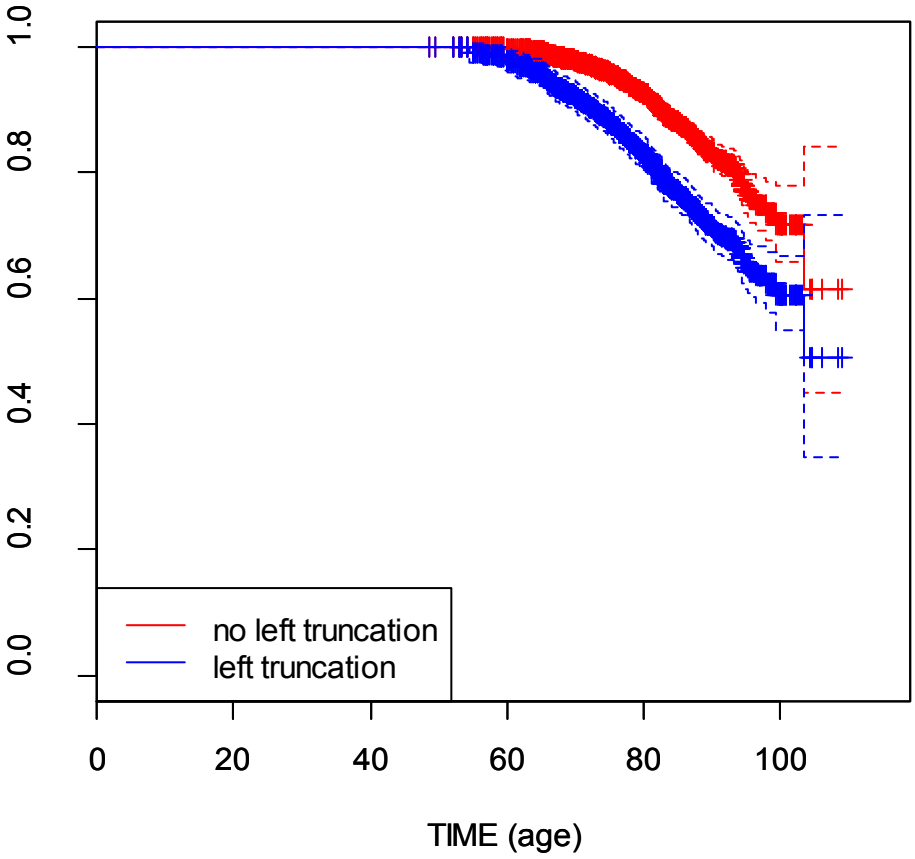


Figure 3.2 Kaplan-Meier estimate of survival distribution function of age at all types of strokes onset

Table 3.1 describes the parameter estimates, standard error of parameter estimates and p value of risk factors of the final Cox proportional hazard model when age at all kinks of strokes onset is used as survival time. Table 3.2 represents the same items when follow-up time to all types of strokes are used.

Table 3.1 Cox proportional hazard model with survival time, age of having all types of strokes

Risk Factor Variables	Parameter Estimate	Standard. Error	p-value
Moderate alcohol consumption	-0.282	0.143	0.0493
Moderate-to-heavy physical activity	-0.211	0.246	0.3916
Any cardiac disease	0.399	0.132	0.0026
Black race	-0.357	1.302	0.7839
Hispanic race	-2.268	1.188	0.0563
Male sex	0.448	0.139	0.0013
Systolic blood pressure (mmHg)	0.001	0.007	0.8395
Diabetes	0.814	0.130	0.0000
Former smoking	0.146	0.143	0.3062
Current smoking	0.478	0.176	0.0067
HDL cholesterol (mg/dl)	0.003	0.004	0.3990
LDL cholesterol (mg/dl)	-0.002	0.001	0.0896
Systolic blood pressure (mmHg) × Black race	0.004	0.008	0.6379
Systolic blood pressure (mmHg) × Hispanic race	0.016	0.008	0.0453

Table 3.2 Cox proportional hazard model with survival time, follow-up time to all types of strokes occurrence

Risk Factor Variables	Parameter Estimate	Standard. Error	p-value
Moderate alcohol consumption	-0.288	0.143	0.045
Moderate-to-heavy physical activity	-0.191	0.246	0.437
Any cardiac disease	0.385	0.133	0.004
Age	0.050	0.007	0.000
Black race	-0.424	1.300	0.744
Hispanic race	-2.331	1.184	0.049
Male sex	0.454	0.140	0.001
Systolic blood pressure (mmHg)	0.001	0.007	0.871
Diabetes	0.831	0.130	0.000
Former smoking	0.148	0.143	0.300
Current smoking	0.472	0.177	0.008
HDL cholesterol (mg/dl)	0.003	0.004	0.411
LDL cholesterol (mg/dl)	-0.002	0.001	0.098
Systolic blood pressure (mmHg) × Black race	0.004	0.008	0.603
Systolic blood pressure (mmHg) × Hispanic race	0.016	0.008	0.037

The significant set of covariates is almost same in two models. Covariates, moderate-to-heavy activity, any cardiac disease, male sex, diabetes, current smoking, and LDL cholesterol level are significant. The parameter estimates of two models are similar even though the exact values and their p values are slightly different. The hazard ratio of an individual with diabetes having a stroke compared to an individual without diabetes is 2.257. This means the hazard ratio increases 2.257 times of baseline hazard when a subject suffers from a diabetes. It is observed that any cardiac disease, male sex, and current smoking also increase the relative risk into 1.47, 1.58, and 1.60 times, respectively. On the other hand, the Hispanic race significantly decreases the risk of having a stroke. It decreases the relative risk into 0.103. The risk factor, age at study enrollment is added in the model using follow-up time since it is not related to outcome. The effect of age at study enrollment on the follow-up time is very significant. The hazard ratio increases 1.65 times when a subject's age at study enrollment is 10-years older. In the Framingham study, the interaction for systolic blood pressure and antihypertensive medications was included in the model and the NOMAS represented the significance of an interaction for diastolic blood pressure and antihypertensive medications. There were no significant interaction for each sex or race-ethnic group (Sacco et al., 2009) whereas in this study, the 2-way interaction of Hispanic race and systolic blood pressure is involved having a significant effect with p value of 0.045.

The log-log Kaplan-Meier curves are drawn for nine binary covariates in order to confirm the proportional hazard assumption (Fig. 3.3 (a) – Fig. 3.3 (i)). The interval of two curves in each plot represents the parameter estimate of each covariate.

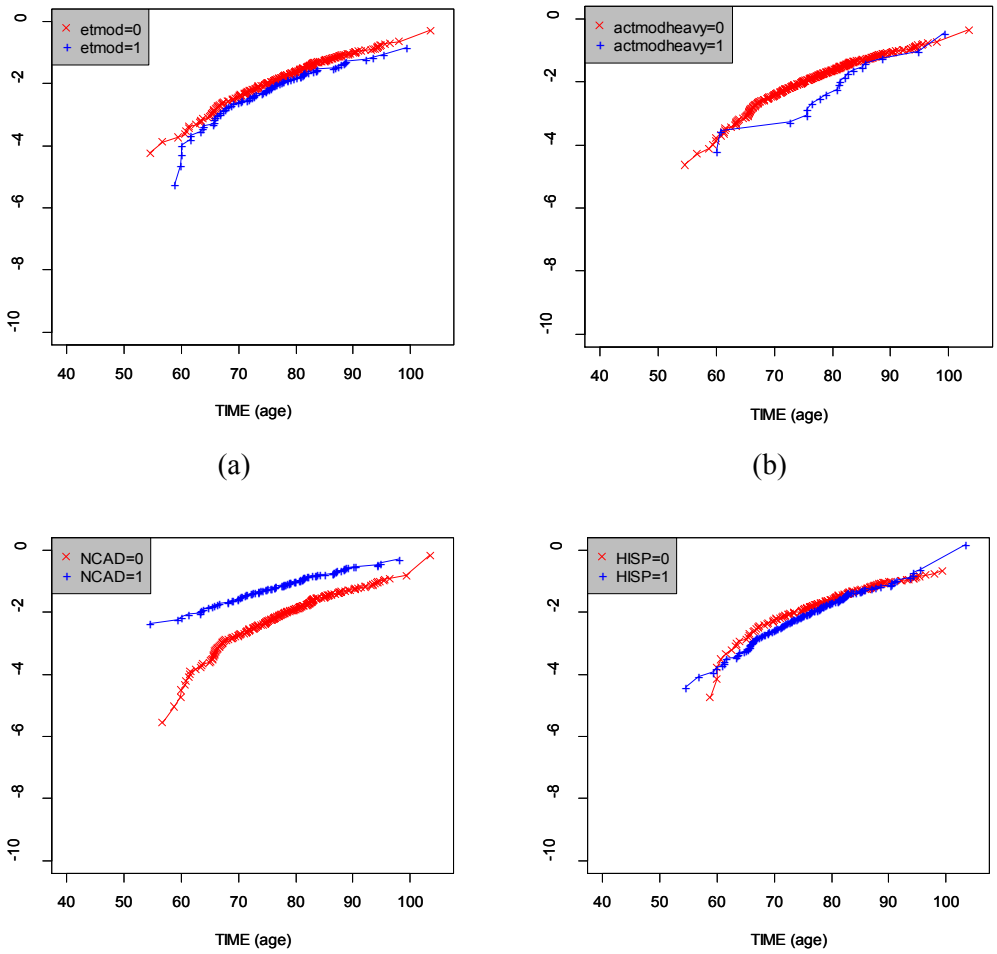
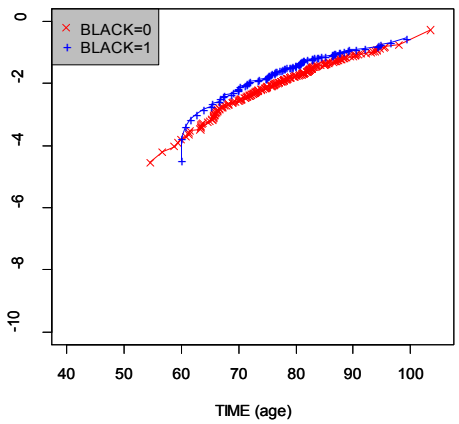
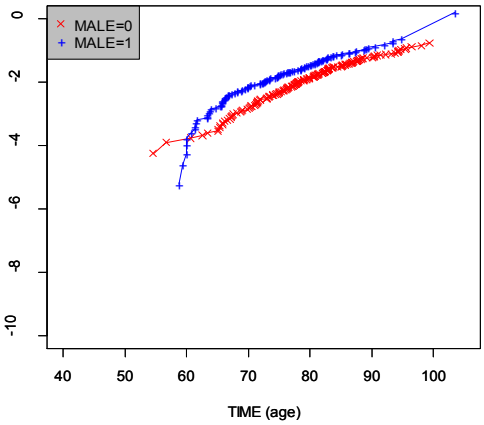


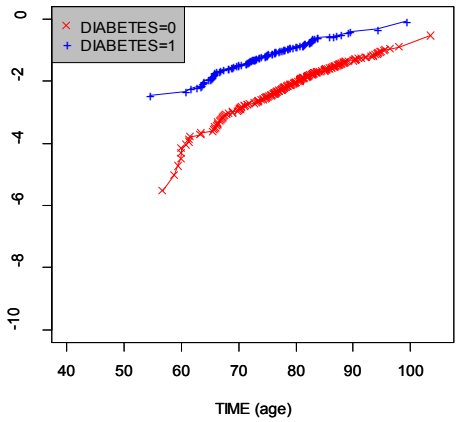
Figure 3.3 Log-log Kaplan-Meier curves of covariates based on the survival function of age at all types of strokes. (a) Moderate alcohol consumption, (b) Moderate to heavy activity, (c) Any cardiac disease, (d) Hispanic race (continued).



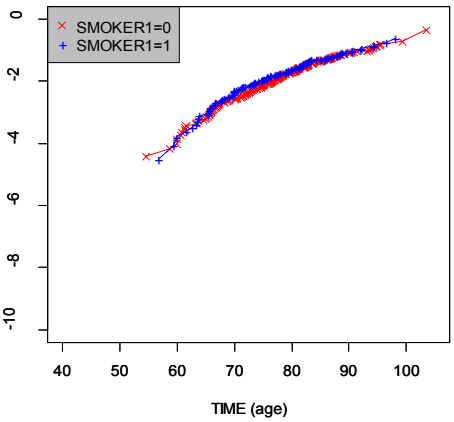
(e)



(f)

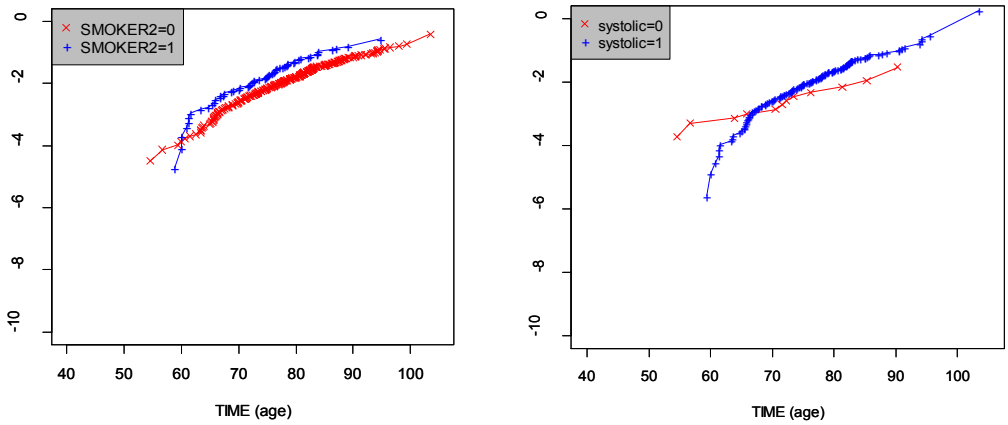


(g)



(h)

Figure 3.3 Log-log Kaplan-Meier curves of covariates based on the survival function of age at all types of strokes. (e) Black race, (f) Male sex, (g) Diabetes, (h) Former smoking (continued).



(i) (j)

Figure 3.3 Log-log Kaplan-Meier curves of covariates based on the survival function of age at all types of strokes. (i) Current smoking, (j) Systolic indicator variable in only Hispanic race subjects.

The plots of any cardiac disease (Fig. 3.3 (c)) and diabetes (Fig. 3.3 (g)) show similar patterns that the interval of two curves decrease as age increases. For moderate to heavy activity (Fig. 3.3 (b)), the plot shows that the stroke risk effect significantly changes according to the value of age. It indicates that the effect of parameter is high around subject age 70 years whereas it decreases significantly after about age 80 years. In the plots of Hispanic race (Fig. 3.3 (d)), black race (Fig. 3.3 (e)), male sex (Fig. 3.3 (f)) and current smoking (Fig. 3.3 (i)), the initial part of two curves are crossed over. Most of two curves in each log-log Kaplan-Meier plot are not parallel and this imply that the proportional hazard assumption in the Cox's model can be biased. In addition, using the 'systolic' indicator variable, the log-log Kaplan-Meier plot is drawn only for the individuals in the group of Hispanic race (Fig. 3.3 (j)). This plot also express the

cross of two curves, which suggests the bias of the proportional hazard assumption.

In order to certify the time-varying effect of each covariate, the extended version of Cox's models are fitted. Each model includes only one covariate with a time-varying coefficient and other covariates in the model are treated as variable having a constant stroke risk effect. For each model, plot is drawn to express the time-varying risk effect of a covariate with a time-varying coefficient. The model with a time-varying coefficient of 'systolic' indicator variable is also fitted and the graph of this time-varying effect is plotted. These are represented in the Figure 3.4 (a) - Figure 3.4 (h).

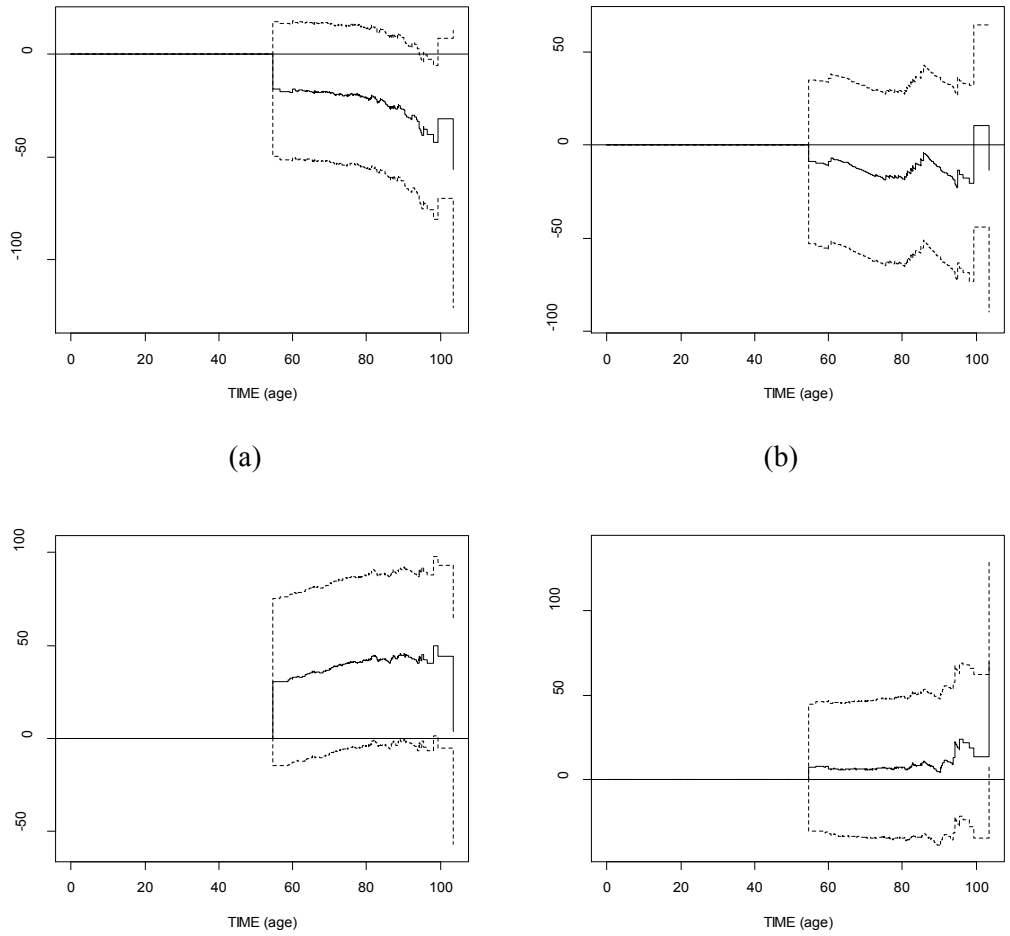
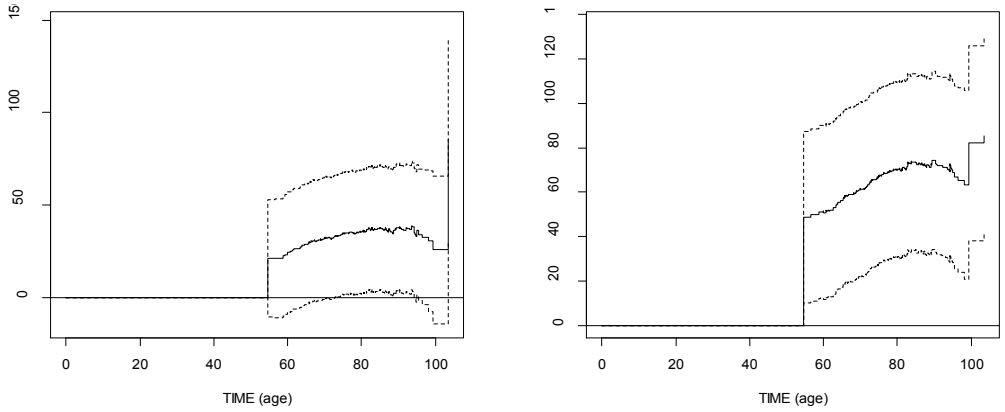
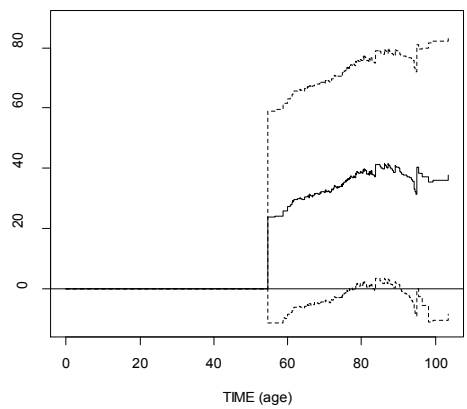


Figure 3.4 Cumulative coefficients plots of covariates in Cox's model with time-varying coefficients using survival time, age at all types of strokes onset. (a) Moderate alcohol consumption, (b) Moderate to heavy activity, (c) Any cardiac disease, (d) Hispanic race (continued).

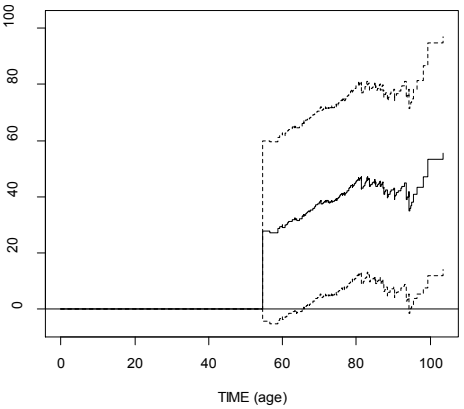


(e)

(f)



(g)



(h)

Figure 3.4 Cumulative coefficients plots of covariates in Cox's model with time-varying coefficients using survival time, age at all types of strokes onset. (e) Male sex, (f) Diabetes, (g) Current smoking, (h) Systolic indicator variable.

The horizontal axis indicates the age of a subject and vertical axis of the graphs expresses the value of cumulative coefficients. It can be interpreted that the slope at a certain age represents the stroke risk effect of each covariate. Thus, the positive slope means that the existence of variable characteristic causes increase in stroke risk and if the slope of the curve is negative, this variable decreases the occurrence of stroke. The plots of male sex (Fig. 3.4 (e)), diabetes (Fig.3.4 (f)) and current smoking (Fig.3.4 (g)) show the slopes of initial part are positive. However, their slopes of curves change from positive to negative after the age 80 years. The plot of any cardiac disease (Fig. 3.4 (c)) shows generally constant slope of the curve. For moderate alcohol consumption (Fig. 3.4 (a)), the effect of risk factor decreases the stroke hazard and this stroke decreasing effect becomes gradually large as a subject's age increases. In the plots for moderate-to-heavy, Hispanic race, and systolic indicator variable, the slope of their curves change with high variation. The moderate-to-heavy risk covariate which shows the strong implication of bias of proportional hazard assumption in its log-log Kaplan-Meier curve represents dynamic change in its cumulative coefficients graph (Fig. 3.3 (b)). For Hispanic race, the risk effect is not large until a subject's age is over middle of eighties where it increases largely (Fig. 3.4 (d)). The plot of systolic indicator variable represents constant level of stroke increasing effect until 80 years of an individual's age and after this point, it shows large variation of risk effect (Fig. 3.3 (h)). Even though some cumulative coefficients plots represent the dynamic variations of coefficients, none of risk factors had a p value < 0.10 in the Kolmogorov-Smimov test and Cramer von Mises test, suggesting that there is no significant risk covariate with time-varying effect.

3.2. Ischemic Stroke

The Cox proportional hazard models are fitted using the ischemic stroke as an event of interest. The two models, using age and follow-up time of having an ischemic stroke are fitted in this chapter. The Table 3.3 describes the parameter estimate, exponential of parameter estimate, standard error and p value of risk factors in the final Cox proportional hazard model with age at ischemic stroke onset as an outcome of interest. The Table 3.4 represents the same items when follow-up time at ischemic stroke onset is used.

Table 3.3 Cox proportional hazard model with survival time, age of having an ischemic stroke

Risk Factor Variables	Parameter Estimate	Standard. Error	p-value
Moderate alcohol consumption	-0.409	0.160	0.011
Moderate-to-heavy physical activity	-0.151	0.262	0.563
Any cardiac disease	0.428	0.142	0.003
Black race	-1.001	1.417	0.480
Hispanic race	-3.188	1.290	0.013
Male sex	0.411	0.151	0.006
Systolic blood pressure (mmHg)	-0.003	0.008	0.625
Diabetes	0.908	0.139	0.000
Former smoking	0.176	0.154	0.254
Current smoking	0.507	0.192	0.008
HDL cholesterol (mg/dl)	0.001	0.004	0.743
LDL cholesterol (mg/dl)	-0.002	0.001	0.201
Systolic blood pressure (mmHg) × Black race	0.008	0.009	0.390
Systolic blood pressure (mmHg) × Hispanic race	0.022	0.008	0.012

The set of significant risk factors is slightly different when compared to the models using all kinds of strokes as an outcome of interest. The models using all types of strokes (Table 3.1, Table 3.2) contain the LDL cholesterol level as a significant covariate. However, the models using ischemic stroke as an outcome (Table 3.3, Table. 3.4) do not include significant variables related to cholesterol levels. Excluding LDL cholesterol level, the significant covariates are same in two model using all types of strokes and ischemic stroke. That is, covariates, moderate-to-heavy activity, any cardiac disease, male sex, diabetes and current smoking are significant risk factors.

Table 3.4 Cox proportional hazard model with survival time, follow-up time to ischemic stroke occurrence

Risk Factor Variables	Parameter Estimate	Standard. Error	p-value
Moderate alcohol consumption	-0.414	0.160	0.010
Moderate-to-heavy physical activity	-0.123	0.262	0.639
Any cardiac disease	0.410	0.142	0.004
Age	0.054	0.007	0.000
Black race	-1.071	1.417	0.450
Hispanic race	-3.292	1.287	0.011
Male sex	0.416	0.151	0.006
Systolic blood pressure (mmHg)	-0.004	0.008	0.582
Diabetes	0.929	0.139	0.000
Former smoking	0.179	0.154	0.244
Current smoking	0.503	0.192	0.009
HDL cholesterol (mg/dl)	0.001	0.004	0.742
LDL cholesterol (mg/dl)	-0.002	0.001	0.218
Systolic blood pressure (mmHg) × Black race	0.008	0.009	0.364
Systolic blood pressure (mmHg) × Hispanic race	0.023	0.008	0.009

The hazard ratio of an individual with diabetes having an ischemic stroke compared to an individual without diabetes is 2.480 (Table 3.3). This value is larger than that of all stroke model (Table 3.1), 2.257. It is observed that any cardiac disease, male sex, and current smoking also increase the relative risk of ischemic stroke into 1.53, 1.51, and 1.66 times (Table 3.3). These quantities are also higher than those in the model with age of having all types of strokes, 1.47, 1.58, and 1.60 times, respectively (Table 3.1). Meanwhile, the Hispanic race significantly decreases the risk of having a stroke. It decreases the relative risk into 0.04 and this value is lower than that of model in Table 3.1, 0.103. The risk factor, age at study enrollment is added in the model using follow-up time (Table 3.4). The effect of age at study enrollment on follow-up time to occurrence of ischemic stroke is very significant as before the model using follow-up time to all kinds of strokes occurrence (Table 3.2). The parameter estimate of age is a little higher in the ischemic stroke model. That is, the ischemic stroke hazard ratio increases 1.72 times when a subject's age at study enrollment is 10-years older. In addition, having a same result with the model using all stroke outcome, the 2-way interaction of Hispanic race and systolic blood pressure is involved as a significant factor with p value of 0.011 in the model with ischemic stroke event.

For comparing two models using age and follow-up time to ischemic stroke occurrence, the significant set of covariates is almost same in two models as before. The parameter estimates of two models are similar even though the exact values and their p values are slightly different.

In order to confirm the proportional hazard assumption, the log-log Kaplan-Meier curves are plotted for nine binary covariates (Fig.3.5 (a) - Fig.3.5 (i)). These graphs using the Kaplan-Meier estimates of survival function of ischemic stroke occurrence represent nearly same pattern compared to the plots in the Figure 3.3. The plots of any

cardiac disease (Fig. 3.5 (c)) and diabetes (Fig. 3.5 (g)) show similar patterns that the interval of two curves decrease as age increases. For moderate to heavy activity (Fig. 3.5 (b)), the plot shows that the stroke risk effect significantly changes according to the value of age. It indicates that the effect of parameter is high around subject age 70 years whereas it decreases significantly after about age 80 years. In the plots of Hispanic race (Fig. 3.5 (d)), black race (Fig. 3.5 (e)), male sex (Fig. 3.5 (f)) and current smoking (Fig. 3.5 (i)), the initial part of two curves are crossed over. Most of two curves in each log-log Kaplan-Meier plot are not parallel and this implies that the proportional hazard assumption in the Cox' model can be biased. In addition, like as the former analysis, using the 'systolic' indicator variable, the log-log Kaplan-Meier plot is drawn only for the individuals in the group of Hispanic race (Fig. 3.5 (j)). This plot also express the cross of two curves, which suggests the bias of the proportional hazard assumption.

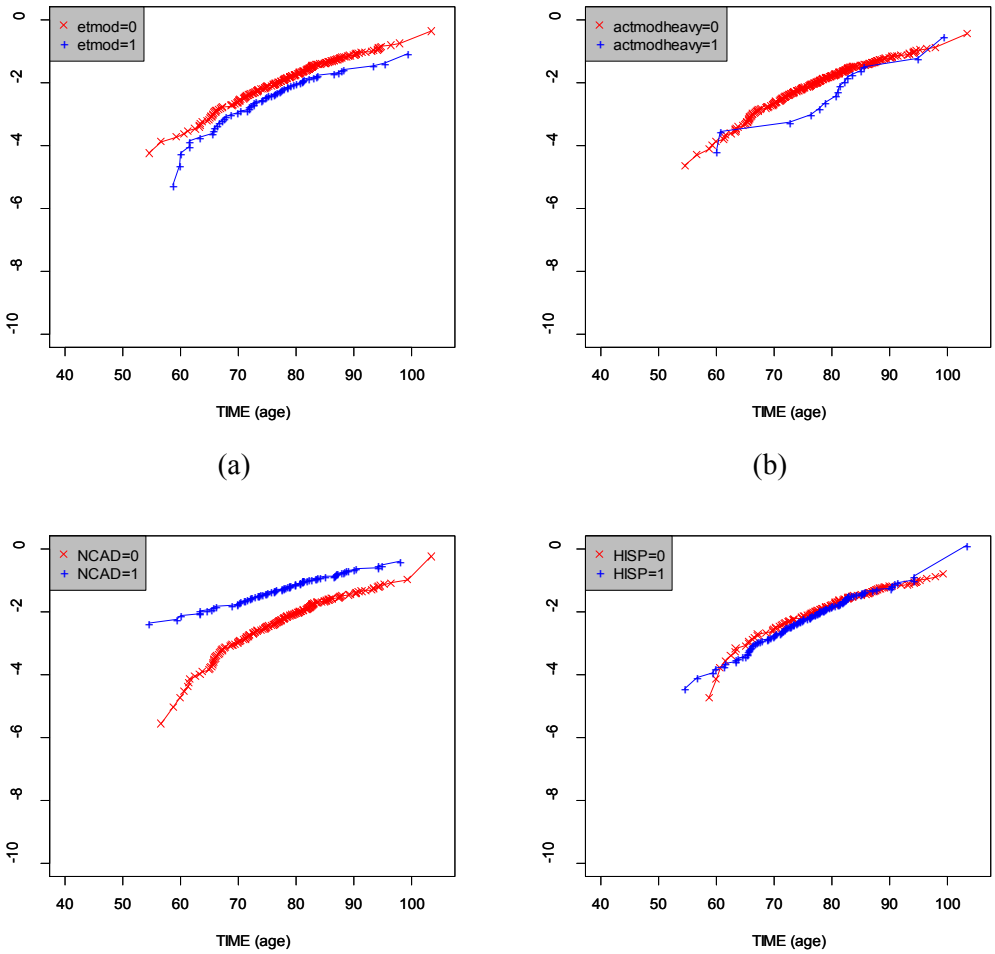
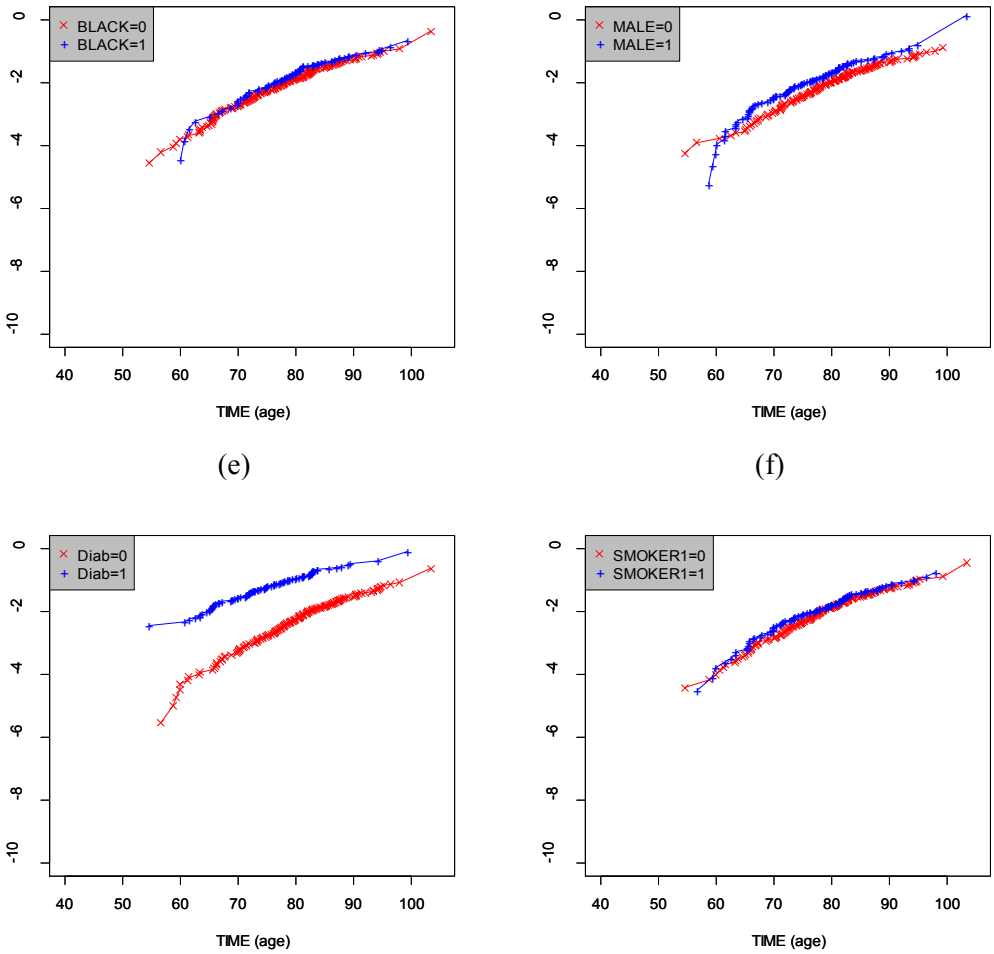


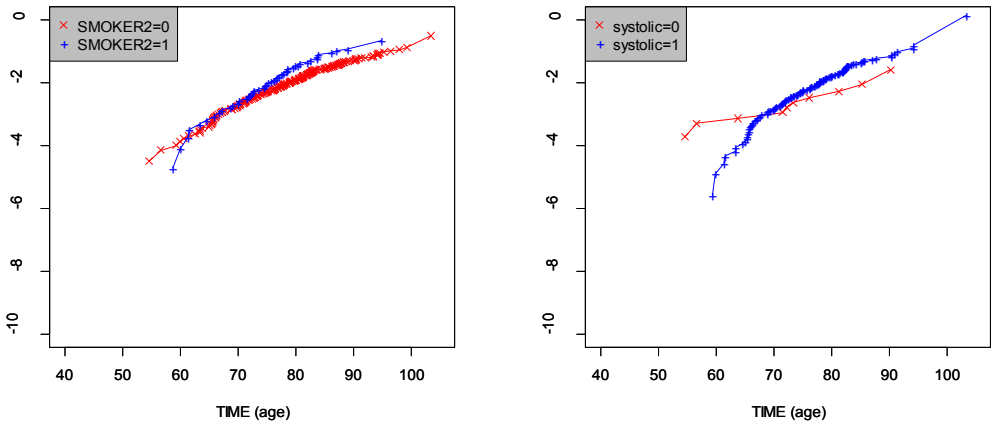
Figure 3.5 Log-log Kaplan-Meier curves of covariates based on the survival function of age at ischemic stroke onset. (a) Moderate alcohol consumption, (b) Moderate to heavy activity, (c) Any cardiac disease, (d) Hispanic race (continued).



(e) (f)

(g) (h)

Figure 3.5 Log-log Kaplan-Meier curves of covariates based on the survival function of age at ischemic stroke onset. (e) Black race, (f) Male sex, (g) Diabetes, (h) Former smoking (continued).



(i) (j)

Figure 3.5 Log-log Kaplan-Meier curves of covariates based on the survival function of age at ischemic stroke onset. (i) Current smoking, (j) Systolic indicator variable in only Hispanic race subjects.

For confirming the time-varying effect of each covariate, the extended version of Cox's models are fitted using the same methods in the previous cases. For each model, plot is drawn to express the time-varying risk effect of a covariate with a time-varying coefficient. The model with a time-varying coefficient 'systolic' indicator variable is also fitted and the graph of this time-varying effect is plotted. These are represented in the Figure 3.6 (a)-(h).

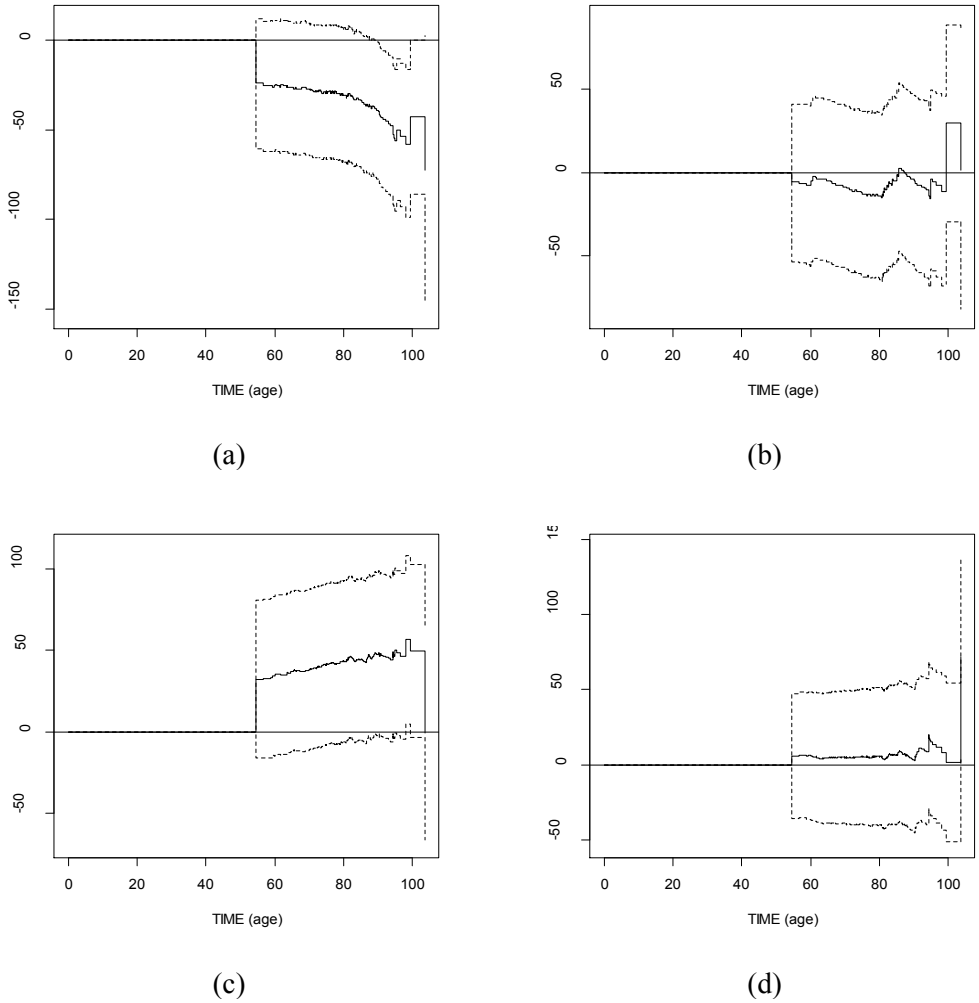
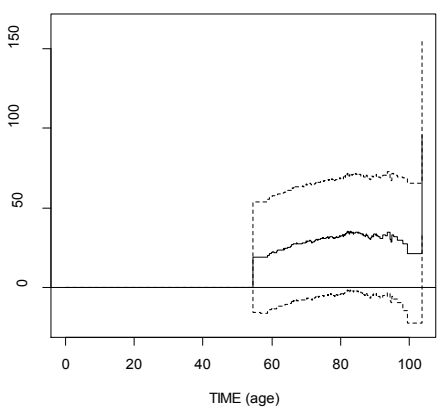
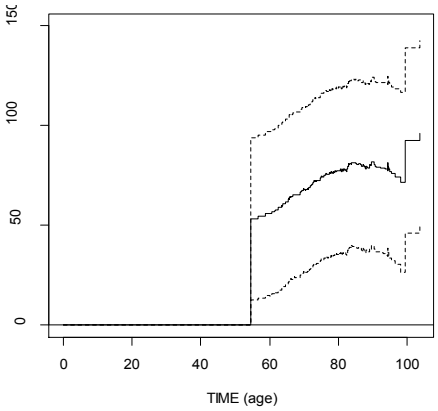


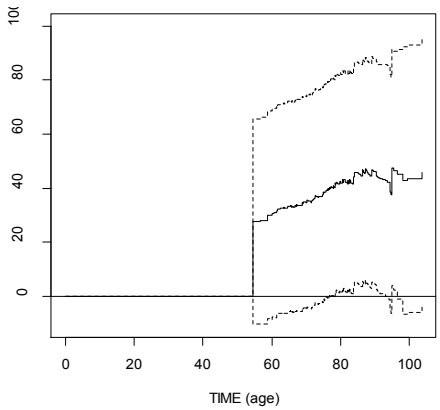
Figure 3.6 Cumulative coefficients plots of covariates in Cox's model with time-varying coefficients using survival time, age at ischemic stroke onset. (a) Moderate alcohol consumption, (b) Moderate to heavy activity, (c) Any cardiac disease, (d) Hispanic race (continued).



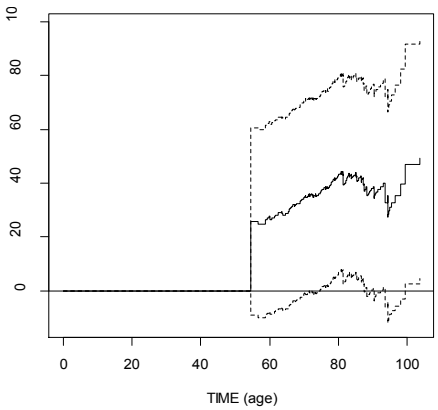
(e)



(f)



(g)



(h)

Figure 3.6 Cumulative coefficients plots of covariates in Cox's model with time-varying coefficients using survival time, age at ischemic stroke onset. (e) Male sex, (f) Diabetes, (g) Current smoking, (h) Systolic indicator variable.

The plots based on the models with ischemic stroke as an event of interest (Fig, 3.6) express similar patterns of time-varying coefficients of covariates in the model with all kinds of strokes (Fig. 3.4). The plots of male sex (Fig. 3.6 (e)), diabetes (Fig.3.6 (f)) and current smoking (Fig.3.6 (g)) show the slopes of initial part are positive. However, their slopes of curves change into negative after the age 80 years. The plot of any cardiac disease (Fig. 3.6 (c)) shows generally constant slope of the curve. For moderate alcohol consumption (Fig.3.6 (a)), the effect of risk factor decreases the ischemic stroke hazard and this ischemic stroke decreasing effect becomes generally large as a subject's age increases. In the plots for moderate-to-heavy, Hispanic race, and systolic indicator variable, the slopes of their curves change with high variation such as the corresponding graphs in Figure 3.4. The moderate-to-heavy risk covariate which shows the strong implication of bias of proportional hazard assumption in its log-log Kaplan-Meier curve (Fig. 3.5 (b)) represents dynamic changes in its cumulative coefficients graph (Fig. 3.6 (b)). For Hispanic race, the risk effect is not large until a subject's age is over middle of eighties where it increases largely (Fig. 3.6 (d)). The plot of systolic indicator variable represents constant level of stroke increasing effect until 80 years of an individual's age and after this point, it shows large variation of risk effect (Fig. 3.6 (h)). The significance of time-varying coefficients is equivalent to the results of models using all kinds of strokes. It is same with the cases using all types of strokes that although some cumulative coefficients plots represent the dynamic variation of coefficients, any of risk factors did not have a p value < 0.10 in the Kolmogorov-Smirnov test and Cramer von Mises test, which means that there is no significant risk covariate with time-varying effect.

Chapter 4

Conclusion

In this study, we tried to investigate stroke risk factors using the NOMAS dataset obtained from the prospective cohort of 3,298 subjects in Northern Manhattan. In particular, we introduced the age of having strokes as survival time for fitting the Cox proportional hazard models, whereas the NOMAS used follow-up time, years took to stroke occurrence since subjects' enrollment at study. For comparing two models using age and follow-up time, models using follow-up time were also fitted and these two models were constructed for two specific outcomes, all stroke and ischemic stroke, respectively.

When we compare two models using age and follow-up time, the values of parameter estimate were similar and the significant risk set of covariates was equal, excluding the a risk factor, age at study enrollment, only added to the latter model. Even though there were no interaction for race and other risk factors in final model of NOMAS, the interaction effect for Hispanic race and systolic blood pressure was significant in both models using age and follow-up time. For confirming the proportional hazard assumption of the model using the age of having all strokes, log-log Kaplan-Meier curves were drawn for binary covariates. Some of these plots implied the bias of proportional hazard assumption in the

Cox's model. In particular, the moderate-to-heavy risk factor showed dynamic variation of the interval of two curves in the plot. The plot of indicator variable of systolic blood pressure expressed discrete cross pattern of two curves. Even though there was no significant difference between models using age and follow-up time, we could confirm the possibility of existence of the covariates with time-varying coefficients through log-log Kaplan-Meier curves.

The extended version of Cox's models with time-varying coefficients were fitted and cumulative coefficients graphs were represented. These graphs showed the change of risk effects of covariates according to subjects' age. For the variables, moderate-to-heavy activity and systolic blood pressure indicator variable, the risk effects of them change with high variation. Despite of the existence of these variables, there was no significant risk factor with time-varying coefficients referring to the result of tests. It is expected that if more covariates used in NOMAS are available, the significant risk variables could be founded since we detected the high variation of risk factor effect through the cumulative coefficients curves.

When we compared the models using all types of strokes and ischemic stroke only, LDL cholesterol level was solely included in the models using all kinds of strokes as a significant risk factor. The set of other significant covariates was same in two models. The values of parameter estimate were different from each other. For ischemic stroke, the decreasing and increasing effects of risk factors were larger than in the case of all types of strokes models.

Bibliography

Breslow, N. E. (1972) Discussion on Professor Cox's paper in "Regression models and life tables". *Journal of the Royal Statistical Society, Series B* 34, 187-220.

Cox, D. R. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34 (2), 187-220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62 (2), 269-276.

D'Agostino, R. B., P. A. Wolf, A. J. Belanger, and W. B. Kannel (1994) Stroke risk profile: adjustment for antihypertensive medication. The Framingham Study. *Stroke*, 25 (1), 40-43.

D'Agostino, R. B., S. Grundy, L. M. Sullivan, and P. Wilson (2001) Validation of the Framingham coronary heart disease prediction scores. *JAMA: the journal of the American Medical Association*, 286 (2), 180-187.

D'Agostino, R. B., R. S. Vasan, M. J. Pencina, P. A. Wolf, M. Cobain, J. M. Massaro, and W. B. Kannel (2008) General cardiovascular risk profile for use in primary care the Framingham Heart Study. *Circulation*, 117 (6), 743-753.

Goldstein, L. B., R. Adams, M. J. Alberts, L. J. Appel, L. M. Brass, C. D. Bushnell, and R. L. Sacco, (2006) Primary prevention of ischemic stroke a guideline from the American Heart Association/American Stroke Association Stroke Council. *Stroke*, 37 (6), 1583-1633.

Kaplan, E. L. and P. Meier (1958) Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53 (282), 457-481.

Martinussen, T. and T. H. Scheike (2006) Dynamic regression models for survival data. *Springer*.

Sacco, R. L., K. Anand, H. S. Lee, B. Boden-Albala, S. Stabler, R. Allen, and M. C. Paik (2004) Homocysteine and the risk of ischemic stroke in a Triethnic Cohort the Northern Manhattan Study. *Stroke*, 35 (10), 2263-2269.

Sacco, R. L., M. Khatri, T. Rundek, Q. Xu, H. Gardener, B. Boden-Albala, ... and M. C. Paik (2009). Improving global vascular risk prediction with behavioral and anthropometric factors the multiethnic NOMAS (Northern Manhattan Cohort Study). *Journal of the American College of Cardiology*, 54 (24), 2303-2311.

Sims, N. R., and H. Muyderman (2010) Mitochondria, oxidative metabolism and cell death in stroke. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1802* (1), 80-91.

Tsai, W. Y. (1990) Testing the assumption of independence of truncation time and failure time. *Biometrika*, 77 (1), 169-177.

Wilson, P. W., R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel (1998) Prediction of coronary heart disease using risk factor categories. *Circulation*, 97 (18), 1837-1847.

국 문 초 록

뇌졸중 위험요인 분석을 위하여 개선된 생존시간을 이용하는 시간변동계수를 포함한 Cox 비례위험모형

A Cox Proportional Hazard Model with Time-Varying Coefficients Using Revised Survival Time to Investigate Stroke Risk Factors

세계보건기구의 발표에 따르면 심혈관계질환은 전세계 사망원인 중 1 위를 차지하고 있다. 심혈관계질환 중에서도 뇌졸중의 경우, 그에 따른 사망률은 줄어들고 있지만 발병률은 지속적으로 증가하고 있는 추세이다. 뇌졸중을 일으키는 위험요인에 관하여, Framingham study, Northern Manhattan stroke study (NOMAS)를 비롯한 여러 연구들이 이루어져왔다. 본 연구에서는 NOMAS 를 통하여 얻어진 자료를 바탕으로 Cox 비례위험모형을 이용하여 뇌졸중 발병에 영향을 미치는 위험요인들을 밝히고자 하였다. 기존의 NOMAS 에서는 피실험자가 연구에 가입한 이후부터 뇌졸중 발생까지의 시간을 생존시간으로 이용하는데 반하여, 본 연구에서는 피실험자의 뇌졸중 발생 나이를 생존시간으로 대체하여 모형적합을 시도하였다. 그래프를 이용하여 Cox 모형의 비례위험가정 충족 여부를 검토하였으며, 특정 위험요인들의 위험효과가 피실험자의 나이 변화에 따라 달라짐을 확인할 수 있었다. 시간변동계수를 포함하는 Cox 모형적합을 통하여 이러한 위험요인들의 시간에 따른 위험효과 변화와 그 유의성을 검토하였다. 또한, 모든 종류의 뇌졸중을 대상으로 한 연구에 더불어, 허혈성 뇌졸중 발병에 초점을 두어 두 연구결과 간의 차이를 밝히고자 하였다.

주요어: 뇌졸중, 북맨하탄 연구, Cox 비례위험모형, 생존시간, 비례위험가정, 시간변동계수

학 번: 2012-20231