**Final Exam of “Event History & Survival Analysis”**

**2017, International Summer School of Renmin University of China**

This is an open-book exam and you may consult to any source of information, including the class notebook and textbooks. However, you must complete this exam independently without discussing any portion of the exam with your classmates or consulting with anyone else.

**Part I. Multiple Choice Questions (Only One Answer is Correct, 1 Point for Each)**

1. Which of the following expressions of Cox regression is ***not*** correct? \_b\_\_\_
   1. 
   2. 
   3. 
   4. 
2. Given a study window of 12 months and a time unit of a quarter(i.e., 3 months), we observed that a study participant took 5 months to experience a change on the event but was randomly censored, the person-time data for this participant in a discrete-time analysis look like the follows: \_\_a\_\_
   1. \_Period \_Y

1 0

2 0

* 1. \_Period \_Y

1 0

2 1

* 1. \_Period \_Y

1 0

2 0

3 0

4 0

5 1

* 1. \_Period \_Y

1 0

2 0

3 0

4 0

5 0

1. Given a study window of 12 months and a time unit of a quarter(i.e., 3 months), we observed that a study participant took 5 months to make a change on the study event, the person-time data for this participant in a discrete-time analysis look like the follows: \_\_b\_\_
   1. \_Period \_Y

1 0

2 0

* 1. \_Period \_Y

1 0

2 1

* 1. \_Period \_Y

1 0

2 0

3 0

4 0

5 1

* 1. \_Period \_Y

1 0

2 0

3 0

4 0

5 0

1. The partial likelihood estimator assumes \_\_a\_\_.
   1. The hazard for any individual is a fixed proportion of the hazard for any other individual
   2. The baseline hazard equals to zero (i.e. h0 = 0)
   3. The baseline hazard equals to one (i.e. h0 = 1)
   4. a and c
   5. All a, b, & c
2. Knowing that the endpoints of a 95% confidence interval for an estimated hazard ratio are [.76719, 1.4396], and the estimated coefficient is .0496761, what is the eof the estimated coefficient [i.e., S.E.()]? \_\_c\_
   1. 1.050931
   2. .1687385
   3. .16056
   4. .26502
3. Knowing that the endpoints of a 95% confidence interval for an estimated hazard ratio are [.76719, 1.4396], and the estimated coefficient is .0496761, what is the point estimate of the hazard ratio? \_\_a\_\_
   1. 1.050931
   2. .1687385
   3. .16056
   4. .26502
4. The likelihood ratio test can be employed to address the following question(s): \_e\_
   1. the statistical significance of a predictor in a Cox regression.
   2. the statistical significance of a categorical variable that is comprised of several dichotomous variables in a Cox regression.
   3. the goodness of model fit to the sample data in a Cox regression.
   4. b& c
   5. All a, b, & c
5. In a Cox regression using income as predictor and income is measured in single Yuan, the researcher found that the estimated coefficient for income is .0007137 and its standard error is .0001051, which of the following interpretation is correct? \_e\_\_\_
   1. Every one-Yuan increase in income increases the speed of change by 7.4%
   2. Every 100-Yuan increase in income increases the speed of change by 7.4%
   3. We are 95% confident that in the population the impact of 100-Yuan increase fastens the speed of change at least 5.2% and at most 9.6%
   4. All a, b, & c
   5. b & c
6. A piecewise exponential (PE) model differs from a discrete-time (DT) model in the following features: \_\_d\_\_.
   1. PE makes a parametric assumption about the event time while DT does not
   2. PE adds a duration variable to DT
   3. DT throws out the information about the event time while PE does not
   4. a & b
   5. All a, b, & c
7. Which of the following study questions require or requires a multilevel survival analysis to control for clustering effect? \_\_e\_
   1. A study of alcoholic relapse using multiple episodes
   2. A study of recidivism using participants’ times of being released from prison and reentering into prison
   3. A study of timing of promotion in which employees are nested within companies
   4. All a, b, & c
   5. Only a & b

**Part II. An Essay Question (70 Points)**

Write an abstract no more than 400 words to describe a longitudinal study that requires survival analysis. Your description should include the following components:

1. Research Questions;
2. Definition of event, duration, time origin, censoring, and study window;
3. Why survival analysis is suitable to the study?
4. Potential source of data.

**Grading Criteria**

|  |  |
| --- | --- |
| **Criterion** | **Rating** |
| 1. Research questions that do require a survival analysis (whether- and when-test) | **1 2 3 4 5** |
| 2. Clear definition about event of interest and right censoring | **1 2 3 4 5** |
| 3. Clear definition about time origin and study window | **1 2 3 4 5** |
| 4. Data collection is feasible | **1 2 3 4 5** |
| 5. Quality of the discussion of statistical problems | **1 2 3 4 5** |
| 6. Writing and organization of the abstract | **1 2 3 4 5** |
| Total |  |

Your score of this problem is a Likert-type scale based on the 6 items listed above. Rating: 1=very poor, 2=poor, 3=fair, 4=good, 5=very good.

Answer：

(1)Research Questions:

In 2009, our country produced vaccine to control influenza A H1N1. In order to ensure the safety and efficiency of the vaccine, the CDC selected 1000 people from different areas of China to get vaccinated in three batches within2 years. To improve the vaccine more precisely, the researchers want to use survival analysis to test whether the batch of getting vaccinated, gender, age, allergy history and disease history affect the occurrence of adverse reactions or not.

(2) Definition

event--------participants have adverse reactions(i.e.,fever, muscle soreness and other cold symtoms) or get H1N1.

duration-------24 months (in months).

time origin-----the first batch began in January 2009，the second batch began in September 2009 and the third batch began in May 2010.

censoring------the participant gets other treatments besides this vaccine/has other diseases/died/is lost contact

study window------vary from the day getting vaccinated from 240 days after vaccination

(3) Reasons of using survival analysis

First, this study researches the influencing factors of the side effects of the vaccine, which related to the survival time of things (i.e., the effective time of vaccine). Second, the starting and ending time of this study can be clearly confirmed according to the outbreak of the flu and the release time of vaccine, which means that the duration and time origin can be confirmed. Third, this study has time variable-----the timing of having adverse reactions or getting H1N1. Also, I think the influencing factors of this study are all time fixed, which can use the Cox proportional hazard model to test easily. Finally, the event and censoring of this study are clear because the adverse reactions of all kinds of vaccine are defined in detail and the accidents are also recorded clearly. So this study is suitable for survival analysis.

(4) Potential source of data

We can get dataset from the CDC. but the data may not include the information of participants. So we can search the data of Disease Control Center in several cities. If the data is still not adequate enough, we can find data from several hospitals in different cities that provide this vaccine, which contains the information of everyone getting vaccinated.

**Part III. Running Stata (20 Points)**

You are provided with a Stata data file of “rearrest.dta”. The data were collected by Henning and Frueh (1996, *Criminal Justice and Behavior 23*: 523-541). The investigators tracked the criminal histories of 194 inmates released from a medium security prison to determine whether and, if so, when, the former inmates were re-arrested. During the period of data collection, which ranged from one day to three years, 106 former inmates (54.6%) were re-incarcerated. Of these former inmates, 158 previously convicted for a property-related offense, and 61 previously convicted person-related crimes (those who had more than one previous conviction for offenses such as simple assault, aggravated assault, or kidnapping). Variables are coded as follows:

ID – subject ID number;

MONTHS – number of months from release to re-incarceration or to the end of study period;

CENSOR – censoring status, 1=censored, 0=not censored;

PERSONAL – history of person-related crimes, 1=yes, 0=no;

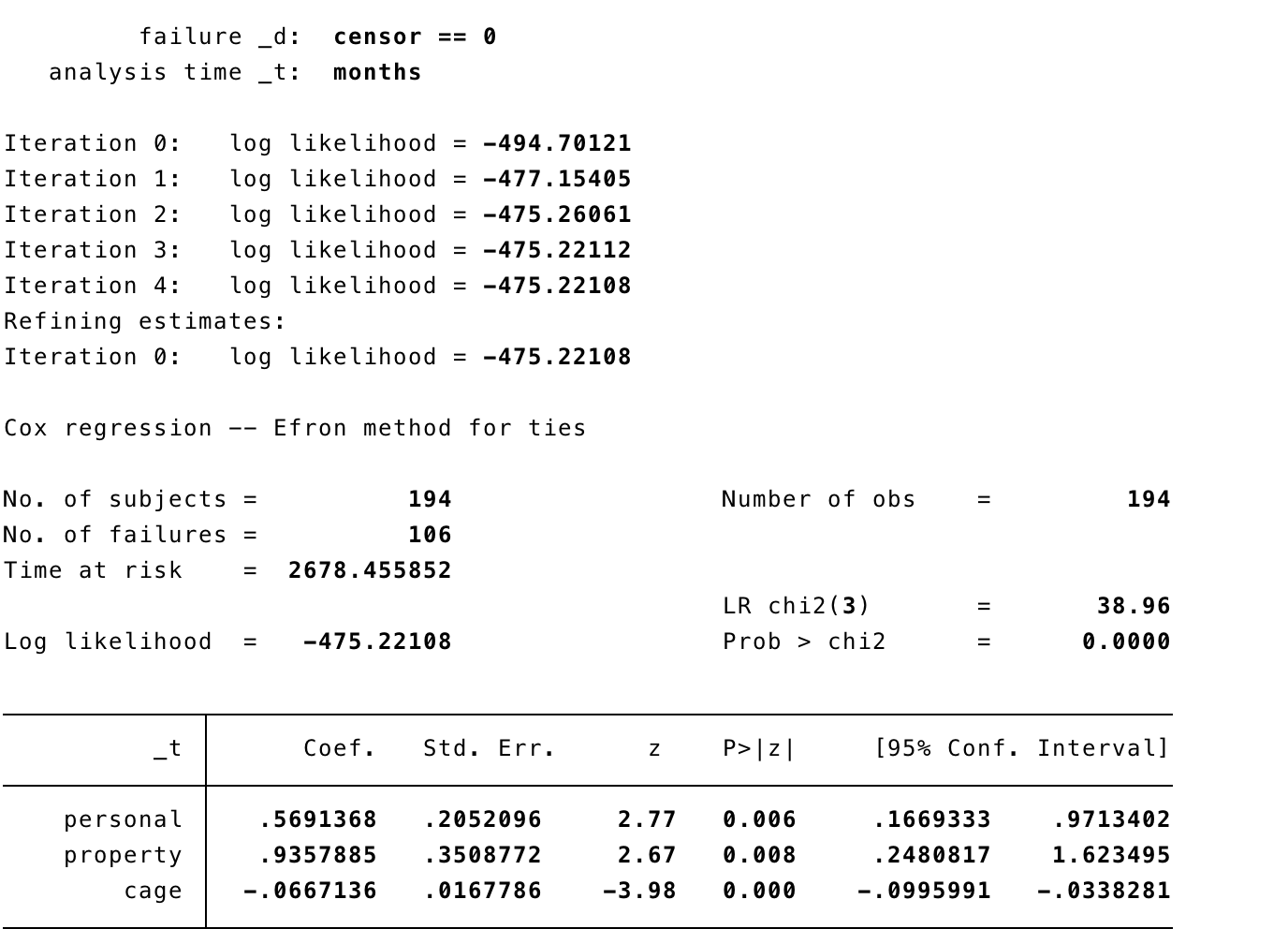
PROPERTY – history of property-related crimes, 1=yes, 0=no

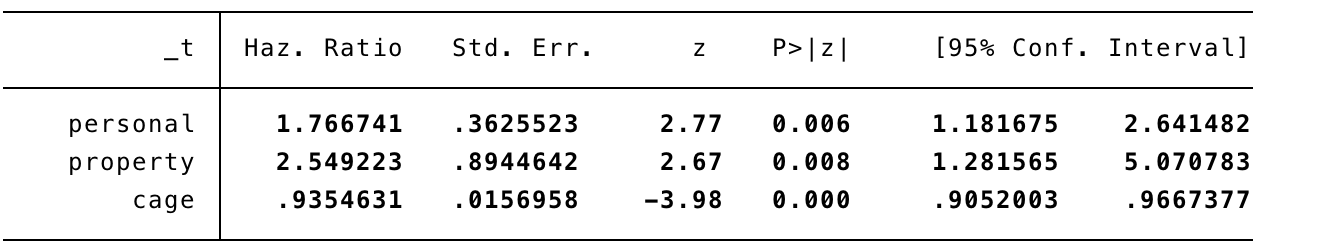
CAGE – Centering variable of age (in years) at time of release (i.e., difference in age upon release between individual and average age of the sample).

Use the data to run a Cox proportional hazard model to test whether age at release, history of person-related crimes, and history of property-related crimes affect the timing of re-incarceration. Write a short paragraph to interpret your findings using results of hazard ratios. Attach your syntax and output.

Answer:

First, I ran the Cox proportional hazard model and the outputs are as follows:





It shows that the overall significance of the model is provided by “LR chi2(3)=38.96, p>chi2 = .0000”. The computation is:

G=2[ Lp(β)- Lp(0)]=2\*[-475.221-(-494.701)]=38.96, Pr(*ᵡ2* (3)>38.96)<.001

The test provides evidence that at least one of the coefficients in this model is significantly associated with survival time.

According to the chart, the current model suggests that the coefficients for *cage* is significant, but the coefficients for *personal* and *property* may not be significant. So I used *stcoxkm* to test the proportionality assumption. The graphs are as follows:

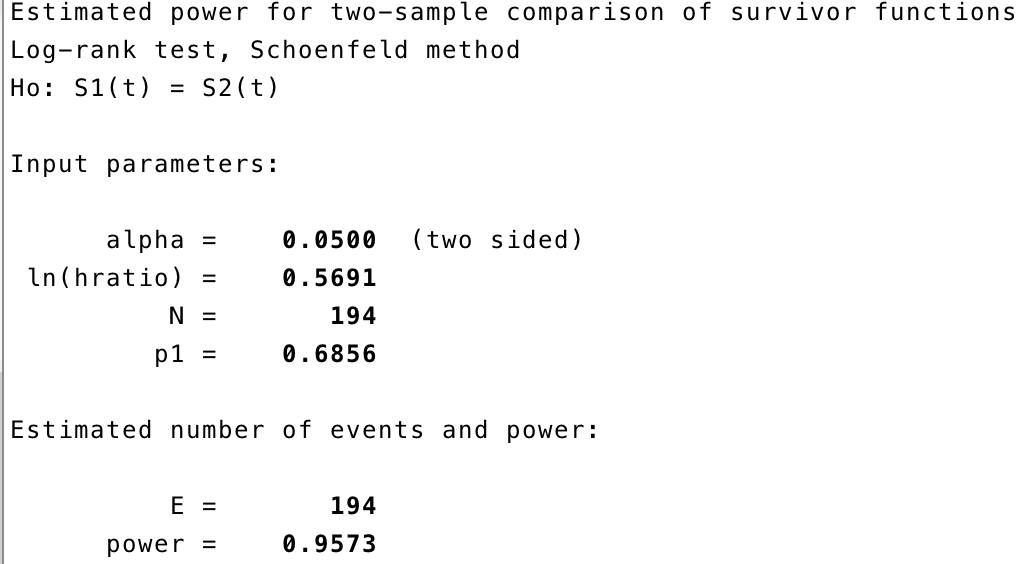
(1) (2)

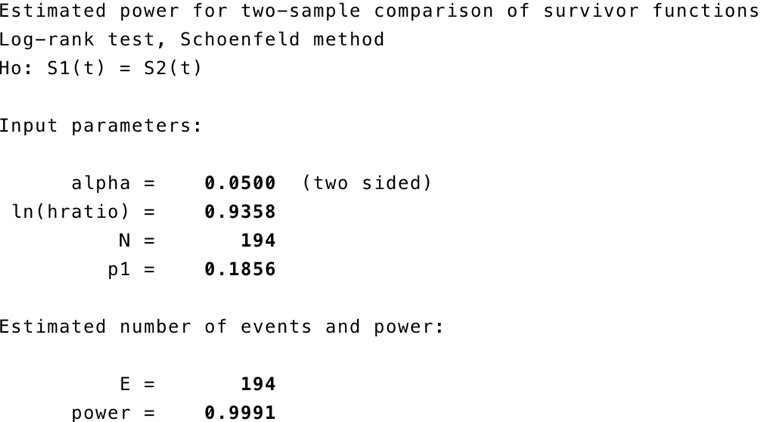
Graph (1) shows that the observed line and predicted line when personal equals 1 are basically parallel, and the observed line and predicted line when personal equals 0 are also parallel. Graph (2) shows that the two pairs of lines are basically parallel. So the test suggests that covariates *personal* and covariates *property* can both be considered to be significant.

Based on the result, we reject the null hypothesis and conclude that age at release(*cage*), history of person-related crimes(*personal*), and history of property-related crimes(*property*) affect the timing of re-incarceration.

Further, the hazard ratio for *personal* is 1.7667, meaning that other things being equal, the hazard rate of re-incarceration for former inmates with person-related crimes is 76.67% higher than those who did not have such crimes. The hazard ratio for *property* is 2.5492, meaning that other things being equal, the hazard rate of re-incarceration for former inmates with property-related crimes is 154.92% higher than those who did not have such crimes. And the hazard ratio for *cage* is 0.9355, meaning that every one year increases in the difference in age upon release between individual and average age of the sample decreases the timing of re-incarceration by 6.45%.

Finally, I used *stpower logrank* tocheck whether or not the studies (*personal* & *property*) have adequate statistical power. The outputs are as follows:





The results provide evidence that both studies have adequate statistical power.

*p.s. ( syntax)*

*cd "/Users/miamiamia/Desktop"*

*//use rearrest.dta*

*use rearrest,clear*

*stset months,failure(censor==)*

*//cox regression*

*stcox personal property cage,efron nohr*

*stcox personal property cage,efron*

*//check the significance*

*stcoxkm, by(personal)*

*stcoxkm, by(property)*

*//adequate power check*

*stpower logrank, n(194) p1(.68556701) hratio(1.7667) schoenfeld*

*stpower logrank, n(194) p1(.18556701) hratio(2.5492) schoenfeld*