Stockholm University
Department of Statistics
(Gebrenegus Ghilagaber)

Take-Home Re-Examination for Analysis of Survival Data with Demographic Applications (Basic-level course, 7.5 HECs, Spring-term 2020).

The Re-exam is handed-out on **Monday 27 April 2020** and replies (with the duly signed Assurance) are expected to be handed-in electronically to the course-coordinator (Gebre@stat.su.se) latest by **Monday 4 May 2020**.

The examination consists of 8 questions that add up to 60 points. **Detailed and well-motivated replies are required in order to get full marks on each question**.

Your replies should include all detailed steps and results (for the theoretical/analytical questions), and explanations as well as most relevant tables and/or figures of results together with their interpretations, conclusions, and implications (for the empirical questions). Input codes and resulting output should be put in an appendix.

The examination will be graded according to the 7-scales that are described in the course description distributed during the start of the course.

Assurance:

I, hereby, assure that I have worked this take-home exam on my own without help from any other person.

Place Date

Signature

Full Name

Person-number

Part 1: Theoretical/Analytical Questions:

Question 1 (6 p)

a. (4p) Derive the log-rank test statistic for equality of two survival curves using the properties of the hyper-geometric distribution where, at each event time, the observations (in each of two groups) may be considered as events or nonevents.

b. (2p) What assumptions are made (implicitly or explicitly) in deriving the test statistic in (a) above?

Question 2 (4 p)

Consider a logistic regression model (for the probability of an event of interest) with one binary explanatory variable. Show that the maximum likelihood estimates of the parameters (constant and coefficient) are exactly equal to the corresponding odds ratios.

Part 2: Analyses of real-life data set.

Questions 3-8 are based on the following data set:

The file **Surv2020_ReExam.dat** (uploaded in Athena in a text format) contains data on entry into first marriage among a sample of women. The five columns represent the following variables:

Column 1: Years (after age 15) to entry into first marriage or to the survey date
Column 2: Indicator of transition to first marriage (0: Not yet married, 1: Married)
Column 3: Birth Cohort with 7 levels (1 indicating the youngest cohort and 7 the oldest cohort)
Column 4: Residence area with 2 levels (1: Urban area, 3: Rural area)
Column 5: Education with 3 levels (0: None, 1: Primary, 2: Secondary or higher)

Question 3 (8p)

Suppose we are interested in modelling the probability of entry to marriage as a function of the three covariates above (Birth Cohort, Residence, and Education).

a. (4p) Fit appropriate model to estimate the relevant parameters and interpret your results.

b. (2p) Use your estimates in (a) to compute the probability of entry into marriage for a randomly selected woman from the oldest cohort who has no education and lives in rural areas.

c. (2p) What do you think is a drawback, if any, of the model you used in (a) above?

Question 4 (8p)

a. (2p) Estimate the survival curves for the different levels of *Birth Cohort* and test if there is a significant difference between them.

b. (2p) Estimate the survival curves for the different levels of *Residence* and test if there is a significant difference between them.

c. (2p) Estimate the survival curves for the different levels of *Education* and test if there is a significant difference between them.

d. (2p) How do your conclusions in 4(a) - 4(c) compare with those in Question 3(a)?

Question 5 (10p)

a. (2p) Model the intensity of experiencing the event of interest as a function of one covariate (*Birth Cohort*). Use the first level of the covariate as baseline (reference) level. Interpret the results and draw your conclusions.

b. (2p) Model the intensity of experiencing the event of interest as a function of two covariates (*Birth Cohort* and *Residence*). Use the first levels of the covariates as baseline (reference) levels. Interpret the results and draw your conclusions.

c. (2p) Does adding *Residence* in 5(b) improve the model in 5(a)? Justify your answer.

d. (2p) Model the intensity of experiencing the event of interest as a function of three covariates (*Birth Cohort*, *Residence*, and *Education*). Use the first levels of the covariates as baseline (reference) levels. Interpret the results and draw your conclusions.

e. (2p) Does adding *Education* in 5(d) improve the model in 5(b)? Justify your answer.

Question 6 (8p)

a. (2p) Estimate the overall intensity of experiencing the event as well as the mean and median survival times assuming that duration (to first marriage) is **exponentially** distributed.

b. (4p) Assume now that duration (age in years) is **exponentially** distributed but with different parameters for the three levels of *Education*. Use appropriate procedure to test for the equality of the population-intensities of experiencing the event across the three educational levels.

c. (2) Are your results in 6(b) in accordance with those in 4(c)? If not, what do you think the reason can be?

Question 7 (8p)

Suppose now we are interested in modeling the effect of the three covariates on the time until event.

a. (5p) Fit all possible models using all three covariates in the models and interpret your results.

b. (3p) Which model fits the data ‘best’? Justify your answer.

Question 8 (8p)

Group the time to marriage into 3 intervals such that the first interval covers the period before age 20, the second interval covers the period between age 20 and 24, and the third interval covers age 25 and older.

a. (2p) Give the number of events (marriages) and exposure years in each of the three intervals. Give also the total events and exposure years.

b. (2p) Without fitting any model, compute the intensities of marriage in each of the three intervals in (a).

c. (4) Fit appropriate model to the data in (a) and compare your results with those obtained in (b).