Stockholm University
Department of Statistics
(Gebrenegus Ghilagaber)

Home-Examination for Analysis of Survival Data with Demographic Applications (ST303G), Basic-level course, 7.5 Credits, Spring 2021.

The exam is handed-out on **Tuesday 13 April 2021 at 09:00** and replies are expected to be handed-in electronically (according to the separate instructions) latest by **Monday 19 April 2021 at 15:00**.

The examination consists of 9 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 should be included as 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.

For questions about the **content** of the exam, contact the course coordinator via email Gebre@stat.su.se. Incoming questions will be answered continuously during the exam period.

If the course coordinator needs to send out information to all students during the exam, this is done to your registered email address. Therefore, check your email during the exam period.

NOTE! The exam shall be submitted electronically via the department's web site **no later than 15.00 (3 pm) on Monday 19 April 2021**. **The system does not allow submission after the deadline** which is a new setup for this semester. Therefore, start the submission well in advance. The last hour of the exam time is intended for arranging the electronic submission.

Please note that practical help is only available during the **first day** of the exam by email to expedition@stat.su.se. Read carefully the enclosed instructions for exam submission. There, you find all the necessary information about submission, anonymous code, extended writing time etc. If you, despite the instructions have problems submitting the exam, email the exam to tenta@stat.su.se. However, this is only done in exceptional cases. Exams sent in by email after deadline will not be corrected.

NOTE! All forms of cooperation and plagiarism are prohibited. We go over all exams carefully to detect cheating. Suspected cheating is reported to the Disciplinary Board and can lead to suspension.

Part 1: Theoretical/Analytical Questions:

Question 1 (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.

Question 2 (6 p)

The Log-rank test for comparing survival experiences in two groups may be viewed as originating from the hyper-geometric distribution where, at each event time, the observations (in each of two groups) may be considered as belonging to events or nonevents.

a. (4p) Derive the corresponding test statistics of the Log-rank test using the properties of the hyper-geometric distribution (especially its expected value and variance).

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

Part 2: Analyses of real-life data set and interpretation of results.

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

The attached file **Surv2021_ReExam.xlsx** contains data on transition to parenthood among a sample of married women. The seven columns represent the following variables:

Column 1: Cluster Number
Column 2: Sub-Cluster Number
Column 3: Birth Cohort with 7 levels (1 indicating the youngest cohort and 7 the oldest cohort)
Column 4: Residence with 3 levels (1: Metropolitan City, 2: Other Towns, 3: Rural area)
Column 5: Education with 4 levels (0: None, 1: Primary, 2: Middle, 3: Secondary or higher)
Column 6: Months since date of marriage.
Column 7: Indicator of transition to parenthood (0: Not yet parent by survey time, 1: Parent)

Question 3 (6p)

a. (3p) Model the probability of experiencing the event as a function of the three covariates (Birth Cohort, Residence, and Education).

b. (3p) Use the estimated parameters in (a) above to estimate the probabilities of experiencing the event for each level of the three covariates.

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(b)?

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 is **exponentially** distributed.
- **b.** (4p) Assume now that duration is **exponentially** distributed but with different parameters for the three levels of *Residence*. Use appropriate procedure to test for the equality of the population-intensities of experiencing the event across the three residences.
- **c.** (2) Are your results in 6(b) in accordance with those in 4(b)? If not, what do you think the reason can be?

Question 7 (7p)

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

- **a.** (4p) 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 (5p)

In what way(s), if any, do you think the first two columns in the data set (Cluster Number and Sub-cluster Number, respectively) be used to improve the analyses in Questions 3 – 7 above?

Question 9 (6p)

Group the time-interval (months since date of marriage) into five intervals such that the first interval covers months 0 – 60, the second covers months 61-120, the third interval covers months 121-180, the fourth interval months 181-240, and the fifth interval months 241 and above.

- **a.** (3p) Without fitting any model, compute the occurrences (occ), exposures (exp), and occ/exp rates (per 1000) in each interval.
- **b.** (3p) Fit appropriate model for the rate of occurrence as a function of the time interval. How do your estimates of occurrence rates (b) compare with those in (a)