Queen's University
STAT 886, Winter 2022

Final Examination

The final exam is due on **Sunday April 24, 2022** by **23:00**.
Please submit to OnQ.

You must work **independently** for this final exam.

If you need to cite existing work (books, articles, on-line resources), please indicate the sources in a list of references.

Part I. Problems (15%)

Define your notation if necessary and give sufficient details when answering the questions.

1. [15 marks] Consider a multiplicative (non-proportional) hazard model where $z(=0,1)$ is a binary covariate indicating two treatment groups, and the hazard function is of the form

$$h(t|z) = h_0(t)\exp[\beta z + \gamma z g(t)],$$

where $g(t)$ is a specified function of time t. It has been shown that the partial likelihood approach is still valid for this model with parameters (β, γ).
(a) Derive a score test of the hypothesis $H_0 : \beta = 0, \gamma = 0$. (Give the test statistic and describe how to use it to test H_0.)
(b) The test in (a) is a two-sample test which, if $g(t)$ is chosen appropriately, can detect hazards differences between the two groups. Describe how you could develop a similar test of the equality of three or more treatment groups.
(c) Explain why testing $H_0 : \gamma = 0$ provides a test of the proportional hazards model assumption (with $h(t|z) = h_0(t)\exp(\beta z)$). Develop a score test for this purpose.

Notes on score tests: Supposed a maximum likelihood (or partial likelihood) method is applied to a model with a p-dimensional parameter θ. Let $U(\theta)$ and $I(\theta)$ be the score function and information matrix respectively. Score tests can be developed based on the following asymptotic results.
(1) For large sample size n, $U(\theta_0) \approx \text{Normal}_p(0, \mathcal{I}(\theta_0))$, or equivalently, $U^{\mathrm{T}}(\theta_0)\mathcal{I}^{-1}(\theta_0)U(\theta_0) \approx \chi^2_p$, where θ_0 is the true value of θ, and $\mathcal{I}(\theta) = E\{I(\theta)\}$ is the Fisher information matrix.
(2) Suppose the parameter space of θ is Θ in $\mathcal{R}^p$. Consider a test for $H_0 : \theta = \theta_0$. Under H_0, the parameter θ takes values in the restricted parameter space Θ_0 in $\mathcal{R}^{p_0}$, with $p_0 < p$. Let $\hat{\theta}$ be the maximum likelihood estimator of θ under H_0. For large sample size n, $U^{\mathrm{T}}(\hat{\theta})I^{-1}(\hat{\theta})U(\hat{\theta}) \approx \chi^2_{p-p_0}$.

Part II. Analysis of Breast Cancer Data (85%)

The data set “br_surv.txt” is posted at OnQ course site. The data are obtained from a randomized Phase III comparative study of vinorelbine combined with doxorubicin (new treatment) versus doxorubicin alone (standard treatment) on patients with disseminated metastatic/recurrent breast cancer.

The **primary** interest of the clinical trial is to study if the new treatment extends patients' survival time compared to the standard treatment.

Some other variables are also collected on the patients. **Detailed descriptions of the variables** are given on the last 2 pages of this document. It is **also important** to conduct exploratory analysis to study how these variables affect the survival time and build an appropriate model (or models) to describe their association.

Analyze the data and write a report for your study and analysis.

Suggestions and Requirements:

- In your analysis, please give some **priorities** to semi-parametric and non-parametric methods such as Kaplan-Meier estimates, (weighted) log-rank tests and the Cox regression model. Support your analysis with appropriate graphical exploration, model checking and residual analysis.
- Good scientific writing and clear explanation are highly valued. You can describe the problems under study, the data and variables; describe your initial exploration and the models you consider, include selected results and figures for model fit, model assessment and comparison; and interpret your final model (or models) and explain what knowledge is obtained from your analysis in the application context. You are encouraged to do some self learning to understand the unfamiliar cancer-related terminologies in the data description file.

  An **example of report** for a course project is posted below.
- Write your report as an article that explains your thoughts, it should not look like patches of analysis output. Your report should **NOT** include code or raw output from R or SAS. Include tables and figures in the report if needed, to present or summarize your explorations, analysis, and results. Please attach the code and output at the end of your report as a record and proof of your **independent** work.
- Please aim for a clear and concise report. The suggested length is no more than **5 pages of text**. Tables and/or figures should be inserted in the report (but do not count for length).

Marking Scheme:
Total marks: 85;
45 marks on statistical analysis;
40 marks on report writing.

The dataset "br_surv.txt" contains the data on the survival times and some baseline characteristics (i.e., information collected at the time when the patient entered the study) of the patients in a phase III clinical trial. The trial aims to compare two treatments: vinorelbine combined with doxorubicin versus doxorubicin alone, for disseminated metastatic/recurrent breast cancer patients.

The definitions of the variables are:

Id: The identification of the patient (characteristic variable);

Survival: Number of days (from randomization) a patient survived for the patient died or time from the randomization to the last contact if the patient is still alive or has lost to follow-up at the time of final analysis;

Dead: =1 if the patient died; =0 if the patient is still alive or has lost to follow-up (censored);

Arm: =0 if treated by Doxorubicin alone; =1 if treated by Vinorelbine combined with Doxorubicin;

Age: Age at randomization (in years);

Perform: Performance status of the patient at randomization (=0 if full active; =1 if restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature; =2 if ambulatory and capable of all self care but unable to carry out any work activities; =3 if capable of only limited self care; =4 if completed disabled);

Meno: =0 if the patient was pre-menopausal at randomization; =1 if post-menopausal;

Measure: =0 if the tumour was not evaluable; =1 if the tumour is evaluable;

Numsites: number of body sites with tumours observed;

Bone: =1 if the location of the tumour was in bone; =0 if not;

Nodal: =0 if nodal involvement of the tumour at 1st diagnosis was negative; =1 if positive;

Er: =0 if the estrogen receptor of the tumour was negative; =1 if positive;

Timdiag: time since the tumour was originally diagnosed (in days);

Disfree: time since the tumour was removed (in days);

Global: The score of the patient's global quality of life at the beginning of the study (0-100);

Adv: =1 if the patient experienced adverse event at the beginning of the study; =0 if not;

Chemmet: =1 if the patient treated by a chemotherapy for metastatic cancer at the beginning of the study; =0 if not;

Regimen: =0 if no prior regimens taken before randomization; =1 if at least one prior regiments;

WBC: Blood WBC count at the beginning of the study;

Gran: Blood granulocyte count at the beginning of the study;

Platelet: Blood platelet count at the beginning of the study.

******
Some variables have values . for some patients, this means the true information is unknown, or missing.

Suggestions about dealing with missing values:

Do not take . as a level of the covariate (factor).

For a variable with missing values, focus on this variable only and check
if it is important in explaining survival time distributions.
If not important, then do not consider the variable in the rest of your
analysis (as if this variable (column) is removed from the data set).
If important, check how many patients have this variable missing,
if only a few (say <2%), then do not include these patients in
your analysis (as if these patients (rows) are removed from the data set).

Imputation is a typical approach for handling missing data, when there is a noticeable proportion of missing values. It is not covered in this course.

Please clearly describe how you deal with the missing values;
feel free to explore and try any approaches you think are reasonable.
******