Applied Statistics Qualifying Exams Coaching

Michael Howes*

Summer 2023

Contents

*With lots of content credit given to previous applied quals coaches

0 Syllabus

- There is no required homework for this course, but it is recommended that you do the 2011-2022 applied qualifying exams for practice.
- Each session will have two parts. First we will review a topic. Next we will discuss the solutions to a particular year's qual.
- A tentative schedule is shown in Table 1. I'll update this schedule. We'll have to re-schedule one of the classes because of the Fourth of July holiday. The topics can also be adjusted based on your preferences.
- In the two weeks before your exams, you should take the 2022 quals in exam conditions. We'll schedule a time to go over the applied exam together.
- Many of you will find that you don't have time to fully write up solutions to every past qual you are asked to do. This is ok, but you will get a lot of reading every exam and at least writing down a sketch of the answer.
- I will update this document as we go along. I will follow Dan Kluger's notes which are already on Canvas.

Session	Date	Review Material	Past qualifying exam
1	6/27	Advice and sample problems	
2	6/29	Linear models	2011
3	7/03	Exponential families and GLMs	2012
4	7/06	The bootstrap	2013
5	7/11	EM	2014
6	7/13	Cross validation	2015
7	7/18	Linear models: additional topics	2016
8	7/20	Bayesian modelling	2017
9	7/25	CAVI	2018
10	7/27	PCA	2019
11	8/01	Survival Analysis	2020
12	8/03	Optimization and numerical linear algebra	2021
13	8/08	Divergence	
No Class	8/10	Come to the ice-cream social!	
14	8/14	Solutions to last-years exam	**2022**

Table 1: Coaching schedule for the applied qualifying exam. This schedule is open to suggestions. The **2022** qual should be done in exact exam conditions. We will schedule a time to go over the 2022 applied exam together.

1 Survival Analysis

Here we will mostly follow the presentation in Section 3.6 of [Efron, 2022]. We will also include some references to past qualifying exams questions.

1.1 Survival functions and hazard rates

Survival analysis is sometimes called "time-to-event" analysis. Our data contains variables $T_i \geq 0$ which record the time at each the "event" occurred for subject i. Historically, T_i was the time at which the ith subject died. However, survival analysis can be used in less morbid applications. The historic development influences the terminology of survival analysis. We will often interpret $T_i \geq t$ as subject i "surviving" to time t or subject i being "at risk" at time t.

Survival functions

An event time T can be thought of as a non-negative random variable. Our goal is to infer the distribution of T for different subjects. We can do this inference in terms of the *survival function* of T. The survival function of T is the function $S_T : [0, \infty) \to [0, 1]$ given by

$$S_T(t) = \mathbb{P}(T \geq t).$$

In the case when T is continuous, the survival function $S_T(t)$ is equal to $1 - F_T(t)$ where F_T is the CDF of the distribution of T. Note that the distribution of T is completely determined by S_T.

Hazard rates

The distribution of T can also be described by T's *hazard rate* $h_T(t)$. This is the chance of the event occurring at time t conditional on surviving to at least time t. For simplicity, we will consider discrete and continuous distributions separately.[1] In these two cases the hazard rate is defined as follows:

$$h_T(t) = \frac{\mathbb{P}(T = t)}{S_T(t)} \quad \text{when } T \text{ is discrete} \tag{1}$$

$$h_T(t) = \frac{f(t)}{S_T(t)} \quad \text{when } T \text{ is continuous with density } f(t) \tag{2}$$

Note that in the discrete case, $h_T(t) = 0$ for all $t \notin \operatorname{supp} T$. For both discrete and continuous distributions, the survival function $S_T(t)$ can be recovered from the hazard

[1] We will not consider the hazard function of a distribution that is a mixture of discrete and continuous.

rate $h_T(t)$. In particular,

$$S_T(t) = \prod_{u<t}(1 - h_T(u)) \quad \text{when } T \text{ is discrete,} \tag{3}$$

$$S_T(t) = \exp\left(-\int_0^t h_T(u)du\right) \quad \text{when } T \text{ is continuous..} \tag{4}$$

The distribution of T is therefore determined by the hazard rate $h_T(t)$. Our modeling and inference can thus be bases on the hazard rates $h_T(t)$. Equations (3) and (4) show that $S_T(t)$ is a *decreasing* function of $h_T(t)$. Higher hazard rates correspond to earlier event times.

1.2 Censoring

A complication in survival analysis that our data is often *right censored*. This means that for some subjects, we do not observe the survival time T_i. For censored subjects, we only know that $T_i \geq C_i$ where C_i is another random time called the censored time. Typically, C_i will be the time at which the study ends measured from when subject i joins the experiment. We will assume our data is in the following form:

$$\begin{aligned} O_i &= \min\{C_i, T_i\} \in [0, \infty) \\ \delta_i &= I_{\{O_i = T_i\}} \in \{0, 1\}, \quad 1 \leq i \leq N. \end{aligned} \tag{5}$$

The variable O_i is the observed time for subject i. It is either equal to the event time T_i or the censored time C_i. The variable δ_i is an indicator with $\delta_i = 1$ meaning that for subject i is not censored.

1.3 Estimation

Consider data as (5) but with the additional assumption that $T_i \overset{\text{iid}}{\sim} T$ and that $(T_i)_{i=1}^N$ are independent of $(C_i)_{i=1}^N$. One of the main tasks in survival analysis is using the data in (5) to estimate the distribution of T.

The Kaplan–Meier estimate

Let $E = \{T_i : 1 \leq i \leq N, \delta_i = 1\}$ be the set of observed event times. For each $t \in E$ defined the following,

- The risk set at time t, $R(t) = \{i : O_i \geq t\}$.
- The number of at risk subjects at time t, $n(t) = |R(t)|$.
- The number of uncensored times equal to t, $y(t) = |\{i : O_i = t, \delta_i = 1\}|$.

Under our i.i.d. assumptions, the conditional distribution of $y(t)$ given $n(t)$ is

$$y(t) \mid n(t) \overset{\text{ind}}{\sim} \text{Binom}(n(t), h_T(t)) \text{ for } t \in E.$$

For $t \in E$, we can estimate $h_T(t)$ with the MLE,

$$\hat{h}(t) = \frac{y(t)}{n(t)}.$$

From (3), we get an estimate of the survival function for T,

$$\hat{S}(t) = \prod_{u \in E, u < t} (1 - \hat{h}_T(u)).$$

This is called the *Kaplan–Meier estimate* of S_T. We can estimate the variance of $\hat{S}_T$ by *Green-wood's formula*

$$\text{Var}\hat{S}(t) \approx \hat{S}(t)^2 \sum_{u \in E, u \leq t} \frac{y(u)}{n(u)(n(u) - y(u))}.$$

Parametric modelling

The MLE $y(t)/n(t)$ can have high variance, especially for large value of t where we expect $n(t)$ to be small. We can reduce the variance of $\hat{S}(t)$ by putting modelling assumptions on the hazard function $h(t)$. Since we have binomial data $y(t) \mid n(t)$, it is natural to use a GLM. Specifically, now assume that

$$T_i \overset{\text{iid}}{\sim} T, \quad \text{Logit}(h_T(t)) = g(t; \theta),$$

where $g(t : \theta)$ is a parametric family of functions on $[0, \infty)$. If we assume that $g(t; \theta)$ can be represented at $g(t; \theta) = \Phi(t)^\top \theta$, then we get a binomial GLM. Specifically, for each $t \in E$,

$$y(t) \mid n(t) \sim \text{Binom}(n(t), p(t)), \quad \text{Logit}(p(t)) = \Phi(t)^\top \theta.$$

We can fit a GLM to get the MLE $\hat{\theta}$ of θ. This gives the following estimates of h_T and S_T,

$$\hat{h}(t) = \text{Logit}^{-1}(\Phi(t)^\top \hat{\theta}),$$
$$\hat{S}(t) = \prod_{u \in E, u < t} (1 - \hat{h}(u)).$$

Standard GLM theory gives the limiting distribution of $\hat{\theta}$. You can then use the delta method to get asymptotic standard deviations for $\hat{h}(t)$ and $\hat{S}(t)$. Here is code to fit such a model in `R`,

```
glm(Y/n ~ Phi, family = binomial(link = "logit"), weights = n)
```

Both the Kaplan–Meier estimate and the parametric GLM estimate come up in Question 6 on the 2015 qualifying exam.

1.4 Comparing different populations

In the previous section, we considered the problem of estimating $h_T(t)$ for one distribution T. A more common problem is comparing the distribution of T for two different populations. Specifically, suppose now that we have two samples,

$$\{(O_i^0, \delta_i^0) : 1 \leq i \leq N_0\},$$
$$\{(O_i^1, \delta_i^1) : 1 \leq i \leq N_1\}.$$

We assume that each sample is the form (5) with the same independence assumptions as before. We wish to test if $T^0 \overset{\text{dist}}{=} T^1$. This can be done by fitting separate hazard rates to each sample as above. You can then use the estimated standard errors to see the two hazard rates are substantially different.

The Log-rank test

The log-rank test is a non-parametric test of the null $T^0 \overset{\text{dist}}{=} T^1$. As before, define the following quantities

- The set of all uncensored event times across the two groups

$$E = \{O_i^0 : 1 \leq i \leq N_0, \delta_i^0 = 1\} \cup \{O_i^1 : 1 \leq i \leq n_1, \delta_i^1 = 1\}.$$

- The number of at risk subjects in each sample at time t, $n_j(t) = |\{1 \leq i \leq N_j : O_i^j \geq t\}|$, $j = 0, 1$.
- The number of uncensored times in each sample equal to t, $y_j(t) = |\{1 \leq i \leq N_j : O_i^j = t, \delta_i^j = 1\}|$.

Under the null hypothesis $T^0 \overset{\text{dist}}{=} T^1$, we have for all $t \in E$,

$$y_j(t) \mid n_j(t) \sim \text{Binom}(n_j(t), h(t)).$$

If we condition on $y_0(t) + y_1(t)$, we can eliminate the unknown parameter $h(t)$. Specifically, if we set $y(t) = y_0(t) + y_1(t)$ and $n(t) = n_0(t) + n_1(t)$, then

$$y_0(t) \mid n_0(t), n_1(t), y(t) \sim \text{Hypergeometris}(n(t), y(t), n_0(t)).$$

We can thus use this conditional distribution to conduct tests of $T^0 \overset{\text{dist}}{=} T^1$. The log-rank test standardizes and combines the values of $y_0(t)$ in the following way,

$$Z := \frac{\sum_{t \in E} y_0(t) - \mathbb{E}[y_0(t) \mid n_0(t), n_1(t), y(t)]}{\sqrt{\sum_{t \in E} \text{Var}(y_0(t) \mid n_0(t), n_1(t), y(t))}} \approx \mathcal{N}(0, 1), \tag{6}$$

where

$$\mathbb{E}[y_0(t) \mid n_0(t), n_1(t), y(t)] = \frac{y(t)n_0(t)}{n(t)} \quad \text{and}$$
$$\mathrm{Var}(y_0(t) \mid n_0(t), n_1(t), y(t)) = y(t)\frac{n_0(t)}{n(t)}\frac{n_1(t)}{n(t)}\frac{n_0(t) + n_1(t) - y(t)}{n(t) - 1},$$

by the hypergeometric result above.

The method of combining $|E|$ hypergeometric statistics as in (6) is called a "Conchran–Mantel–Haenazel test." The log-rank test can also be extended to multiple groups using the Fisher–Yates or multivariate hypergeometric distribution.

1.5 Proportional hazard models

Comparing the hazard functions for two groups can be thought of a regression problem with a binary feature. We'd also like to see how continuous feature affect the hazard times, or how multiple features affect the hazard times. *Proportional hazards models* allow us to do exactly this. Assume that for subjects $1 \leq i \leq N$, we have

- Features $X_i \in \mathbb{R}^p$,
- Observed times $O_i = \min\{C_i, T_i\} \geq 0$,
- Censor indicators $\delta_i = I_{\{O_i = T_i\}} \in \{0, 1\}$.

Our inference will be conditional on the features $X_1, \ldots, X_N$. We will assume the following proportional hazards model for $T_i \mid X_i$,

$$T_i \mid X_i \overset{\text{ind}}{\sim} h(t; X_i) = h_0(t) \exp(X_i^\top \beta). \tag{7}$$

Larger values of $X_i^\top \beta$ thus correspond to larger higher hazard functions and thus earlier event times. The unknown function $h_0(t)$ is called the baseline hazard rate. As before we assume that given the features X_i, C_i and T_i are independent.

Estimation

As before, we will conduct inference by conditioning on the risk sets $R(t)$. This will have the affect of eliminating the baseline hazard $h_0(t)$ which can be thought of as a nuisance parameter. For simplicity, we will assume that there are no ties in the set E.

An importance consequence of the proportional hazards model (7) is the following. For an individual i with $\delta_i = 1$, the probability that they are the member of $R(t_i)$ that dies at time t is

$$\pi_i(\beta \mid R(t_i)) = \frac{\exp(X_i^\top \beta)}{\sum_{k \in R(t_i)} \exp(X_k^\top \beta)}.$$

We then define the *partial likelihood* to be the product of the above factors

$$L(\beta) = \prod_{i:\delta_i=1} \pi_i(\beta \mid R(t_i)) = \prod_{i:\delta_i=1} \frac{\exp(X_i^\top \beta)}{\sum_{k \in R(t_i)} \exp(X_k^\top \beta)},$$

and the corresponding *log-partial likelihood*,

$$\ell(\beta) = \log L(\beta) = \sum_{i:\delta_i=1} X_i^\top \beta - \log \left(\sum_{k \in R(t_i)} \exp(X_k^\top \beta) \right).$$

We then estimate β by maximizing $\ell(\beta)$. That is

$$\hat{\beta} = \operatorname*{argmax}_\beta \ell(\beta).$$

This is tractable since $\ell(\beta)$ is concave in β. We also have

$$\nabla_\beta \ell(\beta) = \sum_{i:\delta_i=1} (X_i - m_i(\beta))$$
$$\nabla^2_\beta \ell(\beta) = \sum_{i:\delta_i=1} V_i(\beta),$$

where

$$m_i(\beta) = \sum_{k \in R(t_i)} \pi_k(\beta \mid R(T_i)) X_k,$$

and

$$V_i(\beta) = \sum_{k \in R(t_i)} \pi_k(\beta \mid R(T_i))(X_k - m_i(\beta))(X_k - m_i(\beta))^\top.$$

The `R` package `survival` is a standard tool for fitting proportional hazards models. Here is code that use this package

```
S <- Surv(O, delta)
coxph(S ~ X)
```

You can also use `glm` style notation,

```
S <- Surv(O, delta)
coxph(S ~ X1 + X2 + X3)
```

Intercepts are not fitted in proportional hazards models. The unknown baseline hazard makes intercepts unidentifiable.

Inference

A lot of theoretical statistics gives the normal approximation:

$$\hat{\beta} \approx \mathcal{N}\left(\beta, \hat{\Sigma}\right),$$

where $\hat{\Sigma} = \left(-\nabla^2_\beta \ell(\hat{\beta})\right)^{-1}$. This normal approximation can be used to test if $C^\top \beta = 0$ for matrix $C \in \mathbb{R}^{p \times k}$. For example, we can test hypothesis of the form $\beta_j = 0$ for some feature j. You can use the Wald, Score or likelihood tests from 300B.

It turns out that the score test is closely related to the log-rank test above. Suppose that we have a single binary feature $X_i \in \{0, 1\}$, and we want to test

$$T_i \mid X_i = 0 \overset{\text{dist}}{=} T_i \mid X_i = 1.$$

If we use the proportional hazards model (7) with X_i as the only covariate, then this is equivalent to testing $\beta = 0 \in \mathbb{R}$. The score test statistic is

$$S_{\text{score}} = \frac{\nabla_\beta \ell(0)^2}{-\nabla^2_\beta \ell(0)}.$$

That is, the score evaluated at 0 divided by the Fisher information at 0. It turns out that S_{score} is exactly equal to Z^2 where Z is as in (6).

References

Bradley Efron. *Exponential Families in Theory and Practice*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2022. doi: 10.1017/9781108773157.