Survival Analysis - Winter 2016

Assignment 1

The assignment is due Wednesday February 1 before the class. The dataset needed in Questions 4-5 can be found from `http://individual.utoronto.ca/osaarela/finrisk82.csv` and the R code for writing the input data files for the example model from `http://individual.utoronto.ca/osaarela/poissonreg.r`.

1. We are interested in the association between ischaemic heart disease (IHD) incidence and daily energy intake. Let us denote the index level of exposure (energy intake less than 2750 kcals per day) as $Z = 1$ and the reference level of exposure (energy intake at least 2750 kcals per day) as $Z = 0$. Further, let X be an age group indicator taking values $X = 0$ (40 - 59) and $X = 1$ (60 - 69). The observed data tabulated by the age group and exposure are given by:

		Person-years	Observed IHD events
$X = 0:$	$Z = 1$	1190	14
	$Z = 0$	1880	9

		Person-years	Observed IHD events
$X = 1:$	$Z = 1$	667.5	14
	$Z = 0$	888.9	8

A regression model fitted to these data gave the output

```
Call:
glm(formula = d ~ z + x + offset(log(y)), family = poisson(link = "log"))

Deviance Residuals:
       1         2         3         4
 0.03754  -0.04643  -0.03729   0.04979

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.3264     0.2790 -19.088  < 2e-16 ***
z             0.8737     0.3077   2.839  0.00453 **
x             0.5982     0.2985   2.004  0.04506 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 12.8375948  on 3  degrees of freedom
```

Olli Saarela, Dalla Lana School of Public Health, 155 College Street, 6th floor, Toronto, Ontario M5T 3M7, Canada. E-mail: olli.saarela@utoronto.ca

```
Residual deviance:  0.0074342  on 1  degrees of freedom
AIC: 22.977

Number of Fisher Scoring iterations: 3
```

(a) Specify the regression model corresponding to the above output.

(b) Write the likelihood function resulting from this model specification and the data.

(c) Show how the regression coefficient for the exposure can be interpreted in terms of a rate ratio.

(d) Based on the fitted model, calculate

   i. the expected numbers of events in each age/exposure category;

   ii. the 10-year risk of IHD for a 40 year old individual with energy intake less than 2750 kcals per day (you may ignore the presence of competing causes).

2. Type I censoring refers to censoring that is predetermined so that anyone who has not yet experienced the outcome event by the end of follow-up period at time τ is censored at τ. Consider a type I censored sample assumed to be generated by a constant hazard rate e^{α}, with censoring time τ common to all n individuals followed up. The observed data are realizations of (T_i, E_i) for $i = 1, \ldots, n$, where $T_i \equiv \min\{\tilde{T}_i, \tau\}$, and $E_i \equiv \mathbf{1}_{\{T_i = \tilde{T}_i\}}$.

   (a) Derive the likelihood expression for α arising from the contributions of the form $P(t_i \leq T_i < t_i + \mathrm{d}t, E_i = e_i; \alpha)$.

   (b) Show that the same expression as in (a) would be obtained by assuming that the latent event times $\tilde{T}_i$ are exponentially distributed.

   (c) Show that the same expression as in (a) would be obtained by assuming that $\sum_{i=1}^{n} e_i$ is Poisson distributed (conditional on $\sum_{i=1}^{n} t_i$).

   (d) Show that $\sum_{i=1}^{n} e_i$, without conditioning on $\sum_{i=1}^{n} t_i$, is Binomially distributed, and write the resulting likelihood expression for α.

3. Continuing from Q2, using the likelihood expression obtained in (a)-(c), derive the maximum likelihood estimator for α, and a standard error for this.

4. The trend for total mortality in the example dataset was decreasing over calendar time. We might be interested whether this decrease is due in particular to decrease in coronary heart disease (CHD) mortality.

   (a) Using the provided dataset and the `glm` function in R, fit an appropriate Poisson regression model for CHD mortality, adjusting for age, sex and region (eastern/western Finland). Use the first calendar year (1982) as the reference category, and present the estimated calendar time trend (log-rate ratios) and corresponding 95% confidence intervals graphically. Has the CHD mortality decreased over time? (Note: the < 35 agegroup does not have any CHD deaths, so you may have to modify the data slightly before fitting the model.)

(b) Fit also a similar model for non-CHD mortality (that is, mortality due to causes other that CHD) and comment on whether you can observe a similar trend there.

5. Instead of log-rate ratios, present the CHD mortality trend over calendar time in terms of the estimated baseline CHD mortality rates, and the corresponding 95% confidence intervals. (For this, you may need to modify the fitted model.)