BIOSTAT III: Survival Analysis for Epidemiologists in R: Take-home examination

Mark Clements

9–18 November, 2020

Contents

Instructions

- **The examination is individual-based: you are not allowed to cooperate with anyone**, although you are encouraged to consult the available literature. The examiner will use Urkund in order to assess potential plagiarism.
- The examination will be made available by noon on Wednesday 18 November 2020 and **the examination is due by 17:00 on Wednesday 25 November 2020**.
- The examination will be graded and results returned to you by Wednesday 2 December 2020.
- The examination is in two parts. To pass the examination, you need to score at least 6/11 for Part 1 focused on rates and general regression modelling and 13/24 for Part 2 on survival analysis.
- Do not write answers by hand: please use Word, LaTeX, Markdown or a similar format for your examination report and submit the report **as a PDF file**.
- Motivate all answers in your examination report. Define any notation that you use for equations. The examination report should be written in English.
- Email the examination report containing the answers **as a PDF file** to gunilla.nilsson.roos@ki.se. **Write your name in the email, but do NOT write your name or otherwise reveal your identity in the document containing the answers.**

Part 1

The **dm** data-frame includes simulated data on all cause mortality rates for those with and without diabetes in Denmark for 1996–2016 (based on the `DMepi` dataset from the `Epi` package). The dataset has the following columns:

sex a factor with levels M, F

A One-year age class, 0–99 years

P Calendar year, 1996–2016

diab Indicator for persons with diabetes (1=yes, 0=no)

Y Person-years

D Number of deaths

R Rates (=D/Y)

Q1

(a) The age-specific mortality rates stratified by sex and diabetes status for those aged 50 years and over are shown in Figure 1. The crude mortality rates by calendar period stratified by sex and diabetes status for ages 70-74 years are shown in Figure 2. Carefully describe the pattern of rates by age, calendar period, sex and diabetes status. (2 pts)

```
library(ggplot2) # ggplot
ggplot(subset(dm, P==2010 & A>=50), aes(x=A,y=R,col=factor(diab))) +
    geom_line() +
    facet_wrap(~sex) + xlab("Age (years)") + ylab("Mortality rate")
```

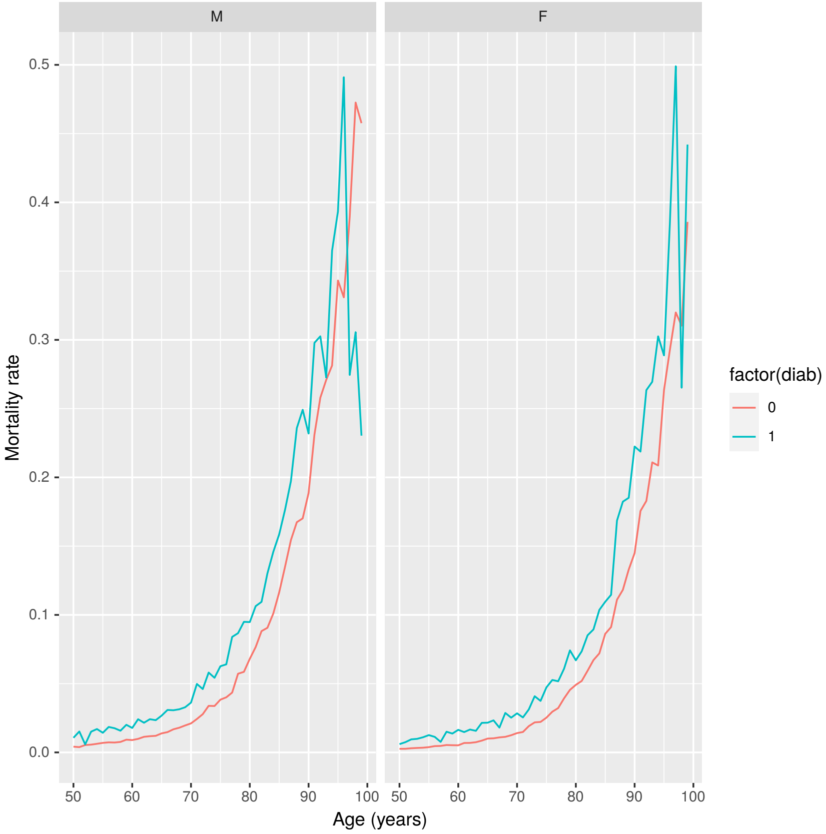

Figure 1: Age-specific mortality rates stratified by sex and presence or absence of diabetes, ages 50 years and over, 2010 calendar year, Denmark.

```
ggplot(subset(dm, A==70), aes(x=P,y=R,col=factor(diab))) +
    geom_line() + ylim(0,0.1) +
    facet_wrap(~sex) + xlab("Age (years)") + ylab("Mortality rate")
```

The following code and output is used to model the mortality rates by diabetes status for males and females separately:

```
fit = glm(D~I(A-70)+I(P-2006)+diab+offset(log(Y)), data=dm, family=poisson,
  subset=(sex=="M"))
summary(fit)
```

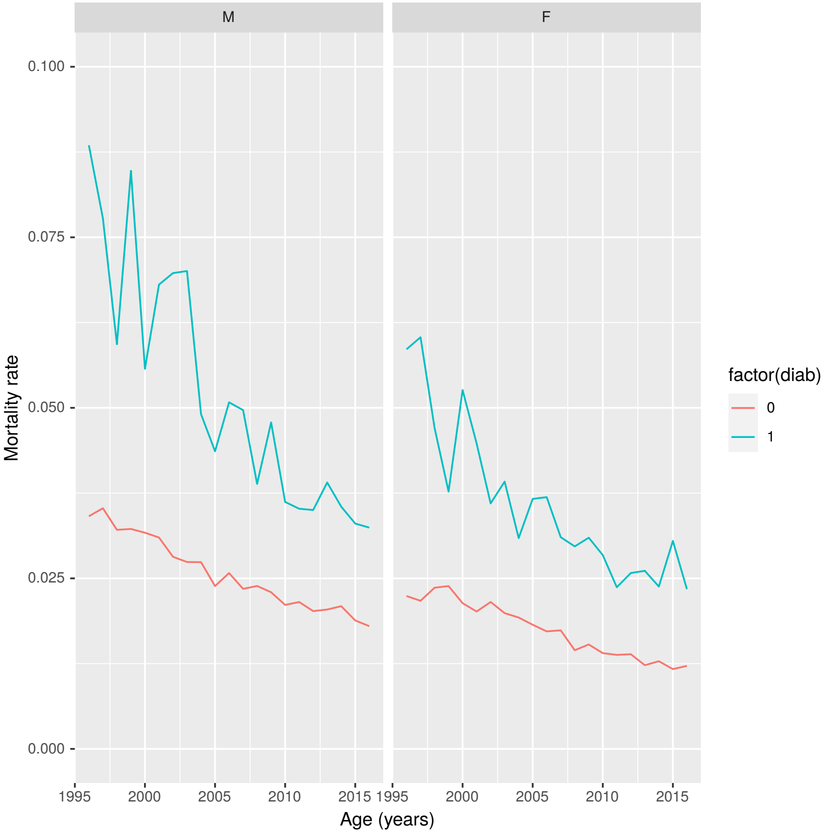

Figure 2: Crude mortality rates stratified by sex and presence or absence of diabetes, ages 70-74 years, Denmark.

```
Call:
glm(formula = D ~ I(A - 70) + I(P - 2006) + diab + offset(log(Y)),
    family = poisson, data = dm, subset = (sex == "M"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.5067  -0.7403   0.0789   1.6214  10.1062

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.581e+00  1.476e-03 -2425.7   <2e-16 ***
I(A - 70)    9.447e-02  9.688e-05   975.1   <2e-16 ***
I(P - 2006) -2.895e-02  2.195e-04  -131.9   <2e-16 ***
diab         5.649e-01  3.641e-03   155.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1606155  on 4196  degrees of freedom
Residual deviance:   16309  on 4193  degrees of freedom
AIC: 37186

Number of Fisher Scoring iterations: 4
```

```
fit = glm(D~I(A-70)+I(P-2006)+diab+offset(log(Y)), data=dm, family=poisson,
  subset=(sex=="F"))
summary(fit)
```

```
Call:
glm(formula = D ~ I(A - 70) + I(P - 2006) + diab + offset(log(Y)),
    family = poisson, data = dm, subset = (sex == "F"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.6231  -0.5669   0.2184   1.5833   9.7857

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.9797988  0.0016456 -2418.5   <2e-16 ***
I(A - 70)    0.1015081  0.0001028   987.2   <2e-16 ***
I(P - 2006) -0.0218066  0.0002171  -100.4   <2e-16 ***
diab         0.5150987  0.0038840   132.6   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1827118  on 4197  degrees of freedom
Residual deviance:   15728  on 4194  degrees of freedom
```

```
AIC: 35513

Number of Fisher Scoring iterations: 4
```

(b) Write out the regression model for males. As a reminder, please explain all of your notation. (2 pts)

(c) What are the mortality rate ratios and 95% confidence intervals for those with diabetes compared with those without diabetes for (i) males and (ii) females? (2 pts)

The following two interaction models can be used to compare the mortality rate ratio of diabetes for males with the mortality rate ratio of diabetes for females.

```
fit2 = glm(D~I(A-70)+I(P-2006)+diab*sex+offset(log(Y)), data=dm, family=poisson)
summary(fit2)


Call:
glm(formula = D ~ I(A - 70) + I(P - 2006) + diab * sex + offset(log(Y)),
    family = poisson, data = dm)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.4248  -0.6975   0.0667   1.5840  10.8122

Coefficients:
              Estimate Std. Error   z value Pr(>|z|)
(Intercept) -3.589e+00  1.469e-03 -2442.520  < 2e-16 ***
I(A - 70)    9.785e-02  7.039e-05  1390.096  < 2e-16 ***
I(P - 2006) -2.540e-02  1.544e-04  -164.546  < 2e-16 ***
diab         5.475e-01  3.620e-03   151.246  < 2e-16 ***
sexF        -3.650e-01  2.048e-03  -178.249  < 2e-16 ***
diab:sexF   -1.737e-02  5.279e-03    -3.291 0.000997 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3433285  on 8394  degrees of freedom
Residual deviance:   35088  on 8389  degrees of freedom
AIC: 75746

Number of Fisher Scoring iterations: 4

fit2 = glm(D~I(A-70)+I(P-2006)+sex+diab:sex+offset(log(Y)), data=dm, family=poisson)
summary(fit2)


Call:
glm(formula = D ~ I(A - 70) + I(P - 2006) + sex + diab:sex +
    offset(log(Y)), family = poisson, data = dm)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.4248  -0.6975   0.0667   1.5840  10.8122

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.589e+00  1.469e-03 -2442.5   <2e-16 ***
I(A - 70)    9.785e-02  7.039e-05  1390.1   <2e-16 ***
I(P - 2006) -2.540e-02  1.544e-04  -164.5   <2e-16 ***
sexF        -3.650e-01  2.048e-03  -178.2   <2e-16 ***
sexM:diab    5.475e-01  3.620e-03   151.2   <2e-16 ***
sexF:diab    5.301e-01  3.872e-03   136.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3433285  on 8394  degrees of freedom
Residual deviance:   35088  on 8389  degrees of freedom
AIC: 75746

Number of Fisher Scoring iterations: 4
```

(d) Write out the regression equation for the second interaction model. (1 pt)

(e) What are the mortality rate ratios and 95% confidence intervals for those with diabetes compared with those without diabetes for (i) males and (ii) females? Why are these estimates different to the estimates in (c)? (2 pts)

(f) Formally test for whether the two mortality rate ratios for males and females in **(e)** are different. Explain how you undertook the test and interpret the findings. (1 pt)

(g) How would you interpret the parameter in the second interaction model `(Intercept)`? (1 pt)

Part 2

Q2

We now use data from the German Breast Cancer Study Group (GBCSG) on a randomised study of hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients (see `https://doi.org/10.1200/JCO.1994.12.10.2086`). The event considered was time to recurrence of breast cancer or death due to breast cancer ("recurrence-free survival"). The main study found no effect associated with the duration of chemotherapy on recurrence-free survival. The help page for the dataset is shown below:

```
library(rstpm2) # brcancer, stpm2
help("brcancer", help_type="text")

brcancer                package:rstpm2                R Documentation

_G_e_r_m_a_n _b_r_e_a_s_t _c_a_n_c_e_r _d_a_t_a _f_r_o_m _S_t_a_t_a.

_D_e_s_c_r_i_p_t_i_o_n:
     See <URL: http://www.stata-press.com/data/r11/brcancer.dta>.

_U_s_a_g_e:

     data(brcancer)

_F_o_r_m_a_t:

     A data frame with 686 observations on the following 15 variables.

     ‘id’ a numeric vector

     ‘hormon’ hormonal therapy

     ‘x1’ age, years

     ‘x2’ menopausal status

     ‘x3’ tumour size, mm

     ‘x4’ tumour grade

     ‘x5’ number of positive nodes

     ‘x6’ progesterone receptor, fmol

     ‘x7’ estrogen receptor, fmol

     ‘rectime’ recurrence free survival time, days

     ‘censrec’ censoring indicator

     ‘x4a’ tumour grade>=2

     ‘x4b’ tumour grade==3

     ‘x5e’ exp(-0.12*x5)

_E_x_a_m_p_l_e_s:

     data(brcancer)
     ## maybe str(brcancer) ; plot(brcancer) ...
```

For data preparation, we categorise for age (x1, with a frequency table) and make a binary indicator for progesterone receptor (x6) to create the data-frame brcancer2:

```
brcancer2 = transform(brcancer,
      x1cat=cut(x1,c(0,45,60,85),right=FALSE), # age groups
      x6ind=ifelse(x6>=20,1,0)) # is progesterone receptor >= 20 nmol?
```

We now define the event time as the time from randomisation to time of recurrence or death

– that is, we are modelling for recurrence-free survival. There were 299 events and the event times are in days from randomisation.

(a) The Kaplan-Meier estimators for the survival functions by progesterone receptor $\geq$ 20 nmol (0=no, 1=yes) are shown in Figure 3. Carefully describe and interpret the two survival curves. (2 pts)

```
library(survival) # survfit, survdiff, Surv, coxph, cox.zph
sfit = survfit(Surv(rectime, censrec==1)~x6ind, data=brcancer2)
plot(sfit, col=1:2, xlab="Recurrence free survival time (days)", ylab="Survival")
legend("topright", paste("x6ind =", 0:1), col=1:2, lty=1, bty="n")
```

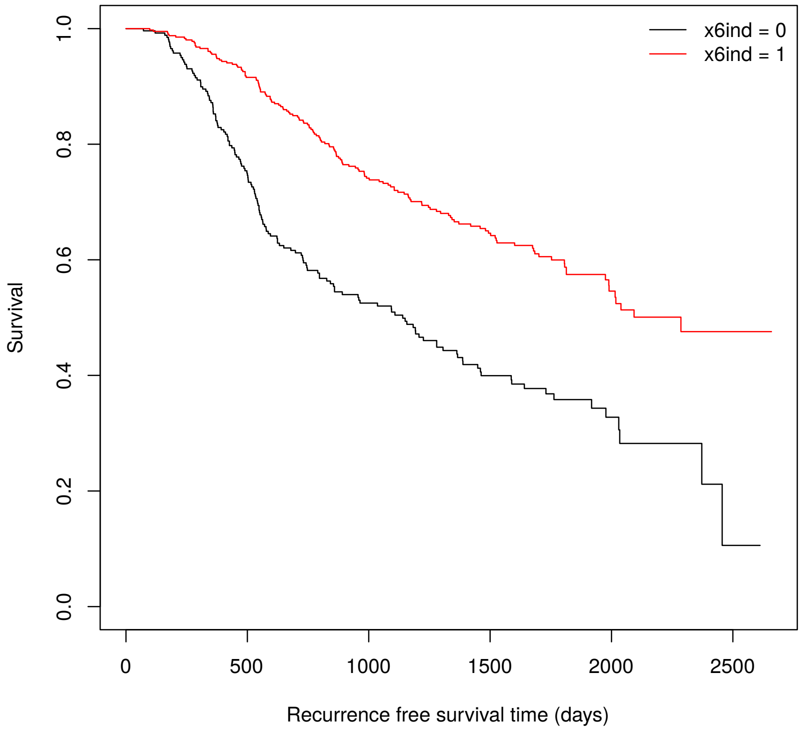

Figure 3: Kaplan-Meier survival curves by progesterone receptor $\geq$ 20 nmol, German Breast Cancer Study Group

(b.i) Write out the regression equation for the Cox model specified in the following code and output. (2 pts)

(b.ii) Based on the following output, discuss whether there is any evidence that progesterone receptor is associated with recurrence-free survival. (2 pts)

```
fit = coxph(Surv(rectime,censrec==1)~x1cat+hormon+x6ind, data=brcancer2)
summary(fit)

Call:
coxph(formula = Surv(rectime, censrec == 1) ~ x1cat + hormon +
    x6ind, data = brcancer2)
```

```
  n= 686, number of events= 299

                coef exp(coef) se(coef)      z Pr(>|z|)
x1cat[45,60) -0.3027    0.7388   0.1502 -2.016  0.04382 *
x1cat[60,85) -0.1359    0.8729   0.1638 -0.829  0.40688
hormon       -0.3523    0.7031   0.1273 -2.767  0.00565 **
x6ind        -0.7701    0.4630   0.1163 -6.624 3.49e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

             exp(coef) exp(-coef) lower .95 upper .95
x1cat[45,60)    0.7388      1.354    0.5504    0.9916
x1cat[60,85)    0.8729      1.146    0.6332    1.2035
hormon          0.7031      1.422    0.5478    0.9023
x6ind           0.4630      2.160    0.3686    0.5814

Concordance= 0.648  (se = 0.016 )
Likelihood ratio test= 56.63  on 4 df,   p=1e-11
Wald test            = 57.77  on 4 df,   p=9e-12
Score (logrank) test = 60.17  on 4 df,   p=3e-12
```

(c) Based on the following Schoenfeld residuals table, is there any evidence for non-proportionality in the modelled covariates? Interpret the table and explain your reasoning. (1 pt)

```
cox.zph(fit)

        chisq df      p
x1cat   3.832  2 0.1472
hormon  0.244  1 0.6212
x6ind   7.180  1 0.0074
GLOBAL 10.622  4 0.0312
```

(d) Based on the following plot of Schoenfeld residuals (Figure 4), how would you expect the hazard ratio for progesterone receptor to vary by time since randomisation? Explain your reasoning. (2 pts)

```
plot(cox.zph(fit)[3])
```

(e) We now fit a flexible parametric survival model adjusting for **x1cat**, **x6ind** and **hormon** (see the following output). How is this model different to the model in **(b)**? (2 pts)

```
fit4 = stpm2(Surv(rectime, censrec==1)˜x1cat+x6ind+hormon, data=brcancer2, df=4)
summary(fit4)

Maximum likelihood estimation

Call:
stpm2(formula = Surv(rectime, censrec == 1) ˜ x1cat + x6ind +
    hormon, data = brcancer2, df = 4)

Coefficients:
```

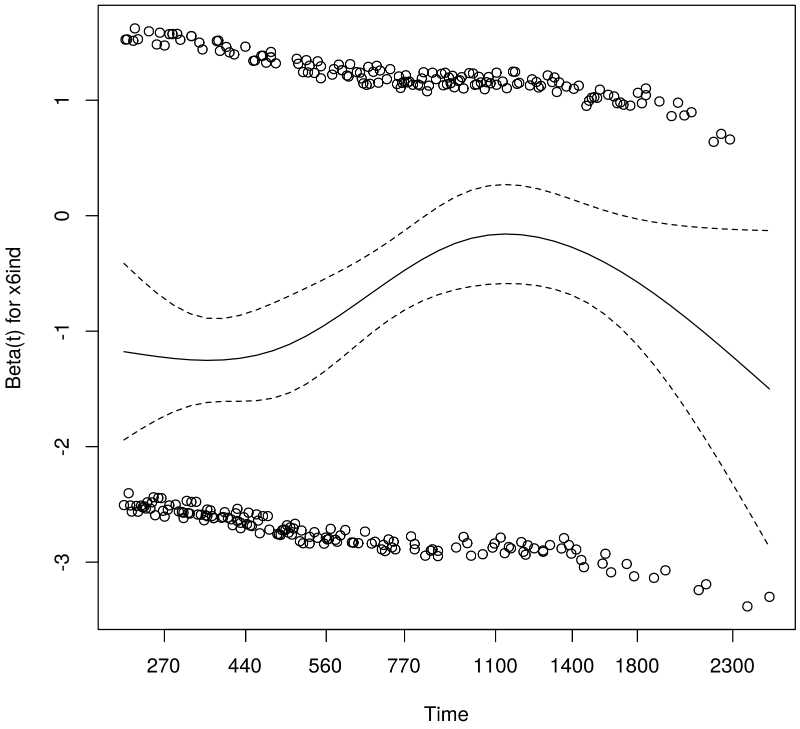

Figure 4: Schoenfeld residual plot for progesterone receptor ≥ 20 nmol, German Breast Cancer Study Group

```
                           Estimate Std. Error z value     Pr(z)
(Intercept)                -6.19938    0.73496 -8.4350 < 2.2e-16 ***
x1cat[45,60)               -0.30442    0.15013 -2.0277  0.042593 *
x1cat[60,85)               -0.13170    0.16372 -0.8044  0.421159
x6ind                      -0.76900    0.11622 -6.6165 3.678e-11 ***
hormon                     -0.35317    0.12728 -2.7748  0.005523 **
nsx(log(rectime), df = 4)1  5.72428    0.71653  7.9889 1.362e-15 ***
nsx(log(rectime), df = 4)2  4.90271    0.48007 10.2126 < 2.2e-16 ***
nsx(log(rectime), df = 4)3 10.16016    1.41344  7.1882 6.564e-13 ***
nsx(log(rectime), df = 4)4  4.79232    0.33088 14.4836 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

-2 log L: 5165.142
```

(f) We now fit a model with time-varying effects for progesterone receptor. We plot the time-varying hazard ratio for progesterone receptor $\geq$ 20 nmol. Carefully interpret the plot in Figure 5. (2 pts)

```
fit5 = stpm2(Surv(rectime, censrec==1)~x1cat+x6ind+hormon, data=brcancer2, df=4,
     tvc=list(x6ind=3))
plot(fit5, type="hr", newdata=data.frame(hormon=0,x1cat="[60,85)",x6ind=0), var="x6ind",
     ylim=c(0,1.5))
```

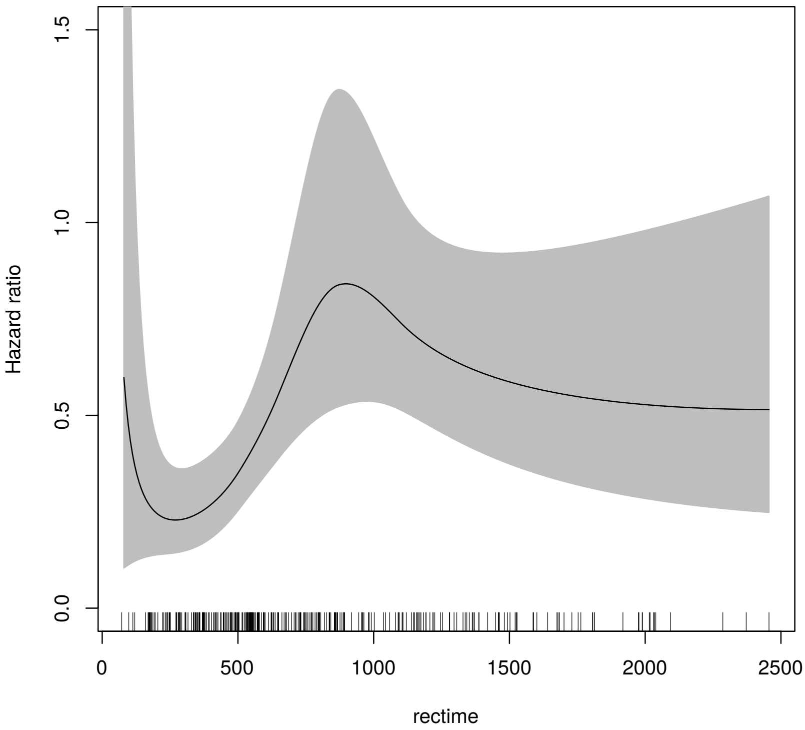

Figure 5: Time-varying hazard ratio for progesterone receptor $\geq$ 20 nmol, German Breast Cancer Study Group

(g) We now present the results as the difference in survival for those with progesterone receptor ≥ 20 nmol compared with those with lower progesterone receptor, assuming no hormonal treatment and ages 60 years and over. Carefully interpret the plot in Figure 6. Is this a marginal or a conditional estimator? (3 pts)

```
plot(fit5, type="sdiff", newdata=data.frame(hormon=0,x1cat="[60,85)",x6ind=0), var="x6ind")
```

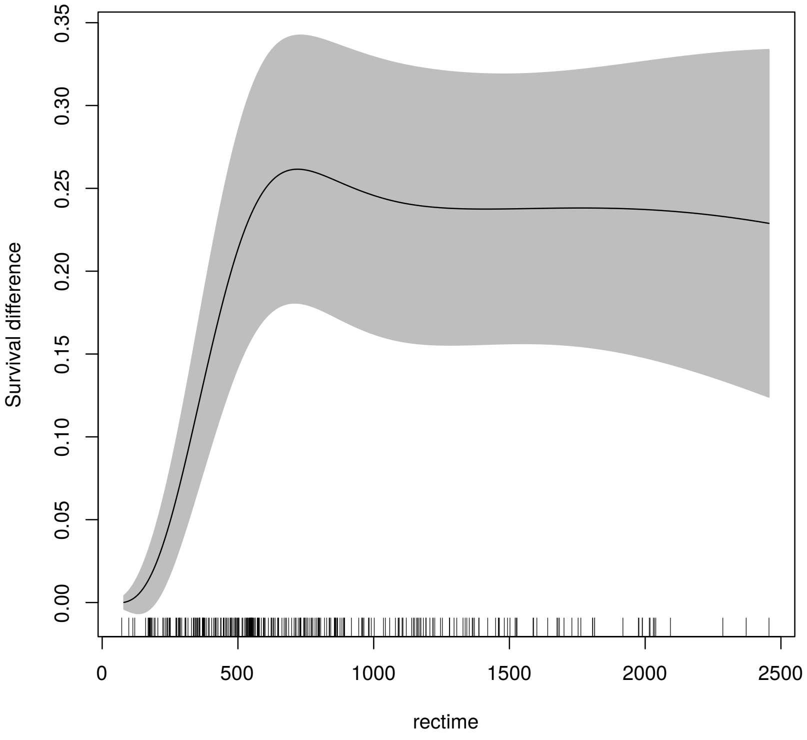

Figure 6: Survival differences for progesterone receptor ≥ 20 nmol compared with < 20 nmol for women aged 60 years and over with no hormonal therapy, German Breast Cancer Study Group

Q3

(a) Consider a randomised study to evaluate the effectiveness of using rehabilitation clinics versus usual care for patients discharged from hospital following a myocardial infarction who survived to 28 days. Effectiveness was measured using all cause mortality with follow-up to five years after study entry. Discuss which time scales you could use for your analysis, describing their advantages and disadvantages. Further describe the data you would collect and which models you would use to analyse those data. (3 pts)

(b) Discuss the difference between left truncation and left censoring. Use examples to explain the differences. (2 pts)

(c) Design an observational cohort study to assess whether aspirin use affects the incidence of prostate cancer. Assume that over-the-counter and prescribed aspirin use can be linked to population and health registers. In your design, you should consider: inclusion and exclusion criteria; entry and exit times; and potential confounders. You should also describe how you would analyse these data, including the primary time scale and which regression model you would use to estimate the target parameter. (3 pts)