C.5 Modern Survival Problem sheet 5: Relative risks and diagnostics

(1) Let $N(t)$ be a counting process with additive hazards $\lambda_i(t) = \lambda_0(t) + \sum_{k=1}^{p} x_{ik}(t)\beta_k(t)$, with $B_k(t) = \int_0^t \beta_k(s)ds$. As in Lecture 12 we define $\mathbf{N}(t)$ to be the vector of the individual counting processes (so it is a binary vector), and similarly $\mathbf{X}(t)$ the matrix of covariates, and $\hat{B}(t)$ the vector of regression coefficient estimators. Define the martingale residual

$$\mathbf{M}_{res}(t) = \int_0^t J(s)d\mathbf{N}(s) - \int_0^t J(s)\mathbf{X}(s)d\hat{\mathbf{B}}(s),$$

where $J(s)$ is the indicator of $\mathbf{X}(s)^T\mathbf{X}(s)$ having full rank, hence of $\mathbf{X}^-(s)$ existing.

(a) Using the fact that

$$J(s)\big(\mathbf{I} - \mathbf{X}(s)\mathbf{X}^-(s)\big)\mathbf{X}(s) \equiv \mathbf{0},$$

show that $\mathbf{M}_{res}$ is a martingale. (That is, every component is a martingale.)
By (9.3) we have that

$$\mathbf{M}(t) := \mathbf{N}(t) - \int_0^t \mathbf{X}(t)d\mathbf{B}(t)$$

is a martingale, and the estimator $\mathbf{B}$ is defined in (9.5) as

$$\hat{\mathbf{B}}(t) = \int_0^t \mathbf{X}^-(s)d\mathbf{N}(s).$$

Thus

$$\begin{aligned}
\mathbf{M}_{res}(t) &= \int_0^t J(s)d\mathbf{N}(s) - \int_0^t J(s)\mathbf{X}(s)\mathbf{X}^-(s)d\mathbf{N}(s) \\
&= \int_0^t J(s)\big(\mathbf{I} - \mathbf{X}(s)\mathbf{X}^-(s)\big)d\mathbf{N}(s) \\
&= \int_0^t J(s)\big(\mathbf{I} - \mathbf{X}(s)\mathbf{X}^-(s)\big)d\mathbf{M}(s) + \int_0^t J(s)\big(\mathbf{I} - \mathbf{X}(s)\mathbf{X}^-(s)\big)\mathbf{X}(s)d\mathbf{B}(s).
\end{aligned}$$

The second term is identically 0, so we have

$$\mathbf{M}_{res}(t) = \int_0^t J(s)\big(\mathbf{I} - \mathbf{X}(s)\mathbf{X}^-(s)\big)d\mathbf{M}(s),$$

which is an integral of martingale increments, hence itself a martingale.

(b) Suppose now that all covariates are fixed and the data are right-censored, and let τ be the final time under consideration (such that $J(\tau) = 1$). Show that

$$\mathbf{X}(0)^T \mathbf{M}_{res}(\tau) = 0.$$

(For time-fixed covariates we define $\mathbf{X}(t) := \mathbf{Y}(t)\mathbf{X}$, where $\mathbf{Y}(t)$ is the matrix with the at-risk indicators $Y_i(t)$ on the diagonal.)
We note that for right-censored data $\mathbf{Y}(s')\mathbf{Y}(s) = \mathbf{Y}(s)$ for $s' \leq s$, and $\mathbf{Y}(s)d\mathbf{N}(s) = d\mathbf{N}(s)$ because $Y_i(s) = 0$ implies that $d\mathbf{N}(s) = 0$. Thus

$$\mathbf{Y}(s)d\mathbf{M}(s) = \mathbf{Y}(s)d\mathbf{N}(s) - \mathbf{Y}(s)\mathbf{X}(s)d\mathbf{B}(s) = d\mathbf{M}(s).$$

For any s with $J(s) = 1$,

$$\begin{aligned}
\mathbf{X}^T d\mathbf{M}_{res}(t) &= \big(\mathbf{X}^T - \mathbf{X}^T\mathbf{Y}(s)\mathbf{X}\mathbf{X}^-(s)\big)d\mathbf{M}(s) \\
&= \big(\mathbf{X}^T - \mathbf{X}^T\mathbf{Y}(s)\mathbf{X}\big(\mathbf{X}^T\mathbf{Y}(s)\mathbf{X}\big)^{-1}\mathbf{X}^T\big)\mathbf{Y}(s)d\mathbf{M}(s) \\
&= \mathbf{X}^T d\mathbf{M}(s) - \mathbf{X}^T\mathbf{Y}(s)d\mathbf{M}(s) \\
&= 0.
\end{aligned}$$

Since this is true for all s, and since $\mathbf{M}_{res}(0) = 0$, it must be true that $\mathbf{X}(0)^T\mathbf{M}_{res}(t) = 0$ for all $t \leq \tau$.

(c) How might this fact be used as a model-diagnostic for the additive-hazards assumption?
The equation $\mathbf{X}^T\mathbf{M}_{res}(\tau) = 0$ means that the n-dimensional vector $\mathbf{M}_{res}(\tau)$ is orthogonal to each of the $p+1$ distinct n-dimensional vectors of the coefficients. There is no linear trend with respect to the covariates. (In other words, in the linear regression model predicting $\mathbf{M}_{res}(\tau)$ as a function of the covariates, the coefficients are all 0.)
If the additive hazards model is true, there should be no nonlinear effect of the covariates on the martingale residuals. So one possible model test is to plot the martingale residuals against nonlinear functions of the residuals — for instance, the square of a covariate, or a product of two covariates — and look for trends. This is described briefly in section 4.2.4 of [ABG08], and more extensively in [Aal93].

(2) Let

$$\begin{aligned}
\mathbf{X}_i(t) &= \text{ vector of observed covariates for individual } i \text{ at time } t;\\
N_i(t) &= \text{ counting process for individual } i \text{ at time } t;\\
\hat{\beta} &= \text{ estimate of Cox regression coefficients;}\\
\hat{A}_0(t) &= \text{ estimate of baseline hazard in Cox model;}\\
\hat{M}_i(t) &= N_i(t) - \int_0^t Y_i(s)e^{\hat{\beta}^T\mathbf{X}_i(s)}d\hat{A}_0(s) \text{ the martingale residuals;}\\
\bar{X}_k(t) &= \frac{\sum_{i=1}^n Y_i(t)X_{ik}(t)e^{\hat{\beta}^T\mathbf{X}_i(t)}}{\sum_{i=1}^n Y_i(t)e^{\hat{\beta}^T\mathbf{X}_i(t)}};\\
U_k(t) &= \sum_{i=1}^n \int_0^t \left[X_{ik}(s) - \bar{X}_k(s)\right] d\hat{M}_i(s).
\end{aligned}$$

U_k is called the *score process*.

(a) Show that

$$U_k(t) = \sum_{t_j \le t} \left(X_{i_j k}(t_j) - \bar{X}_k(t_j)\right).$$

(The summands here are called *Schoenfeld residuals*.)

By definition,

$$\begin{aligned}
U_k(t) &= \sum_{i=1}^n \int_0^t \left[X_{ik}(s) - \bar{X}_k(s)\right] \left(dN_i(s) - Y_i(s)e^{\hat{\beta}\cdot\mathbf{X}_i(s)}d\hat{A}_0(s)\right)\\
&= \sum_{i=1}^n \int_0^t \left[X_{ik}(s) - \bar{X}_k(s)\right] \left(dN_i(s) - Y_i(s)e^{\hat{\beta}\cdot\mathbf{X}_i(s)} \sum_{j=1}^n \frac{e^{\hat{\beta}\cdot\mathbf{X}_j(s)}dN_j(s)}{\sum_{\ell=1}^n e^{\hat{\beta}\cdot\mathbf{X}_\ell(s)}Y_\ell(s)}\right).
\end{aligned}$$

We have

$$\begin{aligned}
&\sum_{i=1}^n \left[X_{ik}(s) - \bar{X}_k(s)\right] Y_i(s)e^{\hat{\beta}\cdot\mathbf{X}_i(s)}\\
&\quad= \sum_{i=1}^n X_{ik}(s)Y_i(s)e^{\hat{\beta}\cdot\mathbf{X}_i(s)} - \sum_{i=1}^n Y_i(s)e^{\hat{\beta}\cdot\mathbf{X}_i(s)} \frac{\sum_{\ell=1}^n Y_\ell(s)X_{\ell k}(s)e^{\hat{\beta}\cdot\mathbf{X}_\ell(s)}}{\sum_{\ell=1}^n Y_\ell(s)e^{\hat{\beta}\cdot\mathbf{X}_\ell(s)}};\\
&\quad= 0.
\end{aligned}$$

Thus

$$\begin{aligned} U_k(t) &= \sum_{i=1}^{n} \int_0^t \big[X_{ik}(s) - \bar{X}_k(s)\big] dN_i(s) \\ &= \sum_{i=1}^{n} \big[X_{ik}(T_i) - \bar{X}_k(T_i)\big] \mathbf{1}_{\{T_i \le t\}} \\ &= \sum_{t_j \le t} \big(X_{i_j k}(t_j) - \bar{X}_k(t_j)\big) . \end{aligned}$$

(b) Show that the score process is the conditional expectation of the partial derivative of the log partial likelihood with respect to the coefficient β_k, conditioned on $\mathcal{F}_t$.
The log likelihood is

$$\ell_P(\beta) = \sum_{t_j} \beta \cdot \mathbf{X}_{i_j}(t_j) - \log \sum_{i=1} n Y_i(t_j) e^{\beta \cdot \mathbf{X}_i(t_j)}.$$

The derivative with respect to β_k is

$$\begin{aligned} \frac{\partial \ell_P}{\partial \beta_k} &= \sum_{t_j} \mathbf{X}_{i_j k}(t_j) - \frac{\sum_{i=1}^n Y_i(t_j) X_{ik}(t_j) e^{\beta \cdot \mathbf{X}_i(t_j)}}{\sum_{i=1}^n Y_i(t_j) e^{\beta \cdot \mathbf{X}_i(t_j)}} \\ &= \sum_{t_j} \big[\mathbf{X}_{i_j k}(t_j) - \bar{X}_k(t_j)\big] \\ &= \sum_{i=1}^{n} \int_0^{\infty} \big[\mathbf{X}_{ik}(s) - \bar{X}_k(s)\big] dN_i(s). \end{aligned}$$

The conditional expectation with respect to $\mathcal{F}_t$ is then

$$\mathbb{E}\left[\frac{\partial \ell_P}{\partial \beta_k} \;\middle|\; \mathcal{F}_t\right] = \sum_{t_j \le t} \big[\mathbf{X}_{i_j k}(t_j) - \bar{X}_k(t_j)\big] + \mathbb{E}\left[\sum_{t_j > t} \big[\mathbf{X}_{i_j k}(t_j) - \bar{X}_k(t_j)\big] \;\middle|\; \mathcal{F}_t\right].$$

If we condition on an event at time $t_j > t$, since i_j is distributed among the elements of $\{1, \ldots, n\}$ in proportion to $Y_i(t_j) e^{\beta \cdot \mathbf{X}_i(t_j)}$,

$$\mathbb{E}\left[\mathbf{X}_{i_j k}(t_j) - \bar{X}_k(t_j) \;\middle|\; \mathcal{F}_{t_j -}\right] = \sum_{i=1}^{n} \mathbf{X}_{ik}(t_j) \frac{Y_i(t_j) e^{\beta \cdot \mathbf{X}_i(t_j)}}{\sum_{i'} Y_{i'}(t_j) e^{\beta \cdot \mathbf{X}_{i'}(t_j)}} - \bar{X}_k(t_j) = 0$$

for all j and k. (We are here conditioning on the past up to a random stopping time t_j, something that was mentioned in section 2.1.10, but not formally introduced.) Since

conditioning on the smaller σ-algebra $\mathcal{F}_t$ may be achieved by first conditioning on the larger, and then on the smaller, by formula 2.5, we see that all the conditional expectations for $t_j > t$ contribute 0 to the sum. Thus, we are left with only the first term

$$\mathbb{E}\left[\frac{\partial \ell_P}{\partial \beta_k} \;\middle|\; \mathcal{F}_t\right] = \sum_{t_j \le t} \left[\mathbf{X}_{i_j k}(t_j) - \bar{X}_k(t_j)\right] = U_k(t).$$

(c) Conclude that $U_k(0) = U_k(\infty) = 0$.
$U_k(0)$ is trivially 0, by definition. $\hat{\beta}$ is chosen to satisfy the equation $\partial \ell_P / \partial \beta_k = 0$. This is the same as $U_k(\infty) = 0$.

(d) Explain why a plot of $U_k(t)$, suitably scaled, would be expected to look like a random walk conditioned to start and end at 0 (a *discrete bridge*) if the proportional hazards assumption holds.
The function starts at 0 and ends at 0. Since it is a martingale, it will behave like a time-changed Brownian motion (by the martingale CLT), except for being conditioned to end at 0.

(3) The dataset `larynx` in the package `KMsurv` includes times of death (or censoring by the end of the study) of 90 males diagnosed with cancer of the larynx between 1970 and 1978 at a single hospital. One important covariate is the stage of the cancer, coded as 1,2,3,4.

(a) Why would it probably not be a good idea to fit the Cox model with relative risk $e^{\beta \cdot \text{stage}}$?
That would treat the categorical variable as though it were quantitative. That would force the relative risks into particular proportions that have no empirical basis. There may be good reason to expect the relative risk to increase with stage, but not to expect particular proportions.

(b) Use a martingale residual plot to show that `stage` does not enter as a linear covariate.
We could fit the model without any covariates — so just find the Nelson–Aalen estimator— and use that as a basis for adding in the stage as a covariate and checking the martingale residuals. Here we will use age as an additional covariate. So we will fit the model $\alpha_i(t) = \alpha_0(t) e^{\beta \cdot \text{age}}$, and check for the behaviour of `stage` as an additional covariate. We show a box plot in figure C.3, showing the distributions of martingale residuals for the 4 different stages. What we see is that the residuals have essentially the same mean for stages 1 and 2, rise substantially for stage 3, and somewhat less for stage 4.

```
require(survival)
require(KMsurv)

data(larynx)
lar.cph=coxph(Surv(time,delta)~age,data=larynx)

       coef exp(coef) se(coef)     z    p
age  0.0233      1.02   0.0145  1.61 0.11

Likelihood ratio test=2.63  on 1 df, p=0.105  n= 90, number of events= 50

lar.fit=survfit(lar.cph)

# The coxph object has a list of times
#   We want to find the index of the time corresponding to individual i.
whichtime=sapply(larynx$time,function(t) which(lar.fit$time==t))

cumhaz=-log(lar.fit$surv[whichtime])

beta=lar.cph$coefficients
relrisk=exp(beta*(larynx$age-mean(larynx$age)))
# Baseline hazard is for mean value of covariate

resids=larynx$delta-cumhaz*relrisk
#Note: We could get the same numbers out as lar.cph$residuals
resids.bystage=lapply(1:4,function(i) resids[larynx$stage==i])
boxplot(resids.bystage,xlab='Stage',ylab='Martingale residual')
```

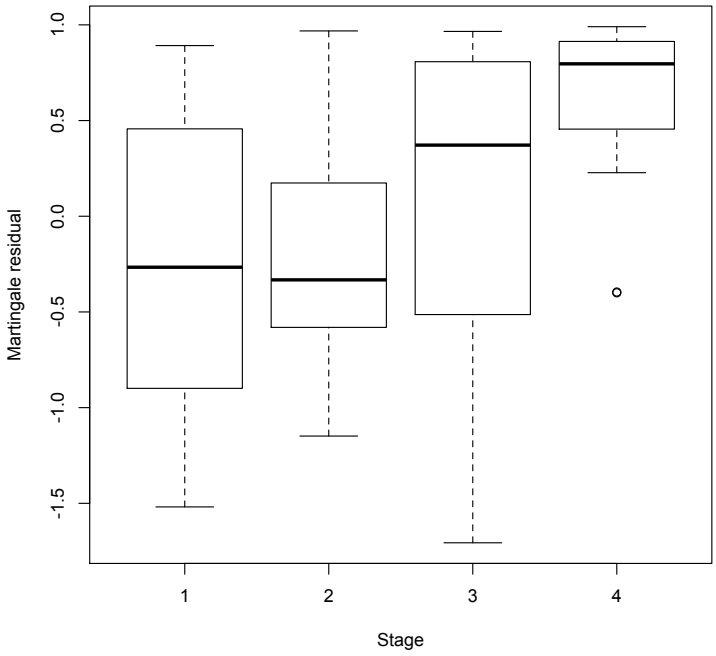

Figure C.3: Box plot of martingale residuals for `larynx` data, stratified by `stage`.

(c) The alternative is to define three new binary covariates, coding for the patient being in stage 2, 3, or 4 respectively (leaving stage 1, where all three covariates are 0, as the baseline group). Fit this model. Are all of these covariates statistically significant?

(d) An equivalent approach is to replace `stage` in the model definition by `factor(stage)`. Show that this produces the same result.
The `R` computation below shows that the coefficient for stage 2 is clearly not statistically significant; the coefficient for stage 3 is borderline ($p = 0.083$); and the coefficient for stage 4 is highly significant ($p = 0.000035$).

```
lar.cph=coxph(Surv(time,delta)~factor(stage)+age,data=larynx)
                 coef exp(coef) se(coef)     z       p
factor(stage)2 0.140      1.15   0.4625 0.303 7.6e-01
factor(stage)3 0.642      1.90   0.3561 1.804 7.1e-02
factor(stage)4 1.706      5.51   0.4219 4.043 5.3e-05
age            0.019      1.02   0.0143 1.335 1.8e-01

Likelihood ratio test=18.3  on 4 df, p=0.00107  n= 90, number of events= 50
```

(e) Try adding year of diagnosis or age at diagnosis as a linear covariate (in the exponent of the relative risk). Is either statistically significant?

In the below code we fit the model including age and stage. Again, only the coefficient for stage 4 is significantly greater than 0.

(f) Use a residual plot to test whether one or the other of these covariates might more appropriately enter the model in a different functional form — for example, as a step function.

The plot is shown in Figure C.4. We see that there seems to be no effect of the `age` variable until age 70, after which it seems to increase linearly.

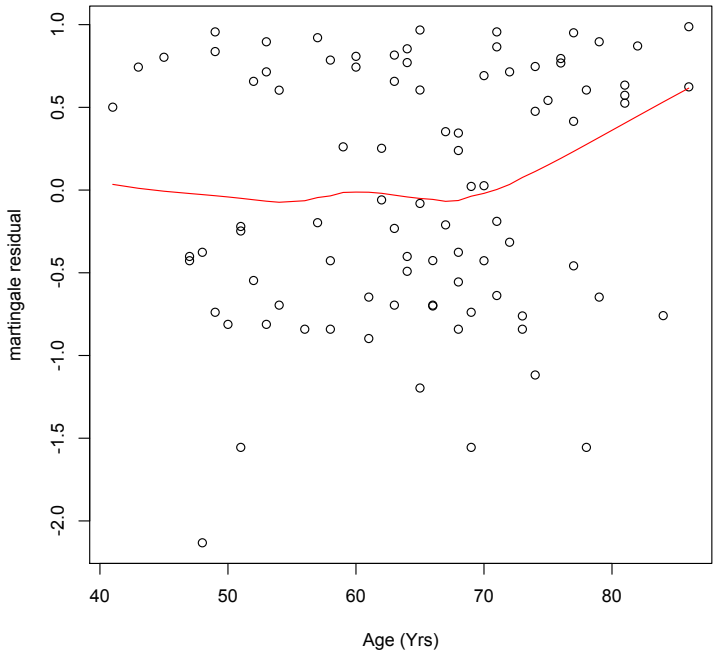

Figure C.4: Plot of martingale residuals against `age` for `larynx` data.

```
########## Residual plot to test age
aord=order(age)
resids=lar.cph2$residuals[aord]
plot(age[aord],resids,xlab='Age (Yrs)',ylab='martingale residual')
lines(lowess(resids~age[aord]),col=2)

########## New model with age starting from 70
newage=pmax(age[aord]-70,0)
lar.cph=coxph(Surv(time,delta)~factor(stage)+newage,data=larynx)
```

(g) Use a Cox-Snell residual plot to test whether the Cox model is appropriate to these data.
There seems to be a marked curvature of the residual plot, suggesting that the model is underestimating the cumulative hazard later on.

```
lar.cph=coxph(Surv(time,delta)~factor(stage),data=larynx)
lar.fit=survfit(lar.cph)

whichtime=sapply(larynx$time,function(t) which(lar.fit$time==t))

cumhaz=-log(lar.fit$surv[whichtime])

beta=lar.cph$coefficients
relrisk=exp(matrix(beta,1,3)%*%rbind(st2-mean(st2),st3-mean(st3),st4-mean(st4)))

coxsnell=c(relrisk*cumhaz)

CS.surv=Surv(coxsnell,delta[aord])
CS.fit=survfit(CS.surv~1)

plot(CS.fit$time,-log(CS.fit$surv),xlab='Time',
ylab='Fitted cumulative hazard for Cox-Snell residuals')
abline(0,1,col=2)
```

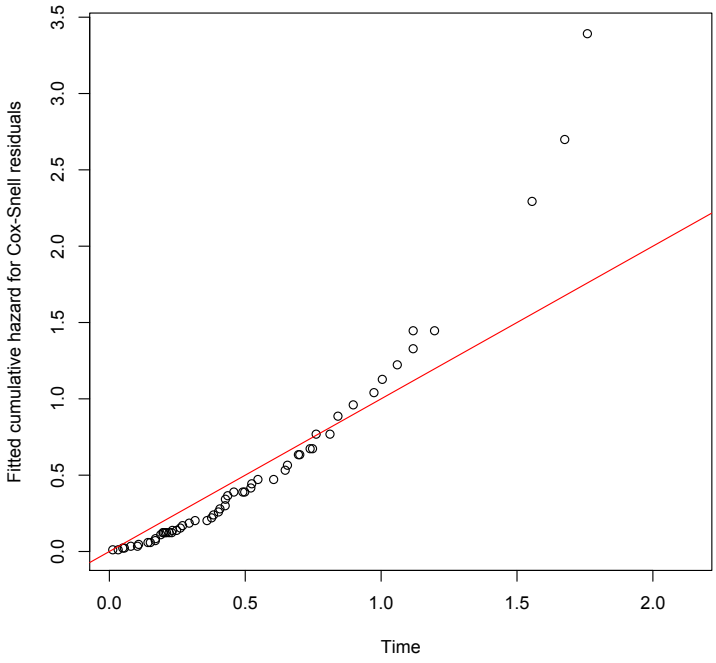

Figure C.5: Cox-Snell residual plot for `larynx` data.