Prostate survival data

```
##    dtime                   status             rx
## 1     72                    alive 0.2 mg estrogen
## 2      1           dead - other ca 0.2 mg estrogen
## 3     40   dead - cerebrovascular 5.0 mg estrogen
## 4     20   dead - cerebrovascular 0.2 mg estrogen
## 5     65                    alive        placebo
## 6     24      dead - prostatic ca 0.2 mg estrogen
## 7     46 dead - heart or vascular        placebo
## 8     62                    alive        placebo
## 9     61                    alive 1.0 mg estrogen
## 10    60                    alive 1.0 mg estrogen
```

Survival distributions depending on treatment

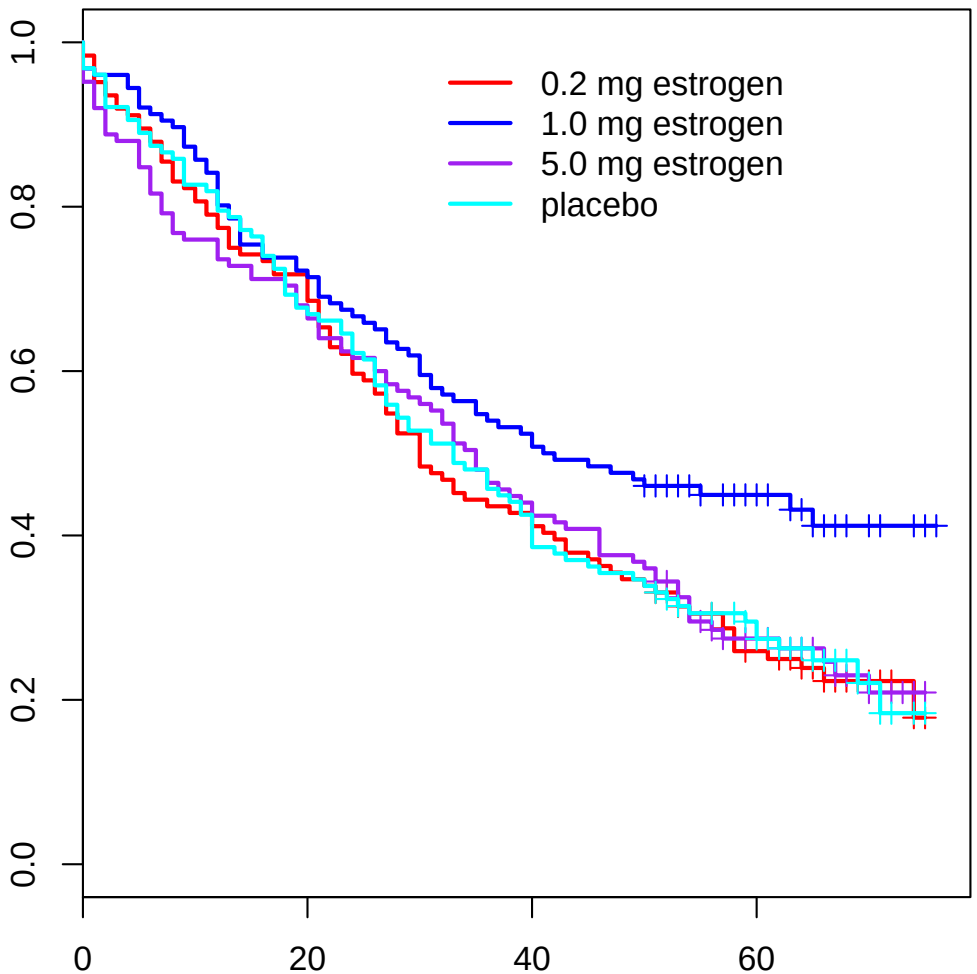
```
prostateSurv <- survfit(Surv(dtime, status != "alive") ~ rx,
                        data = prostate)
```

```
prostateSurv2 <- survfit(Surv(dtime, status != "alive") ~ rx,
                         data = prostate,
                         type = "fleming-harrington")
```

The Fleming-Harrington (or Breslow) estimator of the survival function is $e^{-\hat{\Lambda}(t)}$, where $\hat{\Lambda}(t)$ is the Nelson-Aalen estimator of the cumulative hazard function.

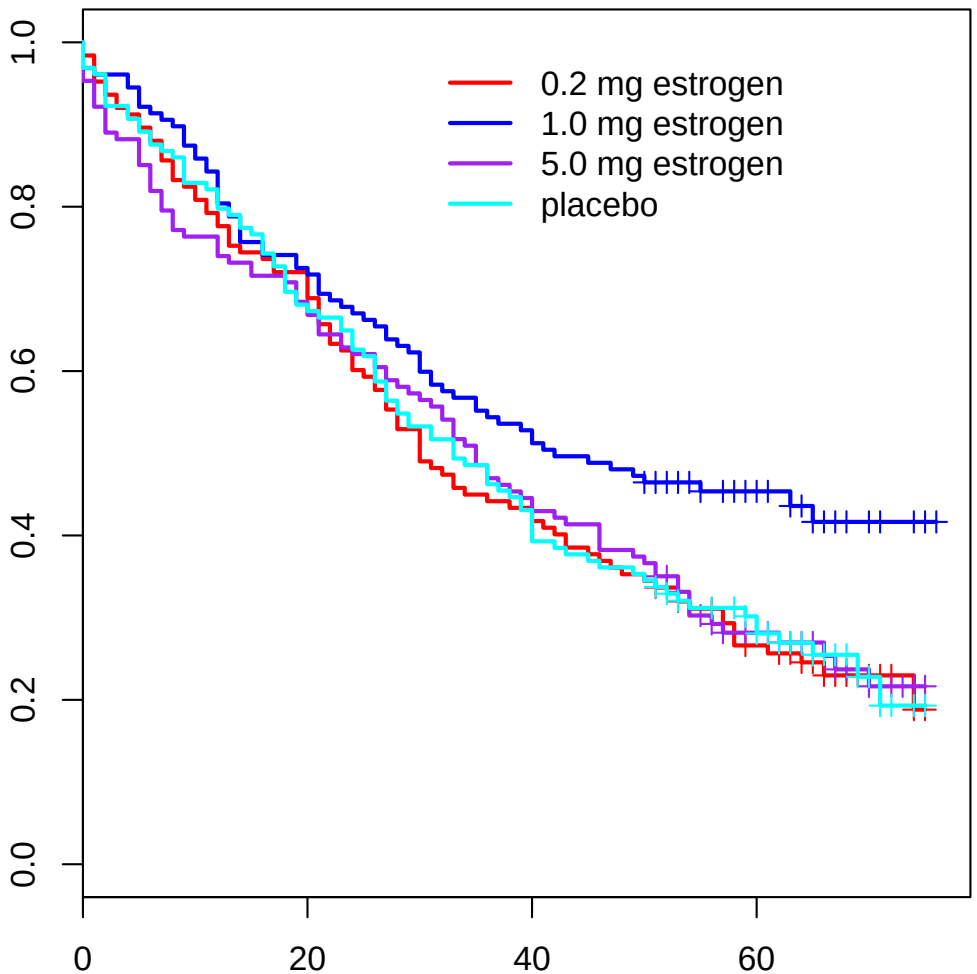
Survival distributions depending on treatment

```
plot(prostateSurv, conf.int = FALSE,
     col = c("red", "blue", "purple", "cyan"), lwd = 2)
```



Survival distributions depending on treatment

```
plot(prostateSurv2, conf.int = FALSE,
     col = c("red", "blue", "purple", "cyan"), lwd = 2)
```



Ties

The Kaplan-Meier and Nelson-Aalen estimators work with tied survival times by replacing the indicator e_i with the increment $\Delta N(T_i)$;

$$\hat{S}(t) = \prod_{s \leq t} \left(1 - \frac{\Delta N(s)}{Y(s)}\right)$$

and

$$\hat{\Lambda}(t) = \sum_{s \leq t} \frac{\Delta N(s)}{Y(s)}.$$

The `survfit` function supports argument `type = "fh2"`, which replaces

$$\frac{\Delta N(s)}{Y(s)}$$

in the Nelson-Aalen estimator by

$$\frac{1}{Y(s)} + \frac{1}{Y(s) - 1} + \ldots + \frac{1}{Y(s) - \Delta N(s) + 1}.$$

Treatment effect

```
survdiff(Surv(dtime, status != "alive") ~ rx,
         data = prostate)

...
##                         N Observed Expected (O-E)^2/E (O-E)^2/V
## rx=0.2 mg estrogen    124       95     84.9     1.212     1.626
## rx=1.0 mg estrogen    126       71     95.9     6.479     9.072
## rx=5.0 mg estrogen    125       93     85.6     0.644     0.867
## rx=placebo            127       95     87.6     0.619     0.839
##
##  Chisq= 9.1  on 3 degrees of freedom, p= 0.0275
```

Proportional hazards model

Assume that

$$\lambda_i(t) = \lambda_0(t)e^{\beta_{\mathrm{rx}}}$$

for a baseline hazard function λ_0 and β_{rx} the log-hazard ratio of treatment rx.

```
prostateCox0 <- coxph(Surv(dtime, status != "alive") ~ rx,
                      data = prostate)
```

The formula specification above uses 0.2 mg estrogen as reference in the dummy variable encoding.

Proportional hazards model

```
summary(prostateCox0)

...
##                      coef exp(coef) se(coef)     z Pr(>|z|)
## rx1.0 mg estrogen -0.4176    0.6586   0.1570 -2.66   0.0078
## rx5.0 mg estrogen -0.0292    0.9712   0.1459 -0.20   0.8413
## rxplacebo         -0.0321    0.9684   0.1451 -0.22   0.8248
...
## Likelihood ratio test= 9.71  on 3 df,   p=0.0212
## Wald test            = 9.01  on 3 df,   p=0.0292
## Score (logrank) test = 9.13  on 3 df,   p=0.0277
```

Proportional hazards model

```
contrasts(prostate$rx) <- "contr.SAS"
prostateCox1 <- coxph(Surv(dtime, status != "alive") ~ rx,
                      data = prostate)
summary(prostateCox1)


...
##                       coef exp(coef) se(coef)     z Pr(>|z|)
## rx0.2 mg estrogen  0.03212   1.03265  0.14512  0.22    0.825
## rx1.0 mg estrogen -0.38552   0.68010  0.15696 -2.46    0.014
## rx5.0 mg estrogen  0.00291   1.00292  0.14594  0.02    0.984
...
## Likelihood ratio test= 9.71  on 3 df,   p=0.0212
## Wald test            = 9.01  on 3 df,   p=0.0292
## Score (logrank) test = 9.13  on 3 df,   p=0.0277
```

Some of the main causes of death

```
tmp <- table(prostate$rx, prostate$status)[, c(1, 2, 3, 6, 7)]
colnames(tmp) <- c("alive",
                   "cer",    # cerebrovascular
                   "heart",  # heart or vascular
                   "prost",  # prostatic cancer
                   "pulm")   # pulmonary embolus
tmp
```

```
##
##                   alive cer heart prost pulm
##   0.2 mg estrogen    29   6    19    42    2
##   1.0 mg estrogen    55  11    14    24    3
##   5.0 mg estrogen    32   7    36    27    7
##   placebo            32   7    27    37    2
```

Potential explanation: Estrogen treatment slows prostate cancer growth. In large dosages estrogen may have side effects, e.g. increasing the risk of death due to blood clots.

Score and information

With $w_i = e^{X_i^T \beta}$ the partial log-likelihood equals

$$\ell_{\text{par}} = \sum_{i:e_i=1} X_i^T \beta - \log \Bigg(\underbrace{\sum_{j:T_i \leq T_j} w_j}_{W_i} \Bigg),$$

with score function

$$\mathcal{U}(\beta) = D\ell_{\text{par}} = \sum_{i:e_i=1} X_i^T - \frac{\sum_{j:T_i \leq T_j} X_j^T w_j}{W_i}$$

and observed information

$$\mathcal{I}(\beta) = -D^2\ell_{\text{par}} = \sum_{i:e_i=1} \frac{\sum_{j:T_i \leq T_j} X_j X_j^T w_j}{W_i^2} - \frac{(\sum_{j:T_i \leq T_j} X_j w_j)(\sum_{j:T_i \leq T_j} X_j w_j)^T}{W_i}.$$

Score and information

The observed information, $\mathcal{I}(\hat{\beta})$, evaluated in the estimated parameter $\hat{\beta}$ is used for quadratic approximations of the ℓ_{par}, for standard error estimates and Wald tests.

The score test

$$\mathcal{U}^T(0)\mathcal{I}(0)^{-1}\mathcal{U}(0)$$

for testing $\beta = 0$ is the log-rank test for a single categorical predictor.

The derivation does not deal explicitly with ties, but treat ties implicitly via the summation over $j : T_i \leq T_j$. This is the Breslow method. A naive implementation using `order` will break ties in an arbitrary (depending on the sorting algorithm) way.

Naive computation of the score

```
orddtime <- order(prostate$dtime)
ei <- (prostate$status != "alive")[orddtime]
X <- model.matrix(prostateCox1)[orddtime, ]
n <- nrow(X)
Y <- length(orddtime):1
U0 <- c(0, 0, 0)  ## Score
I0 <- matrix(0, 3, 3)  ## Information
for(i in 1:length(Y)) {
  if (ei[i]) {
  xx <- X[i:n, ]
  cxx <- colSums(xx)
   U0 <- U0 + X[i, ] - cxx / Y[i]
   I0 <- I0 + (crossprod(xx) / Y[i] - cxx %o% cxx / Y[i]^2)
  }
}
## Recall the coxph score test: 9.13
crossprod(U0, solve(I0, U0))
```

```
##       [,1]
## [1,] 9.027
```

Correct Breslow method

```
Y0 <- rank(prostate$dtime[orddtime], ties.method = "min")
tmp <- outer(Y0, 1:length(Y0), "<=")[ei, ]
X0 <- X[ei, ]
Y <- rowSums(tmp)
U <- c(0, 0, 0)
I <- matrix(0, 3, 3)
for(i in 1:length(Y)) {
  xx <- X[tmp[i, ], ]
  cxx <- colSums(xx)
   U <- U + X0[i, ] - cxx / Y[i]
   I <- I + (crossprod(xx) / Y[i] - cxx %o% cxx / Y[i]^2)
}
crossprod(U, solve(I, U))

##       [,1]
## [1,] 8.966
```

The Breslow method using coxph

```
prostateCox2 <- coxph(Surv(dtime, status != "alive") ~ rx,
                      data = prostate,
                      ties = "breslow")
summary(prostateCox2)
```

```
...
##                       coef exp(coef) se(coef)     z Pr(>|z|)
## rx0.2 mg estrogen   0.0322    1.0327   0.1451  0.22    0.824
## rx1.0 mg estrogen -0.3821    0.6824   0.1570 -2.43    0.015
## rx5.0 mg estrogen   0.0024    1.0024   0.1459  0.02    0.987
...
## Likelihood ratio test= 9.53  on 3 df,   p=0.023
## Wald test            = 8.85  on 3 df,   p=0.0313
## Score (logrank) test = 8.97  on 3 df,   p=0.0298
...
```

```
prostateCox2$score
```

```
## [1] 8.966
```

Ties

In the partial likelihood the weights for the uncensored survival times enter via the factors

$$\frac{w_i}{W_i}.$$

If $T_i = T_j$ are uncensored the denominators $W_k = W_i$ are equal and their contribution is the factor

$$\frac{w_i w_k}{W_i^2}.$$

If the tie is due to rounding or grouping, their contribution should have been

$$\frac{w_i w_k}{W_i(W_i - w_i)} \quad \text{or} \quad \frac{w_i w_k}{W_k(W_k - w_k)}. \tag{1}$$

The naive implementation breaks ties in a unsystematic way, choosing one of the factors arbitrarily.

Ties – Efron's method

Efron's method uses the approximation

$$\frac{w_i w_k}{W_i(W_i - (w_i + w_k)/2)} = \frac{w_i w_k}{W_k(W_k - (w_i + w_k)/2)}$$

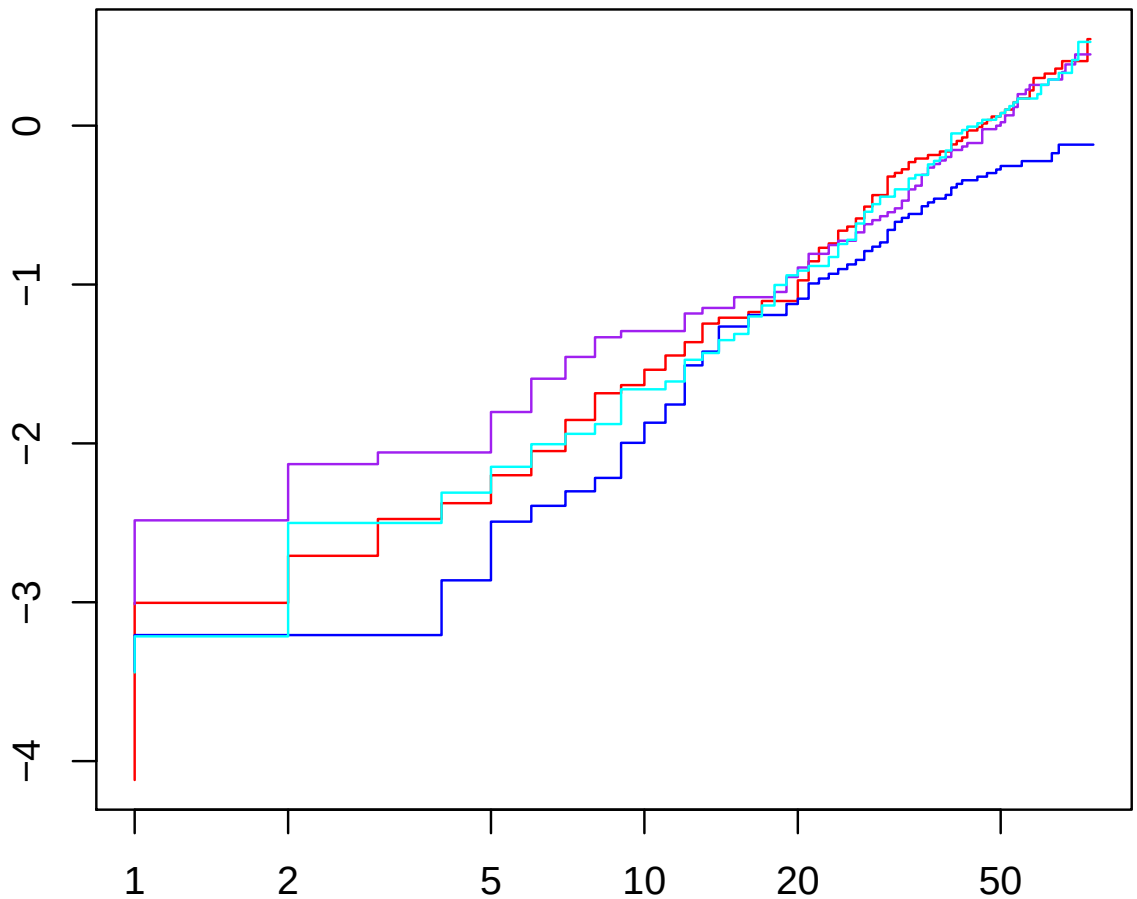
and similar formulas for more than two ties. It is close to the geometric mean of the factors (1).

Efron's method is the default in `coxph`.

The results rarely depend much on whether the Efron or the Breslow (or the naive) method is used, but Efron's method is preferred if the ties are due to a lack of precision and not to a true discrete nature of the survival times.

Checking proportional hazards assumption

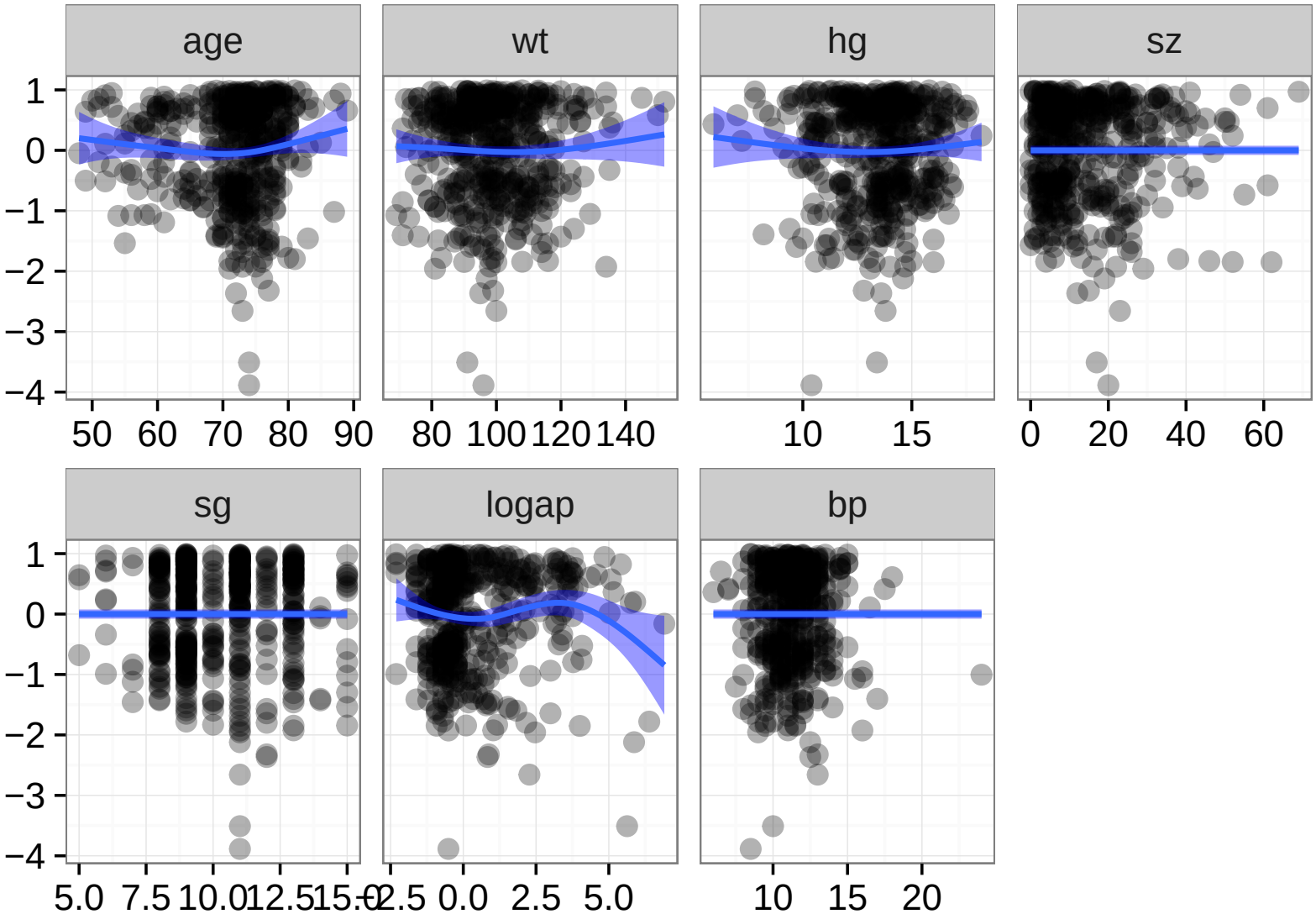
```
par(mar = c(2, 2, 1, 1))
plot(prostateSurv, mark.time = FALSE, conf.int = FALSE,
     col = c("red", "blue", "purple", "cyan"), fun = "cloglog")
```



A prognostic model

```
form <- Surv(dtime, status != "alive") ~ rx + age + wt + pf +
  hx + ekg + hg + sz + sg + logap + bp + bm
prostateCox <- coxph(form, data = subProstate)
```

Martingale residual plot

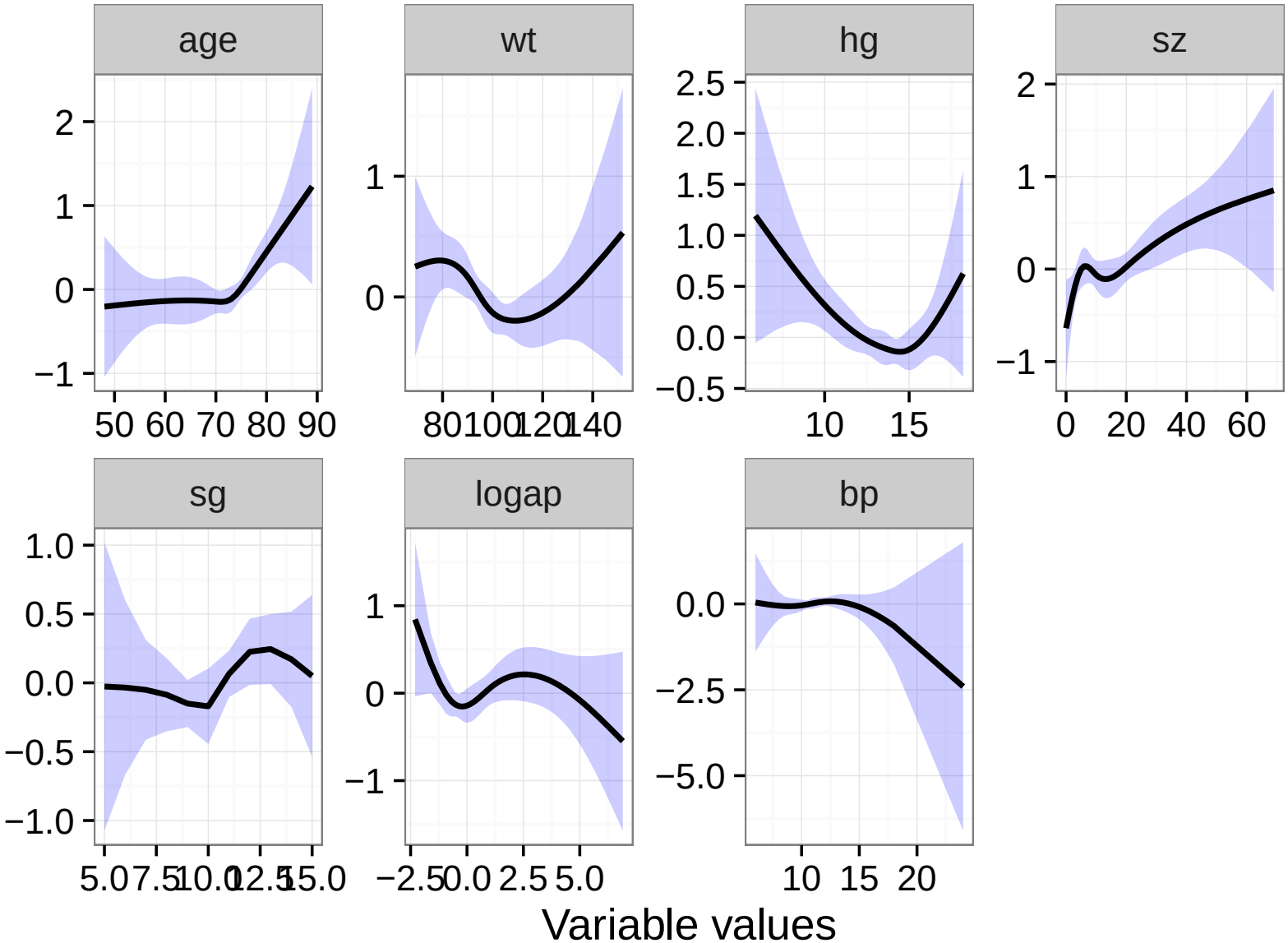


Nonlinear expansions

```
form <- Surv(dtime, status != "alive") ~ rx + ns(age, 4) +
  ns(wt, 4) + pf + hx + ekg +  ns(hg, 4) + ns(sz, 4) +
  ns(sg, 4) + ns(logap, 4) + ns(bp, 4) + bm
prostateCox2 <- coxph(form, data = subProstate)
anova(prostateCox2, prostateCox)

## Analysis of Deviance Table
##  Cox model: response is  Surv(dtime, status != "alive")
...
##   loglik Chisq Df P(>|Chi|)
## 1  -1827
## 2  -1845  36.1 21     0.021
```

Reporting the model



Reporting the model