Stat 486 / 886 Survival Analysis

Chapter 6. Semi-Parametric Regression Analysis

Professor: Wenyu Jiang

Department of Mathematics and Statistics
Queen's University

Acknowledgement: Profs. Jerry Lawless, David Matthews (Waterloo),
Profs. Bingshu Chen, Paul Peng, Dongsheng Tu (Queen's)

Sections of Chapter 6

6.1 Proportional Hazards Models

Proportional Hazards (PH) models

- Another important family of distributions for building regression models for survival data.
- If T is from a PH Model, the hazard function of T given z has the form

$$h(t|z) = h_0(t)\psi(z), \tag{1}$$

with $h_0(t) > 0$, $\psi(z) > 0$,
$h_0(t)$: the baseline hazard function,
$\psi(z)$: the relative risk function, with $\psi(0) = 1$.

Common choice: $\psi(z) = e^{\beta^T z}$. The typical PH model considered for survival data is

$$h(t|z) = h_0(t)\exp(\beta^T z). \tag{2}$$

In a PH model, the effects of the covariates stay the same across time!

Some special properties of PH models

1. Relationships of hazard functions between subjects with covariates z_1, z_2:

   Their hazards are proportional to each other, the ratio is fixed over time!

2. Impact of the covariate on the survival function of T:

Example: Show that the Weibull model is a PH model.

6.2 The Cox Regression Model

A PH model can be completely specified as a parametric model, and analyzed by the maximum-likelihood method.

Cox regression model: a semi-parametric method for analyzing PH models.
Model considered, same as (2) in Section 6.1:

$$h(t|z) = h_0(t)\exp(\beta^T z),$$

where z is a covariate vector (p-dimensional),
β is a vector of regression coefficients in $\mathbb{R}^p$, and
$h_0(t)$ is the baseline hazard function.

Data: (x_i, δ_i, z_i) for subjects $i = 1, \ldots, n$,
where $x_i = \min(t_i, c_i)$ is the observed failure time,
with potential failure time t_i and potential censoring time c_i,
$\delta_i = I(t_i \leq c_i)$ is the censoring indicator,
and z_i the covariate vector, of subject i.

Cox (1972, 1975) proposed the partial likelihood method for model (2).

Idea (Heuristic) of Partial likelihood

Suppose the failure times of the subjects are distinct.

Order the failure time and denote them by $t_{(1)} < t_{(2)} < \ldots < t_{(K)}$, with $K \leq n$.

At time $t_{(j)}$, describe the following probability through an approximation

$$
\begin{aligned}
& P\{\text{subject } (j) \text{ fails at } t_{(j)} | 1 \text{ subject fails at } t_{(j)}\} \\
= & \frac{h_{(j)}(t_{(j)})\Delta t}{\sum_{l=1}^{n} Y_l(t_{(j)}) h_l(t_{(j)})\Delta t} = \frac{h_0(t_{(j)}) e^{\beta^T z_{(j)}}}{\sum_{l=1}^{n} Y_l(t_{(j)}) h_0(t_{(j)}) e^{\beta^T z_l}} \\
= & \frac{e^{\beta^T z_{(j)}}}{\sum_{l=1}^{n} Y_l(t_{(j)}) e^{\beta^T z_l}},
\end{aligned}
$$

where $Y_l(t) = I(x_l \geq t)$ is an indicator that subject l is at risk at t.

Partial likelihood (PL, for inference about β)

Expression of PL, over time,

$$L_c(\beta) = \prod_{j=1}^{K} \frac{e^{\beta^T z_{(j)}}}{\sum_{l=1}^{n} Y_l(t_{(j)}) e^{\beta^T z_l}},$$

where $Y_l(t)$ is the at risk process of subject l at time t.

Or describe PL by risk set, over time,

$$L_c(\beta) = \prod_{j=1}^{K} \frac{e^{\beta^T z_{(j)}}}{\sum_{l \in R(t_{(j)})} e^{\beta^T z_l}},$$

where $R(t)$ is the risk set (collection of subjects at risk) at time t.

Or describe PL by index of subjects,

$$L_c(\beta) = \prod_{i=1}^{n} \left[\frac{e^{\beta^T z_i}}{\sum_{l=1}^{n} Y_l(x_i) e^{\beta^T z_l}} \right]^{\delta_i}. \tag{3}$$

Example of the PL:

Intuitions:
For subject i, the hazard of failure is proportional to $e^{\beta^T z_i}$.
PL is the product of the ratios of hazards across failure times;
the ratio at each failure time t_i (with $\delta_i = 1$) is the hazard that subject i actually failed relative to the total hazards of all subjects at risk at the time.

A Brief History of the Cox Model

- "Regression Models and Life-Tables" by D.R. Cox, published in 1972 (Journal of the Royal Statistical Society, Series B, 1972).
  It is one of the most frequently cited journal articles in statistics.
- Cox introduced the "Partial Likelihood" (Biometrika, 1975).
- Counting processes for statistical inference (Aalen O. Annals of Statistics, 1978).
- Rigorous proofs of asymptotic theories for the Cox model, based on counting processes (Andersen and Gill, Annals of Statistics, 1982).

Remarks on the Cox regression model

- The Cox regression model assumes PH model (2) for failure time, but uses PL for analysis.
- In PL analysis, the baseline hazard function is not specified. This makes it a semi-parametric method.
- Impact of covariates:

- PL method has the properties of the maximum likelihood method (Cox, 1972, 1975; Andersen and Gill, 1982).

- The PL is completely determined by the order of subjects' failure / censoring times. It remains the same for transformed time scale, as long as the transformation function is strictly increasing.

PL Analysis, for β

Estimate, $\hat{\beta}$: obtained by maximizing the PL with respect to β (p-dimensional).

Maximizing (3) is equivalent to maximizing the log PL,

$$l_c(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \beta^T z_i - \log \left[\sum_{l=1}^{n} Y_l(x_i) e^{\beta^T z_l} \right] \right\}.$$

For simplicity of notation, define

$$\begin{aligned}
S^{(0)}(\beta, t) &= \sum_{l=1}^{n} Y_l(t) e^{\beta^T z_l}, \\
S^{(1)}(\beta, t) &= \frac{\partial S^{(0)}}{\partial \beta} = \sum_{l=1}^{n} Y_l(t) z_l e^{\beta^T z_l}, \\
S^{(2)}(\beta, t) &= \frac{\partial^2 S^{(0)}}{\partial \beta \partial \beta^T} = \sum_{l=1}^{n} Y_l(t) z_l z_l^T e^{\beta^T z_l}.
\end{aligned}$$

From the log PL

Score function:
$$U_c(\beta) = \frac{\partial l_c}{\partial \beta} = \sum_{i=1}^{n} \delta_i \left\{ z_i - \frac{S^{(1)}(\beta, x_i)}{S^{(0)}(\beta, x_i)} \right\}, \tag{4}$$

Information matrix:
$$I_c(\beta) = -\frac{\partial^2 l_c}{\partial \beta \partial \beta^T}$$
$$= \sum_{i=1}^{n} \delta_i \left\{ \frac{S^{(2)}(\beta, x_i)}{S^{(0)}(\beta, x_i)} - \frac{S^{(1)}(\beta, x_i)[S^{(1)}(\beta, x_i)]^T}{[S^{(0)}(\beta, x_i)]^2} \right\}. \tag{5}$$

Solving $U_c(\beta) = 0$ gives the maximum partial likelihood estimate (MPLE) $\hat{\beta}$.

Needs to be solved numerically, often by the Newton-Raphson method.

Large sample approximations for the PL method,

$$
\begin{aligned}
&U_c(\beta_0) \approx \mathrm{N}\left(0, I_c(\hat{\beta})\right);\\
&\hat{\beta} \approx \mathrm{N}\left(\beta_0, I_c^{-1}(\hat{\beta})\right);\\
&2[l_c(\hat{\beta}) - l_c(\beta_0)] \approx \chi^2_p,
\end{aligned}
$$

where β_0 is the true value of β.

MPLE has the same properties of the MLE.

- Estimate the variance of MPLE by $\widehat{\mathrm{Var}}(\hat{\beta}) = I_c^{-1}(\hat{\beta})$.
- Perform tests and build confidence intervals for the regression coefficients, based on the above results.
- See Section 3.1 also for component-wise Wald statistic, LR test for a subset of the parameter vector...

Interpretation of Regression Coefficients, Hazard Ratio

PH Model (2) has the form

$$h(t|z) = h_0(t)\exp(\beta_1 z_1 + \ldots + \beta_p z_p).$$

- No intercept β_0 is needed in model (2). Why?

- Model (2) is equivalent to

$$\log \frac{h(t|z)}{h_0(t)} = \beta_1 z_1 + \ldots + \beta_p z_p.$$

- Example: Interpret a regression coefficient (effect of a covariate). Consider 3 covariates: $z_1 =$ age (years), $z_2 =$ smoking status (0 for non-smokers, 1 for smokers), $z_3 =$ body mass index (bmi) of the subject.

  β_1: the log hazard ratio for a 1 year increase in age, for subjects with the same smoking status and bmi.

  e^{β_1}: the hazard ratio for a 1 year increase in age, adjusted for smoking status and bmi.

  β_2: the log hazard ratio for smoking (vs. not smoking), adjusted for age and bmi.

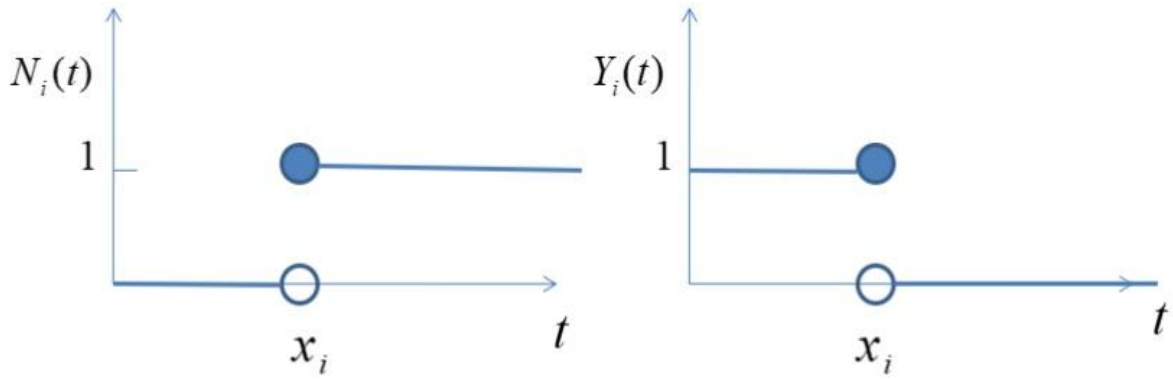
  e^{β_2}: the hazard ratio between smokers and non-smokers, for subjects of the same age and the same bmi.

Estimation of $H_0(t)$ and $S_0(t)$

Describe the life history of a subject by counting processes (Section 4.2 also).

Let $N_i(t) = I(X_i \leq t, \Delta_i = 1)$ for $i = 1, \ldots, n$,
counts the number of failures (0 or 1) for subject i up to time t.

Let $Y_i(t) = I(X_i \geq t)$ be the at risk process, for $i = 1, \ldots, n$,
indicates if subject i is still alive at time t.



$dN_i(t) = I\{X_i \in [t, t+\Delta t), \Delta_i = 1\}$, indicates if a failure occurs at t.

$\sum_{i=1}^{n} dN_i(t) =$ number of failures occur at time t.

Breslow's Estimator for $H_0(t)$

Heuristic explanation:

Breslow's estimator for $H_0(t)$, the baseline cumulative hazard function, has the form

$$\hat{H}_0(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(t)}{\sum_{i=1}^n Y_i(x_l) e^{\hat{\beta}^T z_i}} = \sum_{l:x_l<t} \left\{ \frac{\delta_l}{\sum_{i=1}^n Y_i(x_l) e^{\hat{\beta}^T z_i}} \right\}.$$

With $\hat{H}_0(t)$ above, we can estimate the baseline survival function $S_0(t)$ by

$$\hat{S}_0(t) = e^{-\hat{H}_0(t)}.$$

How to estimate the survival function for a subject with covariate z?
What about its cumulative distribution function?

Example 6.2.1 VA Lung Cancer Data

See Example 5.4.1 for description.
Analyzing the data by the Cox regression models. Which covariates are associated with the patients' survival time?

1. Fit a Cox regression model with all covariates.

Let z_{i1}= performance status (PS), z_{i2}= age,
z_{i3}= years from diagnosis to entry into the study (diag),
$z_{i4} = I(\text{type} = 2)$, $z_{i5} = I(\text{type} = 3)$, $z_{i6} = I(\text{type} = 4)$ for tumor cell categories,
$z_{i7} = I(\text{trt} = 2)$ for treatment categories, for patient i.

Consider T_i, the survival time (in days) of patient i.
Model 1: The hazard function of failure time T_i has the form
$h(t|z_i) = h_0(t)\exp(\beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} + \beta_4 z_{i4} + \beta_5 z_{i5} + \beta_6 z_{i6} + \beta_7 z_{i7})$,
where $h_0(t)$ is the baseline hazard function.

The covariates "age" (p-value=0.56), "diag" (p-value=0.93) and "trt" (p-value=0.29) do NOT significantly affect the survival time distribution..

2. Fit a Cox regression model with performance status and cell type.

Model 2: The hazard function of failure time T_i has the form
$h(t|z_i) = h_0(t)\exp(\beta_1 z_{i1} + \beta_4 z_{i4} + \beta_5 z_{i5} + \beta_6 z_{i6})$.

In the R output for the model, what are $\hat{\beta}_1$ and $e^{\hat{\beta}_1}$?
Also notice the inference results reported for β_1 and e^{β_1}.

```
                          coef exp(coef) se(coef)      z Pr(>|z|)
perfstat                -0.05971   0.94204  0.01343 -4.447 8.72e-06 ***
as.factor(celltype)2     0.36242   1.43681  0.49511  0.732   0.4642
as.factor(celltype)3     1.30454   3.68601  0.61417  2.124   0.0337 *
as.factor(celltype)4     0.40033   1.49232  0.46990  0.852   0.3942
---
                        exp(coef) exp(-coef) lower .95 upper .95
perfstat                    0.942     1.0615    0.9176    0.9672
as.factor(celltype)2        1.437     0.6960    0.5445    3.7917
as.factor(celltype)3        3.686     0.2713    1.1060   12.2840
as.factor(celltype)4        1.492     0.6701    0.5941    3.7484
```

By the Wald tests, both PS and cell type covariates significantly affect the survival time distributions. We can stop here and choose Model 2 to be the final model. Interpretations:
$e^{\hat{\beta}_1} = 0.942$: hazard ratio corresponding to a 1-unit increase in PS, for patients with the same cell type.
What about $\hat{\beta}_1$?

$\hat{\beta}_5 = 1.304$: log hazard ratio of patients with cell type 3, relative to those with cell type 1, given the same PS.

If the main interest is to find the simplest model for describing the survival time distribution, try the following model as well.

3. Fit a Cox regression model with performance status only.

Model 3: The hazard function of failure time T_i has the form
$h(t|z_i) = h_0(t)\exp(\beta_1 z_{i1})$.

The following table summarizes the LR tests for comparing the models.

Model	Log-likelihood	Compared to Model	Observed LR statistic	p-value
1	-87.46			
2	-88.10	1	1.29	0.73
3	-90.15	2		

Compare Models 2 and 3 by testing $H_0 : \beta_4 = \beta_5 = \beta_6 = 0$.

LR statistic: $\Lambda = 2[l(\hat{\beta}_1, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6) - l(\tilde{\beta}_1)] \sim \chi^2_3$.

$\lambda_{obs} =$

Choose Model 3 as the final model is fine too.

Estimate $S(t|z)$ based on the fitted Cox model

Consider Model 3.
Estimated the survival function for patients with median performance status.
(Also for those with PS at the 75th percentile, and mean value.)
The figure on the next page shows these estimated survival functions.
Useful for comparing the survival distributions.

How to do this in R?

- Obtain the "coxph" fit (as an object), i.e., the fit of the Cox model.
- Run "survfit" for the object, setting "newdata" at the specific covariate value (or values). This gives $\hat{S}(t|z)$.
- If "newdata" is not specified, by default, the survival function for those with mean covariate values are estimated.
- Plot $\hat{S}(t|z)$, if needed.
- See the R code at the end of this section.

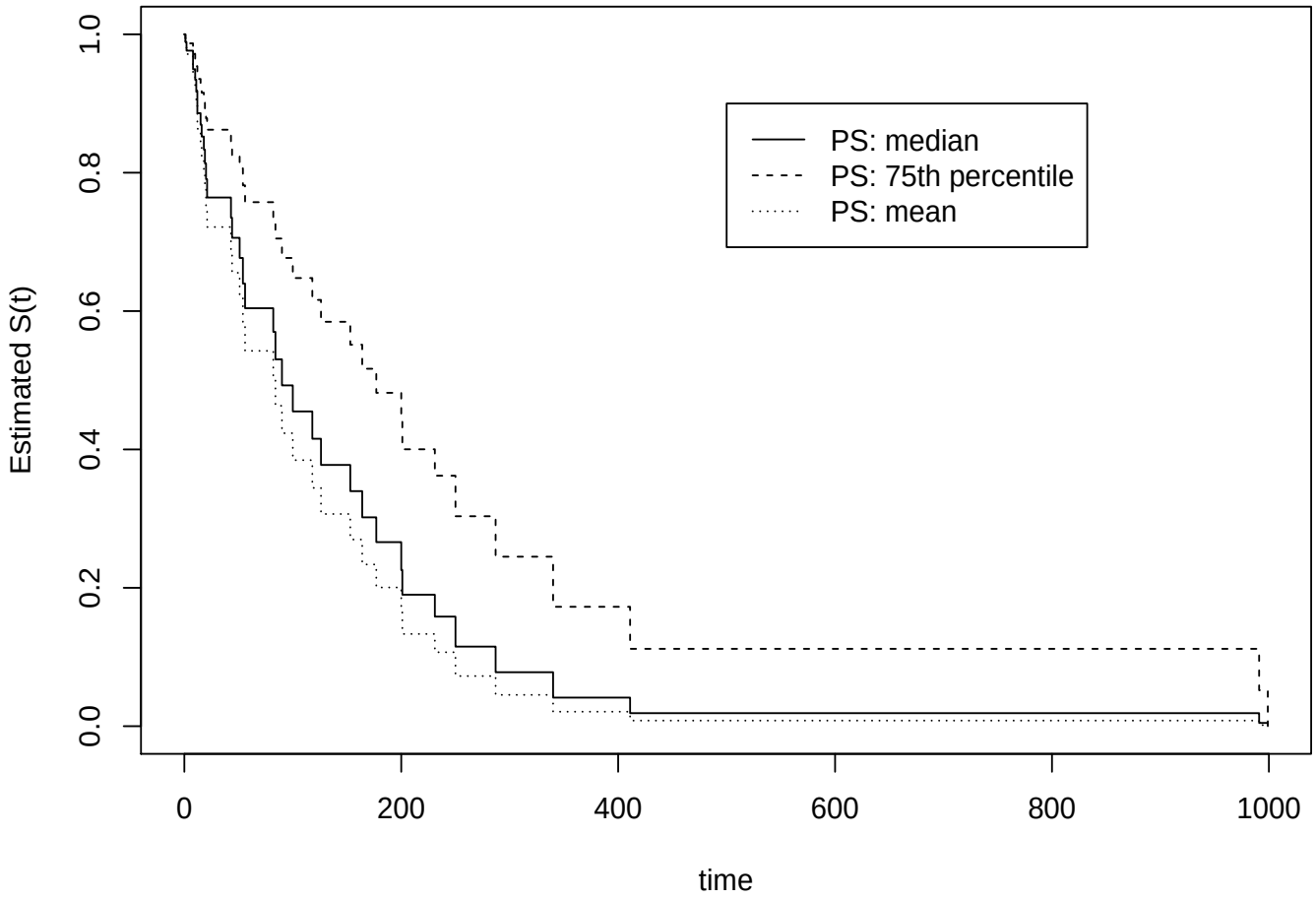
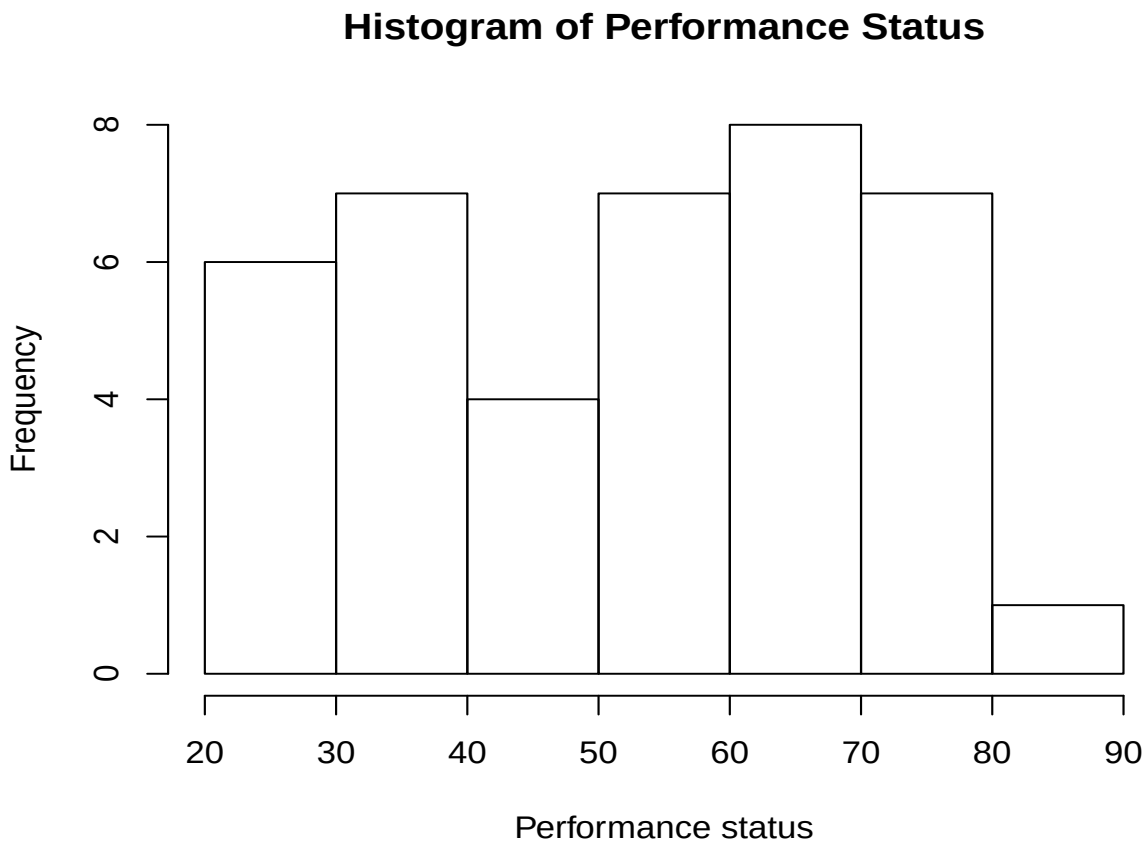


Figure: Estimated survival functions for lung cancer patients with different performance status values, based on Model 3.

How is the performance status variable distributed?



Some comparisons based on Model 2.

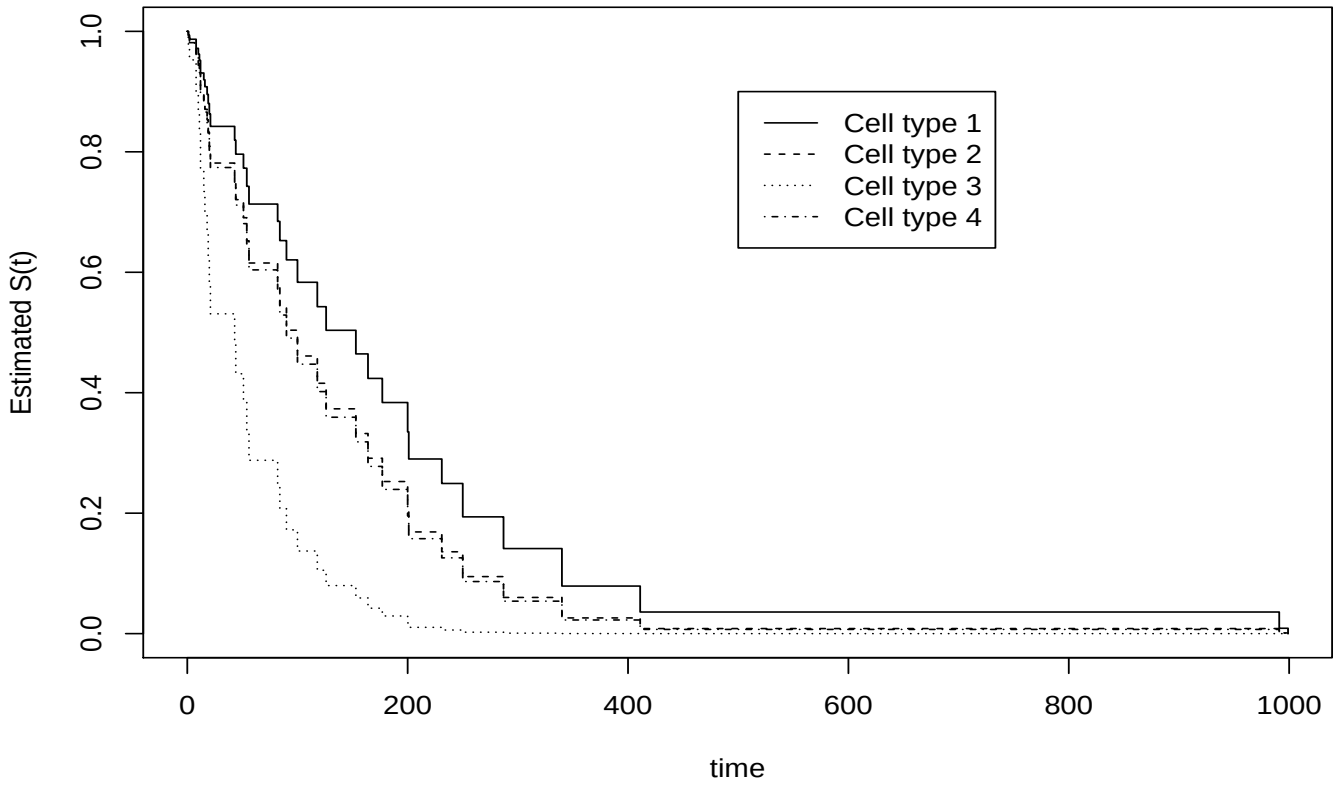


Figure: Estimated survival functions for patients with median performance status, compared across the 4 tumor cell type groups, based on Model 2.

```
>  library(survival)
>  lung<-read.table("eg541.txt", header=T)
>  # Model 1.
>  fit1<-coxph(Surv(time,status)~perfstat+age+msdiag+as.factor(celltype)
+as.factor(trt),data=lung)
>  print(summary(fit1))
  n= 40, number of events= 37
                          coef  exp(coef)   se(coef)      z Pr(>|z|)
perfstat             -0.0601486  0.9416246  0.0138324 -4.348 1.37e-05 ***
age                  -0.0121135  0.9879596  0.0207407 -0.584   0.5592
msdiag                0.0009725  1.0009730  0.0117874  0.083   0.9342
as.factor(celltype)2  0.2768925  1.3190245  0.5457795  0.507   0.6119
as.factor(celltype)3  1.4067875  4.0828183  0.6263071  2.246   0.0247 *
as.factor(celltype)4  0.3183133  1.3748070  0.4864511  0.654   0.5129
as.factor(trt)2      -0.4310210  0.6498453  0.4092120 -1.053   0.2922
---
```

```
                      exp(coef) exp(-coef) lower .95 upper .95
perfstat                 0.9416     1.0620    0.9164    0.9675
age                      0.9880     1.0122    0.9486    1.0289
msdiag                   1.0010     0.9990    0.9781    1.0244
as.factor(celltype)2     1.3190     0.7581    0.4526    3.8443
as.factor(celltype)3     4.0828     0.2449    1.1963   13.9339
as.factor(celltype)4     1.3748     0.7274    0.5299    3.5671
as.factor(trt)2          0.6498     1.5388    0.2914    1.4492
Concordance= 0.764  (se = 0.034 )
Likelihood ratio test= 29.78  on 7 df,   p=1e-04
Wald test            = 26.08  on 7 df,   p=5e-04
Score (logrank) test = 30.59  on 7 df,   p=7e-05
>
> # Log-likelihood: [1] no covariates and [2] with covariates in model.
>  fit1$loglik
[1] -102.34404  -87.45621
> # LR statistics for H0: no covariate effects.
>  2*(fit1$loglik[2]-fit1$loglik[1])
[1] 29.77567
```

```
>  # Model 2.
>  fit2<-coxph(Surv(time,status)~perfstat+as.factor(celltype),
data=lung)
>  print(summary(fit2))
                          coef exp(coef) se(coef)      z Pr(>|z|)
perfstat                -0.05971   0.94204  0.01343 -4.447 8.72e-06 ***
as.factor(celltype)2     0.36242   1.43681  0.49511  0.732   0.4642
as.factor(celltype)3     1.30454   3.68601  0.61417  2.124   0.0337 *
as.factor(celltype)4     0.40033   1.49232  0.46990  0.852   0.3942
---
                        exp(coef) exp(-coef) lower .95 upper .95
perfstat                    0.942     1.0615    0.9176    0.9672
as.factor(celltype)2        1.437     0.6960    0.5445    3.7917
as.factor(celltype)3        3.686     0.2713    1.1060   12.2840
as.factor(celltype)4        1.492     0.6701    0.5941    3.7484
>
>  # LR test comparing Model 1 and Model 2.
>  fit2$loglik
[1] -102.34404  -88.10023
>  2*(fit1$loglik[2]-fit2$loglik[2])  # Observed LR statistic.
[1] 1.28805
>  1-pchisq(2*(fit1$loglik[2]-fit2$loglik[2]),3)   # p-value.
[1] 0.7319727
```

```
>  # Model 3.
>  fit3<-coxph(Surv(time,status)~perfstat,data=lung)
>  print(summary(fit3))
             coef exp(coef) se(coef)      z Pr(>|z|)
perfstat -0.05953   0.94221  0.01269 -4.691 2.72e-06 ***
---
         exp(coef) exp(-coef) lower .95 upper .95
perfstat    0.9422      1.061    0.9191    0.9659
>
> # LR test comparing Model 2 and Model 3.
>  fit3$loglik
[1] -102.3440  -90.1465
>  2*(fit2$loglik[2]-fit3$loglik[2])  # Observed LR statistic.
[1] 4.092541
>  1-pchisq(2*(fit2$loglik[2]-fit3$loglik[2]),3)  # p-value.
[1] 0.2516432
>
> # Distribution of performance status (PS).
> hist(lung$perfstat,main="Histogram of Performance Status",
xlab="Performance status")
>  quantile(lung$perfstat)  # Quantiles of PS.
  0%  25%  50%  75% 100%
  20   40   60   70   90
>  mean(lung$perfstat)      # Mean of PS.
[1] 56.75
```

```
> # Estimate survival functions based on Model 3.
> # For those with median PS.
> sf.cox.50<-survfit(fit3,newdata=data.frame(perfstat=60))
> plot(sf.cox.50,conf.int=F,xlab="time",
ylab="Estimated S(t)")
> # For those with PS at the 75 percentile.
> sf.cox.75<-survfit(fit3,newdata=data.frame(perfstat=70))
> lines(sf.cox.75,conf.int=F,lty=2)
> # Without specified "newdata", by default, S(t) is
> # estimated for those with the mean covariate value.
> sf.cox.mean<-survfit(fit3)
> lines(sf.cox.mean,conf.int=F,lty=3)
> legend(500,0.9,c("PS: median ","PS: 75th percentile",
"PS: mean"),lty=1:3)
```

```
> # Estimate survival functions based on Model 2
> # For those with median PS, and different cell types.
> sf.M2.type1<-survfit(fit2,newdata=
data.frame(perfstat=60,celltype=1))
> plot(sf.M2.type1,conf.int=F,xlab="time",ylab=
"Estimated S(t)")
> sf.M2.type2<-survfit(fit2,newdata=
data.frame(perfstat=60,celltype=2))
> lines(sf.M2.type2,conf.int=F,lty=2)
> sf.M2.type3<-survfit(fit2,newdata=
data.frame(perfstat=60,celltype=3))
> lines(sf.M2.type3,conf.int=F,lty=3)
> sf.M2.type4<-survfit(fit2,newdata=
data.frame(perfstat=60,celltype=4))
> lines(sf.M2.type4,conf.int=F,lty=4)
> legend(500,0.9,c("Cell type 1","Cell type 2",
"Cell type 3","Cell type 4"),lty=1:4)
```

6.3 Residual Analysis

Residuals are commonly used to assess the fit of a regression model to individual subjects.

To assess the fit of a Cox regression model, check

- PH assumption: is the PH model form $h(t) = h_0(t)\exp(\beta^T z)$ appropriate?
- model adequacy: is the linear predictor $\beta^T z$ adequate and appropriate?

There are various types of residuals for the Cox regression models.

1. Martingale residuals (Therneau, Grambsch, Fleming, 1990, Biometrika)

$$\hat{M}_i = \delta_i - \hat{H}_0(x_i)e^{\hat{\beta}^T z_i} = \delta_i - \hat{H}(x_i|z_i)$$

where $x_i = \min(t_i, c_i)$ is the observed failure time of subject i, $\hat{H}_0(t)$ is the Breslow's estimator of $H_0(t)$.

Notes:

- Martingale is a type of stochastic processes with nice known properties.
- $\hat{M}_i \in (-\infty, 1]$. Martingale residuals are skewed.
- It can be shown that $\sum_{i=1}^{n} \hat{M}_i = 0$.

2. Cox-Snell (C-S) residuals

$$\hat{r}_i = \hat{H}(x_i|z_i) + 1 - \delta_i,$$

for $x_i = \min(t_i, c_i)$ and $\delta_i = I(t_i \leq c_i)$.

What should the residuals look like, if the model fits the data well?

What is the relationship between CS residuals and the Martingale residuals?

3. Deviance residuals – transformed martingale residuals

$$d_i = \text{sign}(\hat{M}_i)\left\{-2[\hat{M}_i + \delta_i \log(\delta_i - \hat{M}_i)]\right\}^{\frac{1}{2}}$$

Notes:

- When the Cox model fits the data well, the deviance residuals should roughly have a N $(0,1)$ shaped distribution.
- When the censoring rate is below 25%, the distribution of the d_i's should be close to N $(0,1)$.
- When the censoring rate is above 40%, a large number of residuals are close to 0, which distorts the N $(0,1)$ distribution shape. But residuals should still mostly fall between $[-2,2]$, and be roughly symmetric in range.
- Deviance residuals are useful for assessing both PH assumption and model adequacy.

In R, obtain the residuals by “residuals(fit, type = “deviance”)”.

Diagnostic plots based on deviance residuals

- Plot deviance residuals against each covariate.
  Check for uncaptured trends in the covariate...
- Plot deviance residuals against the risk score.
  Check for overall model fit. Look like observations from N $(0, 1)$?

  What is a risk score?

- A normal Q-Q plot of the deviance redisuals
  (ordered d_i's versus N $(0, 1)$ quantiles).
  Check if d_i's are approximately normally distributed.

Example 6.3.1 VA Lung Cancer Data

See Example 5.4.1 for description, and Example 6.2.1 for statistical analysis by Cox regression models.
Study the deviance residuals for comparing and checking the fit of Models 2 and 3.

Model 3 is the Cox regression model with the performance status (PS) covariate only.
Model 2 is the Cox regression model with PS and cell type.

For each model, the deviance residuals are plotted against the each of the covariates, PS and cell type;
then plotted against the risk score;
the normal Q-Q plot is also presented, along with the line of $y = x$.

See the figures below.

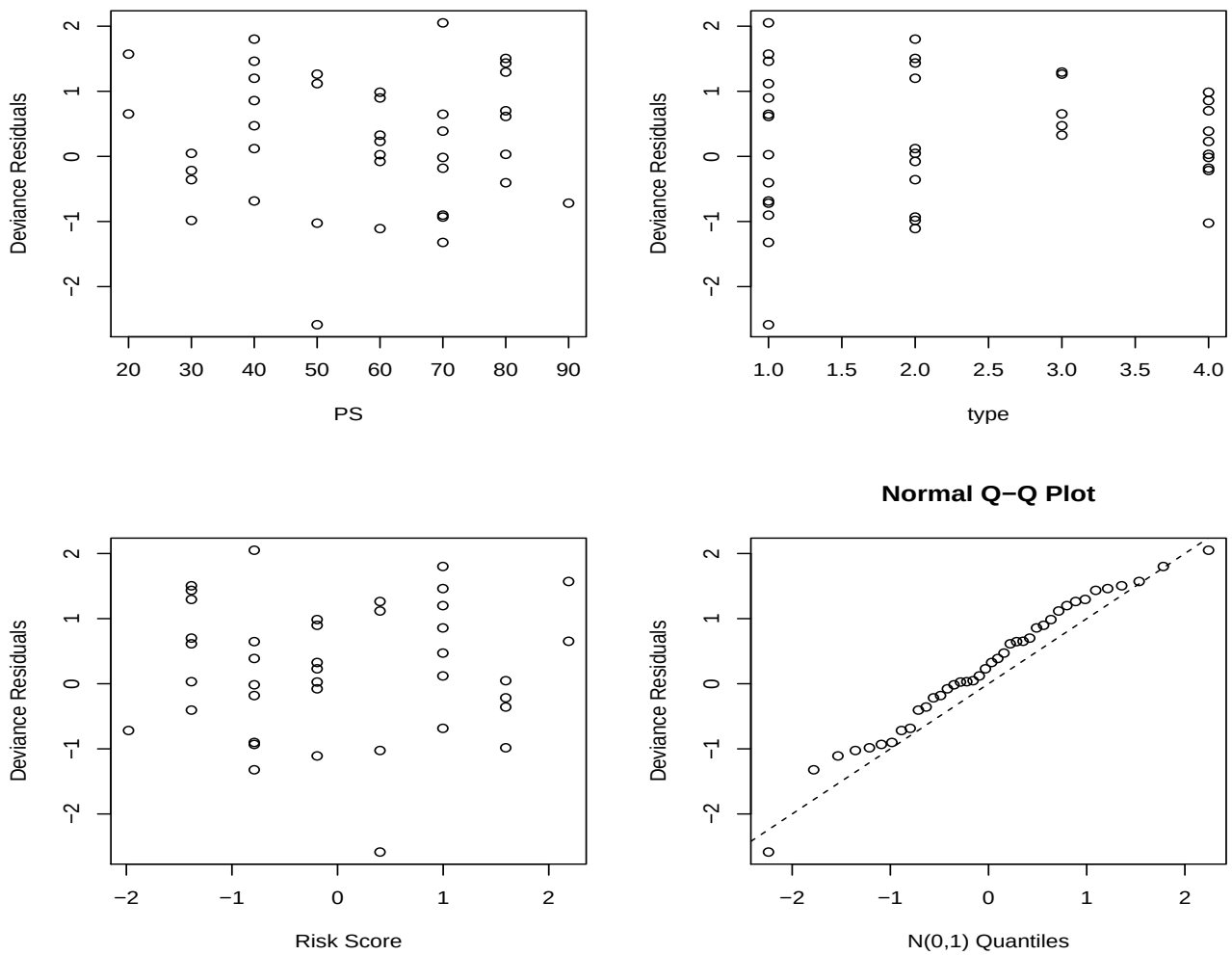


Figure: Residual plots for Model 3. Clock-wise: deviance residuals vs. performance status, cell types, and risk scores; normal Q-Q plot for the residuals.

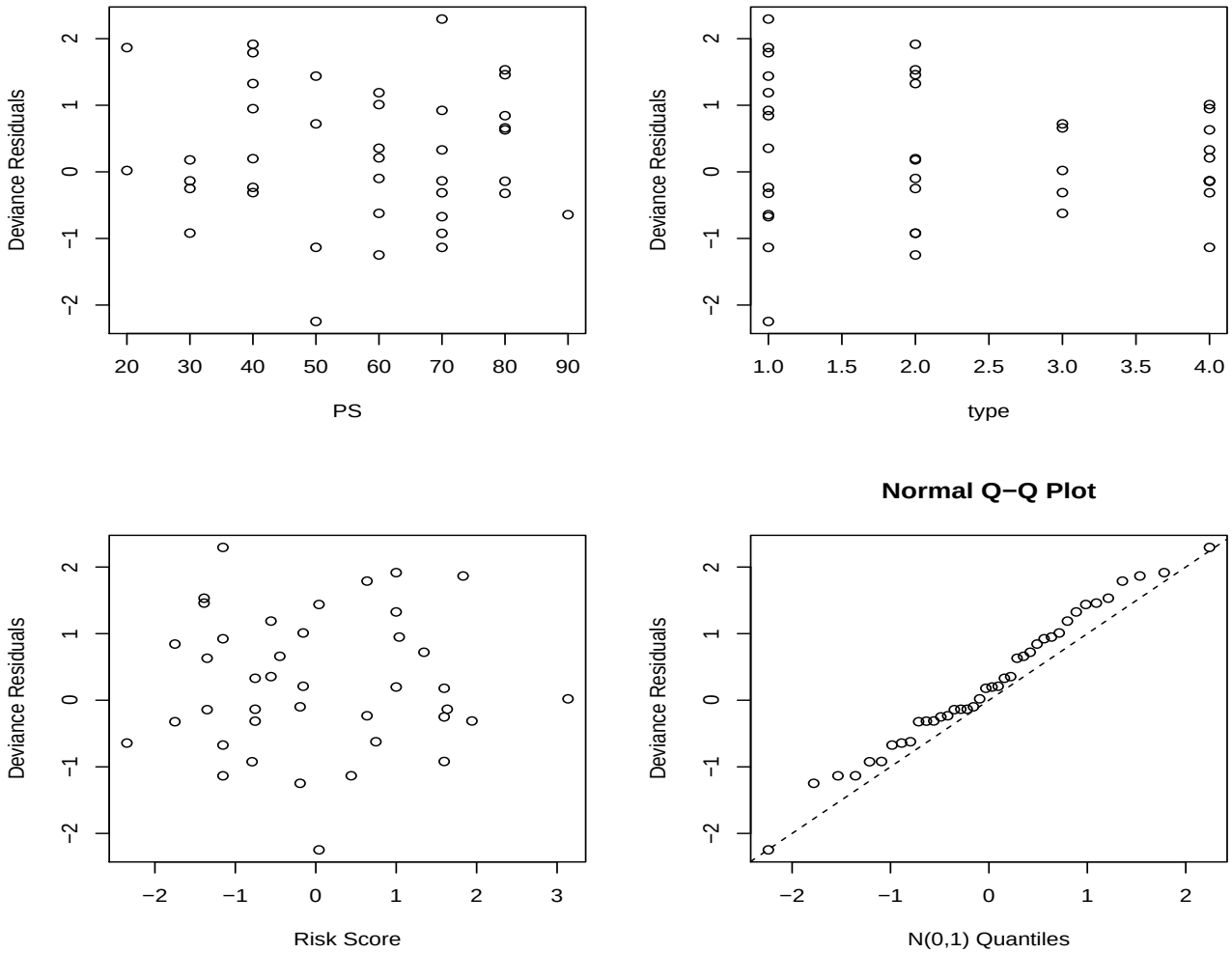


Figure: Residual plots for Model 2. Clock-wise: deviance residuals vs. performance status, cell types, and risk scores; normal Q-Q plot for the residuals.

Discussion

- On the deviance residual vs. cell type plot for Model 3, residuals for subjects with cell type 3 are all positive.
- For Model 2, the residuals for subjects with cell type 3 become more symmetric about zero.
- Deviance residuals in the plots versus performance status and versus risk scores behave roughly like random observations from N $(0, 1)$.
- For Model 3, the Q-Q plot agrees fairly well with the straight line $y = x$; for Model 2, there is some slight improvement in this.
- Model 3 is simpler. It is no worse than Model 2 according to the LR test (shown in Example 6.2.1).
- Model 2 includes all covariates significantly associated with the survival time. Its fit is slightly better according to the residual analysis.
- There are only 5 patients in cell type 3. If the effects of cell types are of interest, a larger study may be needed (more patients in each cell type).

```
>  library(survival)
>  lung<-read.table("eg541.txt",header=T)
>  # Model 3 for Example 6.2.1.
>  fit3<-coxph(Surv(time,status)~perfstat, data=lung)
>  # Calculate deviance residuals.
>  fit3.resids<-resid(fit3, type="deviance")
>  PS<-lung$perfstat
>  type<-lung$celltype
>  par(mfrow=c(2,2))
>  # Residuals vs. covariate.
>  plot(PS,fit3.resids,ylab="Deviance Residuals")
>  plot(type,fit3.resids,ylab="Deviance Residuals")
>  # Residuals vs. risk score (predicted values).
>  plot(predict(fit3),fit3.resids,ylab="Deviance
Residuals",xlab="Risk Score")
>  # Normal Q-Q Plot.
>  qqnorm(fit3.resids,ylab="Deviance Residuals",
xlab="N(0,1) Quantiles")
>  abline(0,1,lty=2)
```

```
>  # Model 2 for Example 6.2.1.
>  fit2<-coxph(Surv(time,status)~perfstat+
as.factor(celltype),data=lung)
>  fit2.resids<-resid(fit2,type="deviance")
>  par(mfrow=c(2,2))
>  plot(PS,fit2.resids,ylab="Deviance Residuals")
>  plot(type,fit2.resids,ylab="Deviance Residuals")
>  plot(predict(fit2),fit2.resids,ylab="Deviance
Residuals",xlab="Risk Score")
>  qqnorm(fit2.resids,ylab="Deviance Residuals",
xlab="N(0,1) Quantiles")
>  abline(0,1,lty=2)
```

4. Schoenfeld residuals

Originally proposed by Schoenfeld (1982, Biometrika).

Recall the score function for partial likelihood (PL) analysis

$$\begin{aligned} U_c(\beta) &= \sum_{i=1}^{n} \delta_i \left[z_i - \frac{S^{(1)}(\beta, x_i)}{S^{(0)}(\beta, x_i)} \right] \\ &= \sum_{i=1}^{n} \delta_i \left[z_i - \frac{\sum_{l=1}^{n} z_l Y_l(x_i) e^{\beta^T z_l}}{\sum_{l=1}^{n} Y_l(x_i) e^{\beta^T z_l}} \right] \\ &= \sum_{i=1}^{n} \delta_i \left(z_i - \sum_{l=1}^{n} z_l w_{il} \right). \end{aligned}$$

Define Schoenfeld residual for the kth covariate by

$$\hat{s}_{ik} = \delta_i \left(z_{ik} - \sum_{l=1}^{n} z_{lk} \hat{w}_{il} \right)$$

for subjects $i = 1, \ldots, n$ with $\delta_i = 1$, for covariates $k = 1, \ldots, p$.

About the Schoenfeld residuals

- $\hat{s}_{ik}$ is the difference between the covariate value z_{ik} of the subject i that failed at time x_i, and its "expected value" across the risk set at x_i.

- $\hat{s}_{ik}$ is defined by subject i's contribution to the score function of the PL.
- Schoefeld residuals are NOT defined for censored subjects.
- Instead of one residual for each subject, there is a separate residual for each failed subject for each covariate.
- The mean of the Schoenfeld residuals is 0. Why?
- Schoenfeld residuals are independent of time if the PH assumption is valid.
- Schoenfeld residuals are used for checking the PH assumption. Time trends in the residual vs. time plot suggest violations of the PH assumption.

Scaled Schoenfeld Residuals and Tests for PH Assumptions

To detect time trends more easily from the Schoenfeld residuals, and test for the validity of PH assumptions, Grambsch and Therneau (Biometrika, 1994) defined the scaled Schoenfeld residuals,

$$s_{ik}^{*} = \hat{\beta}_k + \left\{ I_i^{-1}(\hat{\beta}, x_i) \left[z_i - \frac{S^{(1)}(\hat{\beta}, x_i)}{S^{(0)}(\hat{\beta}, x_i)} \right] \right\}_k,$$

for subjects $i = 1, \ldots, n$ with $\delta_i = 1$, for covariates $k = 1, \ldots, p$, where $I_i(\hat{\beta}, x_i)$ is the piece of $I(\hat{\beta})$ contributed by the failed subject i at time x_i.

The mean of scaled Schoenfeld residuals is at

Idea of tests for PH assumption:
Check PH model $h(t|z) = h_0(t)\exp(\beta^T z)$ by embedding it to a more general model

$$h(t|z) = h_0(t)\exp[\beta(t)^T z], \text{ with } \beta(t) = \beta + \theta g(t)$$

for some function $g(t)$ of time.

A test of PH assumption is equivalent to

a test of $H_0 : \theta = 0$ in the model above.

Details of the tests for PH assumption, based on scaled Schoenfeld residuals

- Grambsch and Therneau developed tests based on the idea above, the test statistics were shown to have chi-square distributions.

- In R, “cox.zph” conducts these chi-square tests.
- Choices of $g(t)$.

- Test for an individual covariate, and the global test (all covariates).

Example 6.3.2 VA Lung Cancer Data

See Example 5.4.1 for description, and Example 6.2.1 for statistical analysis by Cox regression models.
Carry out residual analysis for Model 2, check if the PH assumption is valid.
Make use of Schoenfeld residuals and scaled Schoenfeld residuals.

Model 2 is the Cox model with performance status (PS) and cell type.

Schoenfeld residuals for PS, and each of the cell types 2, 3, 4 are plotted against time. See the figure on the next page.

If Model 2 fits the data well, the means of the residuals should be close to

Are there any time trends?

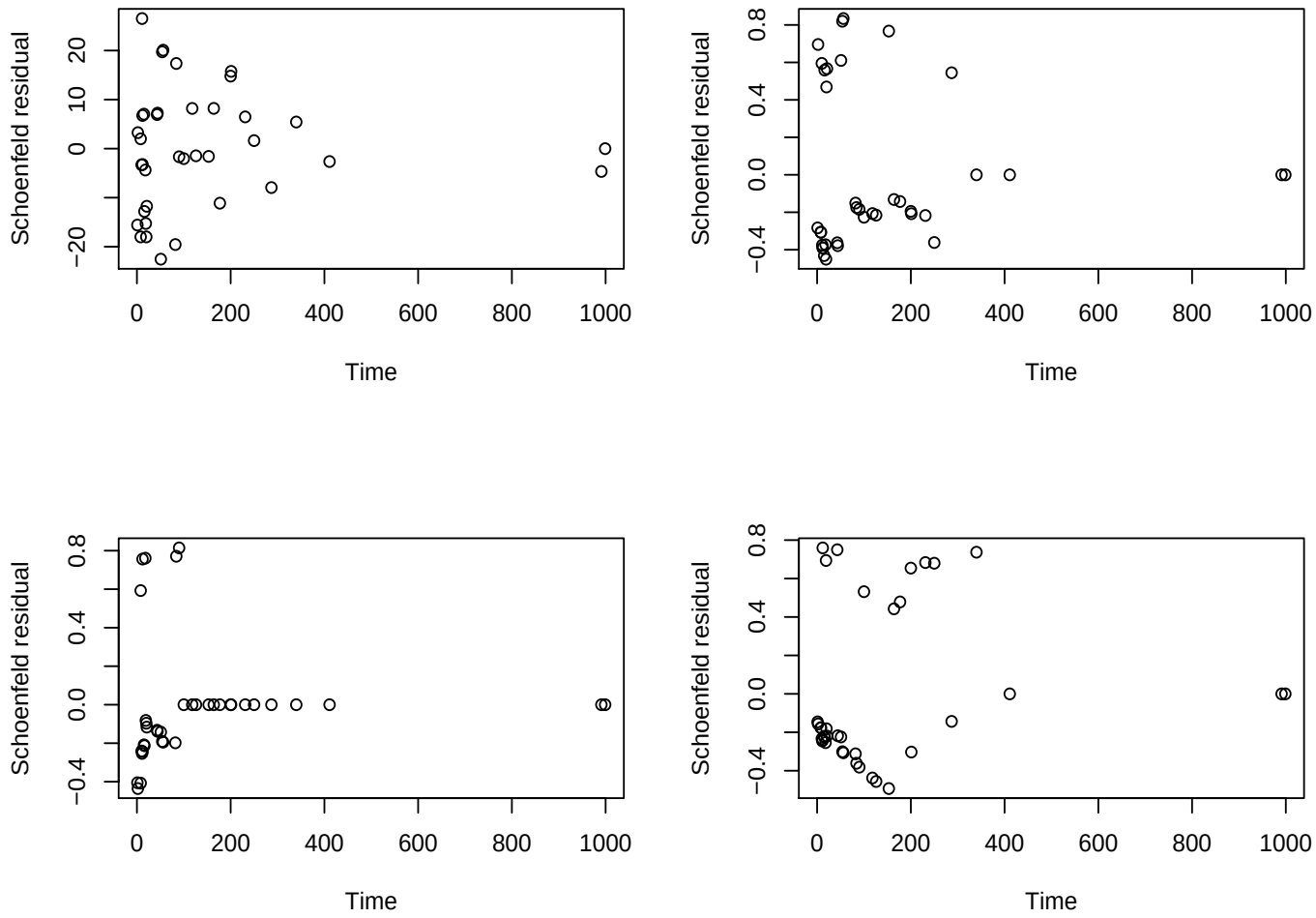


Figure: Schoenfeld residuals (original) versus time, for performance status, and cell types 2, 3, 4 (clock-wise).

Next consider the scaled Schoenfeld residuals.
The following are the tests for time trends in the residuals.
Obtained from "cox.zph" in R. See attached R code and output.

	chisq	df	p
perfstat	1.0144	1	0.314
factor(celltype)2	0.0267	1	0.870
factor(celltype)3	1.8197	1	0.177
factor(celltype)4	3.3349	1	0.068
GLOBAL	11.3961	4	0.022

No time trends are confirmed for the individual covariates (although for cell type 4, the test gives a p-value 0.068). See the residual plots on the next page.

The global test indicates significant time trends (p-value=0.022).

Is the PH assumption valid for Model 2?

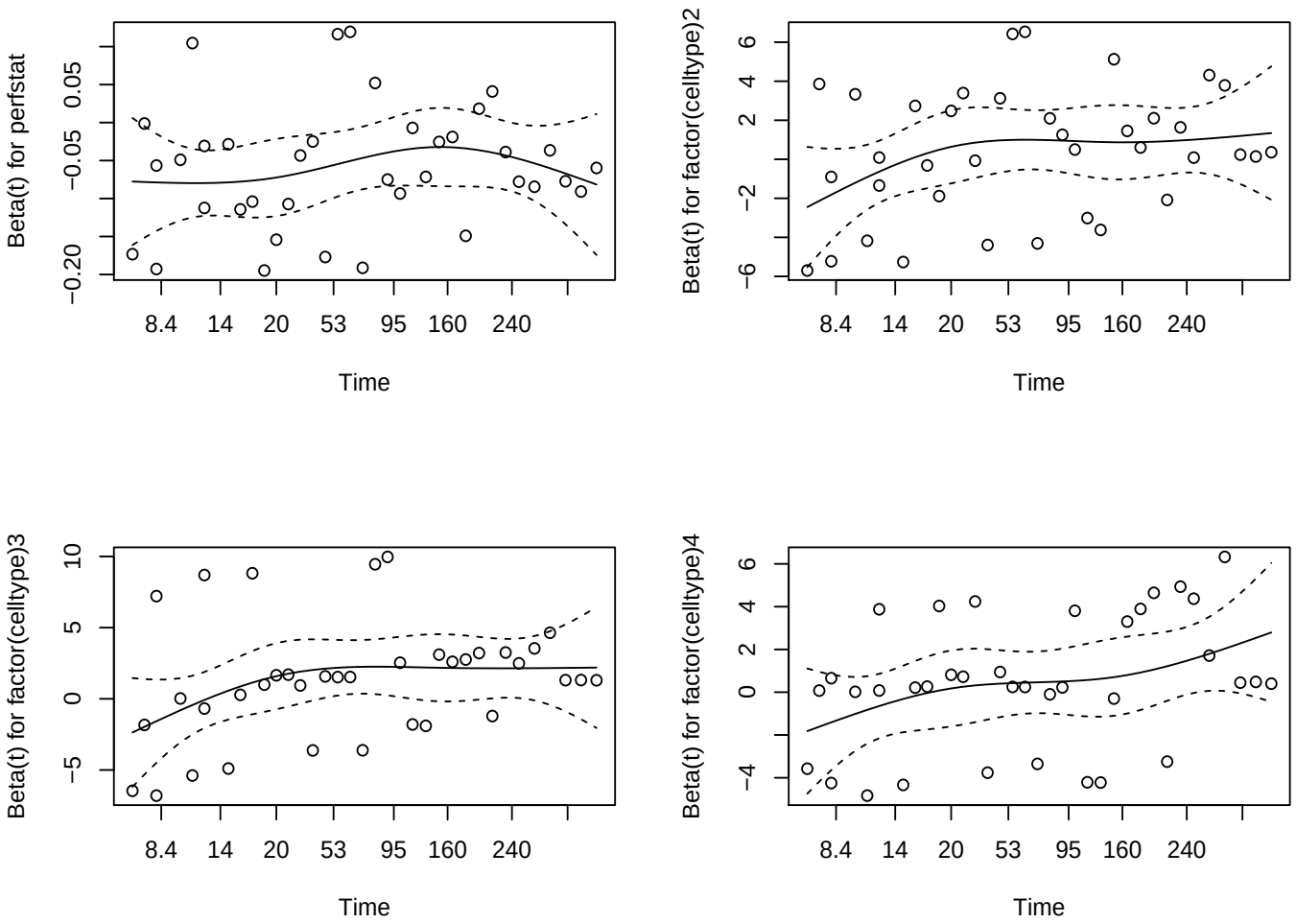


Figure: Scaled Schoenfeld residuals versus time, for performance status, and cell types 2, 3, 4 (clock-wise).

```
>  library(survival)
> lung=read.table("eg541.txt", header=T)
>  # Model 2. See Example 6.2.1.
>  fit2=coxph(Surv(time,status)~perfstat+factor(celltype),data=lung)
>  # Schoenfeld residuals (original).
>  resid.sch=residuals(fit2,type="schoenfeld")
>  print(resid.sch[1:5,])
     perfstat factor(celltype)2 factor(celltype)3 factor(celltype)4
1  -15.580643        -0.2832816        -0.4052975        -0.1456593
2    3.261249         0.6956621        -0.4354232        -0.1564861
8    2.003330        -0.3065055        -0.4070958        -0.1763476
8  -17.996670        -0.3065055         0.5929042        -0.1763476
10  -3.305326         0.5947474        -0.2405465        -0.2331616
> dimnames(resid.sch)
[[1]]
 [1] "1"   "2"   "8"   "8"   "10"  "11"  "12"  "12"  "15" ...
......
[25] "126" "153" "164" "177" "200" "201" "231" "250" "287" ...
[37] "999"
[[2]]
[1] "perfstat"          "factor(celltype)2" "factor(celltype)3"
[4] "factor(celltype)4"
```

```
>  # Failure times in increasing order.
>  t=sort(lung$time[lung$status==1])
>  t
 [1]   1   2   8   8  10  11  12  12  15  16 ... ...
[20]  82  84  90 100 118 126 153 164 177 200 ... 999
>  # Plot Schoenfeld residuals vs. time.
>  par(mfrow=c(2,2))
>  plot(t,resid.sch[,1],ylab="Schoenfeld residual",
xlab="Time")
>  plot(t,resid.sch[,2],ylab="Schoenfeld residual",
xlab="Time")
>  plot(t,resid.sch[,3],ylab="Schoenfeld residual",
xlab="Time")
>  plot(t,resid.sch[,4],ylab="Schoenfeld residual",
xlab="Time")
```

```
>   # Scaled Schoenfeld residuals.
>   resid.scaledsch=residuals(fit2, type="scaledsch")
>
>   # Run cox.zph for scaled Schoenfeld residuals.
>   zph=cox.zph(fit2,terms=F)
>
>   # Plot scaled Schoenfeld residuals versus time.
>   par(mfrow=c(2,2))
>   plot(zph)
>
>   print(zph) # Test if residuals depend on time.
                     chisq df     p
perfstat            1.0144  1 0.314
factor(celltype)2   0.0267  1 0.870
factor(celltype)3   1.8197  1 0.177
factor(celltype)4   3.3349  1 0.068
GLOBAL             11.3961  4 0.022
```

6.4 Further Topics

So far we have introduced the Cox model, and the analysis based on the PL.
The Cox model and method are developed for

PH model,

distinct failure times,

time fixed covariates.

It turns out the Cox model can be

modified to handle ties in failure times,

extended to analyze non-proportional hazards data by

- stratification,
- including time-dependent covariates.

Handling Ties in Cox Model

Common methods for handling tied failure times in the Cox model:

- Exact method (very time consuming if there are a lot of ties).
- Breslow approximation (default method in SAS).
- Efron approximation (default method in R).

The exact method

- assumes ties result from imprecise measurement of time, and there is a true unknown order of the time of the events;
- calculates the exact probability of all possible orders of events;
- is very time consuming!

Example of the exact method:
At time $t = 1$ month, 3 failure events are observed, for subjects with ID's (3, 8, 16) respectively. The exact method assumes these subjects have different failure times, but the measurements are not precise enough to reveal that.

There are $3!$ ways (orders) that the 3 failures can happen:
$O_1 = (3, 8, 16), O_2 = (3, 16, 8), O_3 = (8, 3, 16), O_4 = (8, 16, 3), O_5 = (16, 3, 8), O_6 = (16, 8, 3)$.

At time $t = 1$, the exact method multiplies the term $\sum_{i=1}^{3!} P(O_i)$ to the PL, as the contribution to PL by the 3 tied failure events.
For the order of $O_6 = (16, 8, 3)$, for example,

$$P(O_6) = \left(\frac{e^{\beta^T z_{16}}}{e^{\beta^T z_3} + e^{\beta^T z_8} + e^{\beta^T z_{16}} + \ldots} \right) \left(\frac{e^{\beta^T z_8}}{e^{\beta^T z_3} + e^{\beta^T z_8} + \ldots} \right) \left(\frac{e^{\beta^T z_3}}{e^{\beta^T z_3} + \ldots} \right).$$

What if there are 15 ties at a time?

The Breslow method and Efron method.

- Breslow (1974) and Efron (1977) proposed approximations to the exact method.
- Both are faster in calculation time.
- Breslow method is more popular. It does not do well when the number of ties at a particular time point is large in proportion relative to the number of subjects at risk.
- Efron method is generally the more accurate of the two.
- In R, check descriptions for “coxph()”. We can specify the method by “coxph(..., ties = ...)”.

Stratified Cox Model

Sometimes the PH assumption is violated for the current Cox model fitted to the entire data. This can often be resolved by

- finding the suitable covariates for the model:
  including other important explanatory variables,
  removing unimportant variables
  adding other covariate forms such as higher order terms, and interaction terms between covariates;
  (That is, the form of the predictor $\beta^T z$ in the Cox model matters. For example, PH assumption holds for Model 3 of Example 6.2.1.)
- stratification;
- including time-dependent covariates.

The stratified Cox model:

Split subjects into J groups or strata, $j = 1, \ldots, J$, by variable s. Assume the following hazard function for subject i in strata j

$$h(t|z_i; s_i = j) = h_{0j}(t) \exp(\beta^T z_i), \text{ for } j = 1, \ldots, J,$$

where z_i is the covariate vector of subject i, other than the variable s.

The stratified Cox model

- allows a different baseline hazard function (and baseline survival function) for each strata;
- assumes PH model within each strata;
- relaxes the overall PH assumption;
- assumes a common β across all strata (how to test?)

The analysis of β for the stratified Cox model is based on stratified PL

$$L_c(\beta) = \prod_{j=1}^{J} L_{c_j}(\beta), \text{ where } L_{c_j}(\beta) \text{ is the PL of strata } j.$$

One baseline survival function $S_{0j}(t) = e^{-H_{0j}(t)}$ is estimated for each strata j.
Breslow's estimators by strata.

The following model is often compared to the stratified Cox model. It assumes the hazard function for subject i in strata j has the form

$$h(t|z_i; s_i = j) = h_{0j}(t) \exp(\beta_j^T z_i), \text{ for } j = 1, \ldots, J.$$

It assumes a separate Cox model for each strata.
It allows different covariate effects β_j across strata.

What is the likelihood function for all the data, based on this model?

Example 6.4.2 VA Lung Cancer Data

See Example 5.4.1 for description, and Example 6.2.1 for statistical analysis by Cox regression models and Example 6.3.2 for assessment of PH assumption for Model 2. Try the stratified Cox model, as a way to deal with the possible violation of PH assumption of Model 2.

Stratify by the cell type variable.

1. Fit a stratified Cox regression model with the covariate PS only.

Let z_i= performance status (PS) of patient i,
s_i = be the cell type of patient i, taking values 1, 2, 3, 4.

Model S: The hazard function of patient i of cell type j has the form
$h(t|z_i; s_i = j) = h_{0j}(t)\exp(\beta z_i)$ for $j = 1, 2, 3, 4$,
where $h_{0j}(t)$ is the baseline hazard function of cell type j.
The effect of PS is assumed to be the same across all strata.

For comparison:

2. Fit a Cox regression model for each cell type, with the covariate PS only.

Model S_j: The hazard function of patient i of cell type j has the form $h(t|z_i; s_i = j) = h_{0j}(t)\exp(\beta_j z_i)$ for $j = 1,2,3,4$,
where $h_{0j}(t)$ is the baseline hazard function of cell type j.
The effect of PS is allowed to be different among strata.

Is fitting separate Cox models for the 4 cell types better than the stratified Cox model? Or, is the effect of PS the same across all cell types?

Test $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta$, based on the LR statistic

$$\Lambda = 2[l_{c_1}(\tilde{\beta}) + l_{c_2}(\tilde{\beta}) + l_{c_3}(\tilde{\beta}) + l_{c_4}(\tilde{\beta}) - l_c(\hat{\beta})] \approx \chi^2_3.$$

$$\lambda_{obs} = 2[-17.62 - 13.90 - 3.52 - 11.06 - (-46.44)] = 0.70.$$

p-value=0.87. Cannot reject H_0. Separate Cox models for the 4 cell types are no better than the stratified Cox model.

How to estimate $S_j(t|z)$ for strata j based on the stratified Cox model in R? (How is $S_j(t|z)$ related to the baseline survival function of strata j?)

- Run "survfit" for the fit of the stratified Cox model, setting "newdata" at the specific covariate value (or values). This gives $\hat{S}_j(t|z)$ for strata $j = 1, \ldots, J$. Plot if needed.
- If "newdata" is not specified, by default, the survival functions by strata are estimated at the mean covariate values (mean of each covariate is over the entire data).

The figure on the next page gives the estimated survival functions for cell types 1, 2, 3, 4, for those with mean performance status (56.25 over all patients).

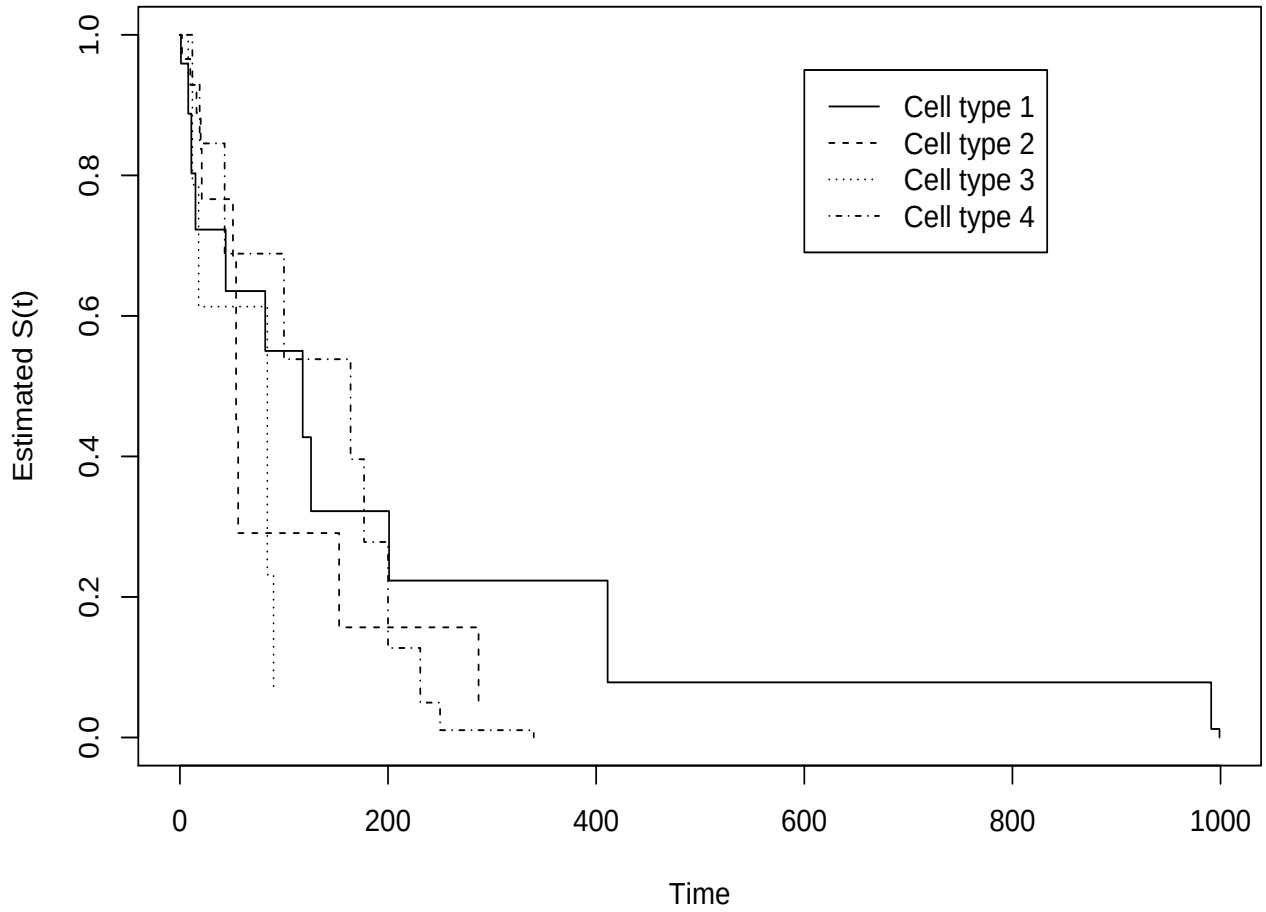


Figure: Estimated survival functions for cell types 1, 2, 3, 4, for those with mean performance status (56.25 over all patients).

```
>  library(survival)
>  lung<-read.table("eg541.txt",header=T)
>  # Stratified Cox model.
>  fits=coxph(Surv(time,status)~perfstat+strata(celltype),data=lung)
>  print(fits)
             coef exp(coef) se(coef)     z        p
perfstat -0.06530   0.93678  0.01529 -4.27 1.95e-05
Likelihood ratio test=22.71  on 1 df, p=1.887e-06
n= 40, number of events= 37
>  fits$loglik
[1] -57.79794 -46.44464
>
>  # Cox model for each cell type.
>  fits1=coxph(Surv(time,status)~perfstat+strata(celltype),data=lung,
subset=celltype==1)
>  print(fits1)
             coef exp(coef) se(coef)      z       p
perfstat -0.06369   0.93829  0.02457 -2.592 0.00954
Likelihood ratio test=7.77  on 1 df, p=0.005323
n= 14, number of events= 12
>  fits1$loglik
[1] -21.50234 -17.61926
```

```
>  fits2=coxph(Surv(time,status)~perfstat+strata(celltype),data=lung,
subset=celltype==2)
>  print(fits2)
             coef exp(coef) se(coef)      z      p
perfstat -0.05260   0.94876  0.02666 -1.973 0.0485
Likelihood ratio test=5.02  on 1 df, p=0.02511
n= 11, number of events= 10
>  fits2$loglik
[1] -16.40370 -13.89568
>
>  fits3=coxph(Surv(time,status)~perfstat+strata(celltype),data=lung,
subset=celltype==3)
>  print(fits3)
             coef exp(coef) se(coef)      z     p
perfstat -0.05930   0.94242  0.04292 -1.382 0.167
Likelihood ratio test=2.54  on 1 df, p=0.111
n= 5, number of events= 5
>  fits3$loglik
[1] -4.787492 -3.517180
```

```
>  fits4=coxph(Surv(time,status)~perfstat+strata(celltype),data=lung,
subset=celltype==4)
>  print(fits4)
             coef exp(coef) se(coef)      z      p
perfstat -0.08792   0.91584  0.03433 -2.561 0.0104
Likelihood ratio test=8.09  on 1 df, p=0.004455
n= 10, number of events= 10
>  fits4$loglik
[1] -15.10441 -11.06013
>
> # Test to compare stratified Cox model and
> # separate Cox model for each cell type.
> 2*(fits1$loglik[2]+fits2$loglik[2]+fits3$loglik[2]+fits4$loglik[2]
-fits$loglik[2])
[1] 0.7047782
> 1-pchisq(0.705,3)
[1] 0.8720273
```

```
>  # Estimate survival functions by strata,
>  # at (the default) mean PS value, based on
>  # the stratified Cox model.
>  plot(survfit(fits),xlab="Time",ylab="Estimated S(t)",
lty=1:4)
>  legend(600,0.95, legend=c("Cell type 1","Cell type 2",
"Cell type 3","Cell type 4"),lty=1:4)
>  # Same plot as
>  # plot(survfit(fits,newdata=data.frame(perfstat
=mean(lung$perfstat)),lty=1:4))
>  # Mean performance status, over all patients.
>  mean(lung$perfstat)
[1] 56.75
```

Remarks:

- If the PH assumption is violated for a covariate in the Cox model, one way to relax the assumption is to stratify by this covariate and fit the stratified Cox model.
- Stratified Cox models are suitable for studies where subjects from different strata (for example, hospitals) may have different baseline survival distributions.
- The stratified Cox model estimates "weighted" log hazard ratios (or effects of the covariates), each effect is weighted across all strata.
- The effect of the stratifying variable cannot be studied. The stratifying variable usually should be a "nuisance" variable. For example, do not stratify by the treatment variable in a clinical trial.

Interaction Term in Cox Regression Model

Two covariates z_1, z_2 may jointly affect the survival time distribution, so that the hazard function has the form

$$h(t|z) = h_0(t)\exp(\beta_1 z_1 + \beta_2 z_2 + \beta_{12} z_1 z_2).$$

The corresponding coefficients are called
the main effects: β_1, β_2
and the interaction effect: β_{12}.

What does interaction effect describe?

Interaction between continuous and categorical covariates

Recall the covariates defined for the lung cancer data (Example 6.2.1),
z_{i1}= performance status (PS), z_{i2}= age, ...,
$z_{i7} = I(\text{trt} = 2)$ for treatment categories, for patient i.

Model 4: The hazard function of failure time T_i has the form
$h(t|z_i) = h_0(t)\exp(\beta_1 z_{i1} + \beta_2 z_{i2} + \beta_7 z_{i7} + \beta_{27} z_{i2} z_{i7})$.

Interpret the interaction effect, β_{27}, between age and treatment.

Interaction between 2 categorical covariates

Suppose we are interested in two age groups, and define $z_{i2}^* = I(\text{age} >= 60)$.

Model 5: The hazard function of failure time T_i has the form
$h(t|z_i) = h_0(t)\exp(\beta_1 z_{i1} + \beta_2^* z_{i2}^* + \beta_7 z_{i7} + \beta_{27}^* z_{i2}^* z_{i7})$.

Interpret the interaction effect, β_{27}^*, between age group and treatment.

Exercise: Fit Model 4 and Model 5 to the lung cancer data. Is the interaction effect significant in each model?

Remarks:

- Interaction terms may be useful for improving the model, in the presence of model inadequacy, or violations of PH assumption.
- By convention, if a model includes the effect of the interaction between two covariates ($\beta_{12}z_1z_2$), it must include the main effects for both covariates ($\beta_1 z_1$ and $\beta_2 z_2$).
- Refrain from using overly complicated interaction terms in your models, they can be hard to interpret.

Treatment-biomarker interaction, personalized medicine

Is treatment effect the same for patients of any ages?

Model 4:

Model 5: