Applied Statistics

J. Blanchet and J. Wadsworth

Institute of Mathematics, Analysis, and Applications
EPF Lausanne

An MSc Course for Applied Mathematicians, Fall 2012

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Outline

Part I

Introduction and Basic Notions

Introduction

- In **Survival Analysis** we are concerned with modelling times until events
- **Time to event modelling** is widely applied in many fields of scientific research; applications arise in contexts where we are interested in modelling the time until a certain change of state occurs in a certain subject.
- Time to event modelling takes different names depending on the sort of application; for example:
  - **Survival Analysis**: Biostatistics (e.g. time until death);
  - Reliability Analysis: Engineering (e.g. time until failure);
  - Duration Analysis: Econometrics (e.g. time in unemployment).
  - . . .

Example problem

Example

- Suppose interest lies in testing the efficacy of a new drug for extending the lifetimes of terminally ill patients
- From the start of the trial we would record the times until death of the patients
- We are then interested in estimating the distribution of these times, and whether the distribution is different for those taking the drug

Features of survival / reliability data

- Data on event times are collected during an **observation period**
- Usually the observation period lasts for a **fixed time length**
- During the period, 'events' (e.g. death) will occur for some subjects, but not necessarily all
- Survival data may therefore be **right-censored**, meaning the data we collect are

$$X_i = \min\{T_i, c\}$$

where T_i is the time of the event and our observation period is $(0, c]$.

Features of survival / reliability data

- Define a **non-censoring indicator**

$$\delta_i = I(T_i < c)$$

- Then we can write

$$X_i = T_i \delta_i + c(1 - \delta_i)$$

- The observation is censored if the true survival time $t_i > c$, otherwise we observe t_i
- In some cases **data may also be censored at other times** if, for example, a patient is lost from a study, and all we know is that they were alive when they dropped out.

Introduction: basic notation

- Let T denote a positive random variable which models the time to the event of interest
- Define the **survival function**

$$S(t) \equiv 1 - F(t) = P(T > t)$$

- The probability density function is $f(t) = -\mathrm{d}S(t)/\mathrm{d}t$.

Introduction: basic notation

- The **hazard function** assesses the instantaneous risk of demise at time t, conditional on survival to that time:

$$\lambda(t) = \lim_{h \to 0} \frac{P(t < T \leq t + h | T > t)}{h} = \frac{f(t)}{S(t)} = -\frac{\mathrm{d} \log S(t)}{\mathrm{d}t}.$$

  Hence,

$$S(t) = \exp\left(-\int_0^t \lambda(u)\mathrm{d}u\right).$$

- The density of T can be rewritten as

$$f(t) = \lambda(t)S(t).$$

Example problem

Example

- In the example above:
  - T: time until death
  - $S(t)$: probability of being alive at time t
  - $\lambda(t)$: approximate risk of dying in a short period $(t + h)$, given that an individual as remained alive until time t.

Part II

Kaplan–Meier Product Limit Estimator

Kaplan–Meier Estimator

- We begin with **nonparametric estimation of the survival function**
- The estimator we study is the **Kaplan–Meier estimator**, which is the Nonparametric Maximum Likelihood Estimator (NPMLE) of the population survivor function $S(t)$ (Kaplan and Meier, 1958)
- The Kaplan–Meier estimator is analogous to the empirical distribution function, but **accounts for censored data**
- If censoring is not accounted for we will get a **biased estimate**
- In a time to event context, treating censored data as non-censored data will naturally lead to an underestimation of the real duration of interest.

Kaplan–Meier Estimator

Let

- $t_{(i)}$ denote the ordered times recorded for both censored and uncensored observations, i.e.,

$$0 \leq t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)},$$

  $\delta_{(i)}$ its corresponding non-censoring indicator (1 if uncensored, 0 if censored).
- $n_{(i)}$ denote the number of subjects at risk just prior to time $t_{(i)}$
- $d_{(i)}$ denote the number experiencing an 'event' (e.g. death) at time $t_{(i)}$
- The Kaplan–Meier product-limit estimator is defined as:

$$\widehat{S}(t) = 1 - \widehat{F}(t) = \prod_{\{i:t_{(i)} \leq t\}} \left[\frac{n_{(i)} - d_{(i)}}{n_{(i)}} \right]^{\delta_{(i)}},$$

Kaplan–Meier Estimator

Example

Deaths (events): 0.5; 1.2; 2.2; 3.4; 8.3 months
Losses from study: 1.8; 6.0 months.

	Recorded Times	Status	KM-estimate
$n_{(i)}$	$t_{(i)}$	$\delta_{(i)}$	$\widehat{S}(t_{(i)})$
7	0.5	1	0.857
6	1.2	1	0.714
5	1.8	0	0.714
4	2.2	1	0.534
3	3.4	1	0.357
2	6.0	0	0.357
1	8.3	1	0.000

Kaplan–Meier Estimator

`Surv{survival}` + `survfit{survival}`

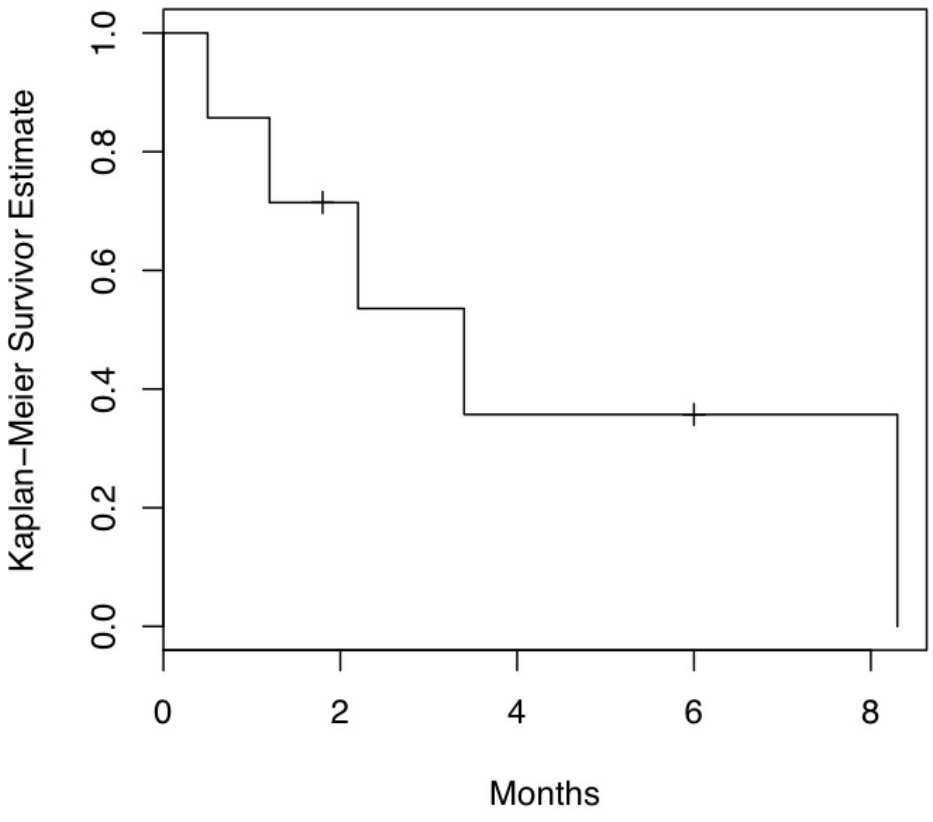


Figure: The Kaplan–Meier curve

Kaplan–Meier Estimator

Surv{survival} + survfit{survival}

```
 ## install.packages("survival")
 library(survival)
load("surv.Rdata")
 time <- surv[,1]
 status <- surv[,2]

 ## obtain KM-estimates
 fit <- survfit(Surv(time, status)~1, type="kaplan-meier",
                conf.type="none")

 ## Plot the KM survival function
 plot(fit,xlab="Months",ylab="Kaplan--Meier Survivor Estimate")
```

Kaplan–Meier Estimator

- Note how the sizes of the jumps change after censoring
- If no censored event occurs between two uncensored events, then the denominator in the fraction stays the same e.g.

$$\cdots \times \left(\frac{10-1}{10}\right)^{1} \times \left(\frac{9-1}{9}\right)^{1} \times \cdots$$

- If there is censoring between two uncensored events then this will not be the case

$$\cdots \times \left(\frac{10-1}{10}\right)^{1} \times \left(\frac{9-1}{9}\right)^{0} \times \left(\frac{8-1}{8}\right)^{1} \times \cdots$$

- This is effectively how the censoring is accounted for: we do not observe a decrease at the time point that the censoring occurs, but the size of the next decrease is larger than the previous one

Part III

Regression models for Time to Event Data

Introduction

- Let T denote the time to the event of interest.
- We want to link T to some observed covariates $\boldsymbol{z}$.

Example

- T: months in unemployment;
- z_1: sex;
- z_2: age;
- z_3: qualification;
- $S(t; \mathbf{z})$: probability of being unemployed more than t months for a person with a given sex, a given age and a given qualification (i.e. given $\mathbf{z}$);
- $\lambda(t; \mathbf{z})$, $f(t; \mathbf{z})$.

- How can we link T to $\mathbf{z}$?

Proportional-Hazards Model

- **Proportional hazards model** specifies that

$$\lambda(t;\mathbf{z}) = \lambda_0(t)e^{\mathbf{z}\beta},$$

  where $\lambda_0(t)$ is a **baseline hazard function**.
- The covariates act multiplicatively on the hazard function.
- The conditional density function of T given $\mathbf{z}$ is

$$f(t|\mathbf{z}) = \lambda_0(t)e^{\mathbf{z}\beta}\exp\left\{-e^{\mathbf{z}\beta}\int_0^t \lambda_0(u)\,du\right\}$$

- The conditional survivor function for T given $\mathbf{z}$ is

$$S(t|\mathbf{z}) = [S_0(t)]^{\exp(\mathbf{z}\beta)}, \quad \text{where } S_0(t) = \exp\left\{-\int_0^t \lambda_0(u)\,du\right\}.$$

- Proportional hazards property: the hazard ratio

$$\frac{\lambda(t;\mathbf{z}_i)}{\lambda(t;\mathbf{z}_j)} = \frac{\lambda_0(t)e^{\mathbf{z}_i\beta}}{\lambda_0(t)e^{\mathbf{z}_j\beta}} = \frac{e^{\mathbf{z}_i\beta}}{e^{\mathbf{z}_j\beta}}$$

  is independent of time t.

Modeling the baseline function
Parametric models

- What about $\lambda_0(t)$?
- First possibility: impose some parametric model.
- Among the most common models are:
  - **exponential regression model**: $\lambda_0(t) = \lambda$ and then

$$\begin{aligned} f(t|\mathbf{z}) &= \lambda e^{\mathbf{z}\beta} \exp(-\lambda t e^{\mathbf{z}\beta}), \quad t \geq 0. \\ T|\mathbf{Z} = \mathbf{z} &\sim \text{Exp}(\lambda e^{\mathbf{z}\beta}) \end{aligned}$$

  - **Weibull regression model**: $\lambda_0(t) = a\lambda^a t^{a-1}$, $a, \lambda > 0$

$$f(t|\mathbf{z}) = a\lambda^a t^{a-1} e^{\mathbf{z}\beta} \exp\{-(\lambda t)^a e^{\mathbf{z}\beta}\}, \quad t \geq 0.$$

- However it imposes a strong structure to the data, which may be incorrect.

Modeling the baseline function
Parametric models

- Second possibility: $\lambda_0(t)$ is left arbitrary; it can take any form. This is **Cox Proportional-Hazards model**.
- This approach has the advantage of not making arbitrary, and possibly incorrect, assumptions about the form of the baseline hazard function.

Estimation of Parametric Proportional-Hazards model

Censored Likelihood

- Let θ denotes the set of parameters to be estimated.
  E.g. $\theta = \{a, \lambda, \beta\}$ in Weibull regression model.
- The conditional pdf and conditional survival functions are denoted $f(t|\mathbf{z}, \theta)$ and $S(t|\mathbf{z}, \theta)$.
- We consider n iid random pairs (X_i, δ_i) where $X_i = \min\{T_i, c\}$ and $\delta_i = I(T_i < c)$.
- The likelihood function for observations $\{(x_1, \delta_1), \ldots, (x_n, \delta_n)\}$ is

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\mathbf{z_i}, \theta)^{\delta_i} \; S(x_i|\mathbf{z_i}, \theta)^{1-\delta_i}.$$

- We seek θ maximizing

$$\log L(\theta) = \sum_{u} \log f(x_i|\mathbf{z_i}, \theta) + \sum_{c} \log S(x_i|\mathbf{z_i}, \theta).$$

  where u and c mean sums over uncensored and censored observations.

Estimation of Cox Proportional-Hazards Model

Partial Likelihood

- As the baseline hazard function is not specified, the likelihood function cannot be fully specified. Cox (1975) defines a likelihood based on conditional probabilities which are free of $\lambda_0(t)$; he called it **partial likelihood function**.
- For the iid observations $\{(x_1, \delta_1), \ldots, (x_n, \delta_n)\}$, the partial likelihood function is given by

$$L_p(\beta) = \prod_{i=1}^{n} \left(\frac{e^{z_i \beta}}{\sum_{j \in \mathcal{R}(x_i)} e^{z_j \beta}} \right)^{\delta_i},$$

  where $\mathcal{R}(x_i) = \{j / x_j \geq x_i\}$ (set of individuals who are "at risk" for failure at time x_i, i.e. who are alive just before x_i).
- Efron (1974) proposed a modified version of L_p which is considered to give better results when there are ties in the time data.
- Cox's estimates maximize the log-partial likelihood, which gives $\hat{\beta}$.
- Estimation of $\lambda_0(t)$ is done after-the-fact and based on $\hat{\beta}$: Breslow or Kalbfleisch-Prentice estimator.

Partial Likelihood Inference

- Analogous to standard likelihood theory, it can be shown (though not easily) that

$$\frac{\hat{\beta} - \beta}{se(\hat{\beta})} \sim N(0, 1).$$

- The variance (and se) of $\hat{\beta}$ can be obtained by inverting the second derivative of the partial likelihood,

$$var(\hat{\beta}) = \left\{ -\frac{\partial^2 l(\beta)}{\partial \beta^2} \right\}.$$

As $var(\hat{\beta})$ is a function of the true β which is unknown, we calculate the "observed information" matrix by replacing β by its estimate $\hat{\beta}$ in $var(\hat{\beta})$ above.

Example: Recidivism

- The data set is originally from Rossi et al. (1980).
  Rossi, P. H., R. A. Berk & K. J. Lenihan. 1980. Money, Work and Crime: Some Experimental Results. New York: Academic Press.
- The file 'Rossi.txt' contains data from an experimental study of recidivism of 432 male prisoners, who were observed for a year after being released from Maryland state prisons in the 1970s. Half the released convicts were given financial aid; half did not receive aid.
- We seek to see if allowing financial aid after release from prison ('fin=1') decreases recidivism rate.

```
Rossi <- read.table("Rossi.txt", header=T)
Rossi[1:5, 1:10]
```

```
  week arrest fin age race wexp mar paro prio educ
1   20      1   0  27    1    0   0    1    3    3
2   17      1   0  18    1    0   0    1    8    4
3   25      1   0  19    0    1   0    1   13    3
4   52      0   1  23    1    1   1    1    1    5
5   52      0   0  19    0    1   0    1    3    3
```

Example: Recidivism

- "week": week of first arrest after release or censoring
- “arrest“: 1 if arrested, 0 if not arrested
- ”fin”: financial aid; no/yes
- “age“ in years at time of release
- ”race” black or other
- “wexp“ full-time work experience before incarceration: no or yes
- ”mar” marital status at time of release: married or not married
- “paro“ released on parole? no or yes
- ”prio” number of convictions prior to current incarceration
- “educ“ level of education: 2 = 6th grade or less; 3 = 7th to 9th grade; 4 = 10th to 11th grade; 5 = 12th grade; 6 = some college

Example: Recidivism

coxph {survival}

```
mod <- coxph(Surv(week, arrest) ~ fin + age + race + wexp + mar
     + paro + prio,data=Rossi)
summary(mod)

n= 432, number of events= 114

         coef exp(coef) se(coef)      z Pr(>|z|)
fin  -0.37942   0.68426  0.19138 -1.983  0.04742 *
age  -0.05744   0.94418  0.02200 -2.611  0.00903 **
race  0.31390   1.36875  0.30799  1.019  0.30812
wexp -0.14980   0.86088  0.21222 -0.706  0.48029
mar  -0.43370   0.64810  0.38187 -1.136  0.25606
paro -0.08487   0.91863  0.19576 -0.434  0.66461
prio  0.09150   1.09581  0.02865  3.194  0.00140 **
```

The covariates 'age' and 'prio' (prior convictions) have highly statistically significant coefficients, while the coefficient for 'fin' (financial aid - the focus of the study) is marginally significant.

Example: Recidivism

`survfit {survival}`

```
plot(survfit(mod), xlab="Weeks", ylab="Proportion Not Rearrested")
```

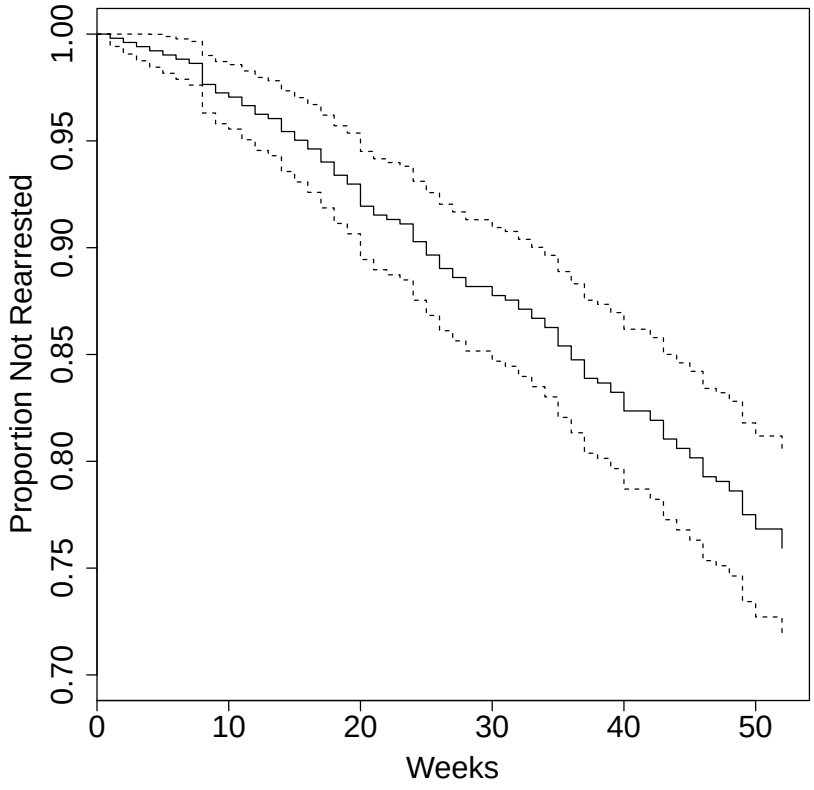


Figure: Estimated survival function $\hat{S}(t)$ at the mean values of the covariates.

Example: Recidivism

To distinguish between those receiving (fin = 1) and not receiving (fin = 0) financial aid, we create an artificial data set with fin=0/1 and the other covariates fixed at their mean values.

```
Rossi.fin <- data.frame(
      fin=c(0,1), age=rep(mean(Rossi$age),2),
      race=rep(mean(Rossi$race),2),wexp=rep(mean(Rossi$wexp),2),
      mar=rep(mean(Rossi$mar),2), paro=rep(mean(Rossi$paro),2),
      prio=rep(mean(Rossi$prio),2)
)

Rossi.fin
  fin      age      race      wexp       mar      paro     prio
1   0 24.59722 0.8773148 0.5717593 0.1226852 0.6180556 2.983796
2   1 24.59722 0.8773148 0.5717593 0.1226852 0.6180556 2.983796
```

Example: Recidivism

`survfit {survival}`

```
plot(survfit(mod, newdata=Rossi.fin), conf.int=T, col=c("red","blue"))
```

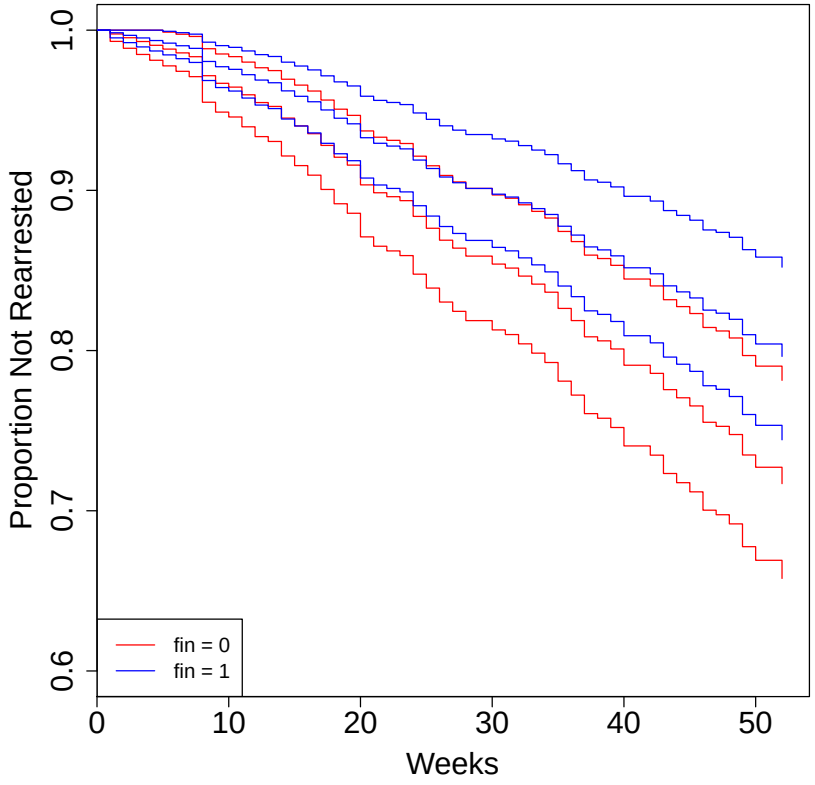


Figure: Estimated survival function $\hat{S}(t)$ for those receiving (fin = 1) and not receiving (fin = 0) financial aid. Other covariates are fixed at the mean values.

Cumulative hazard function

- The **cumulative hazard function** at time t represents the total hazard an individual is exposed to up to time t:

$$\Lambda(t;\mathbf{z}) = \int_0^t \lambda(u;\mathbf{z})\,\mathrm{d}u$$

- For Proportional-Hazards model, this gives

$$\Lambda(t;\mathbf{z}) = e^{\mathbf{z}\beta} \int_0^t \lambda_0(u)\,\mathrm{d}u$$

- For the exponential regression model,

$$\Lambda(t;\mathbf{z}) = e^{\mathbf{z}\beta}\lambda t$$

  The plot of $\Lambda(t;\mathbf{z})$ vs. t should give a straight line through the origin.

- For the Weibull regression model,

$$\Lambda(t;\mathbf{z}) = e^{\mathbf{z}\beta} \int_0^t a\lambda^a u^{a-1}\,\mathrm{d}u = e^{\mathbf{z}\beta}(\lambda t)^a.$$

  The plot of $\log \Lambda(t;\mathbf{z})$ vs. $\log t$ should give a straight line.

Cumulative hazard function

```
bash<-basehaz(mod)
plot(bash[,"time"],bash[,"hazard"],type="l")
plot(log(bash[,"time"]),log(bash[,"hazard"]),type="l")
```

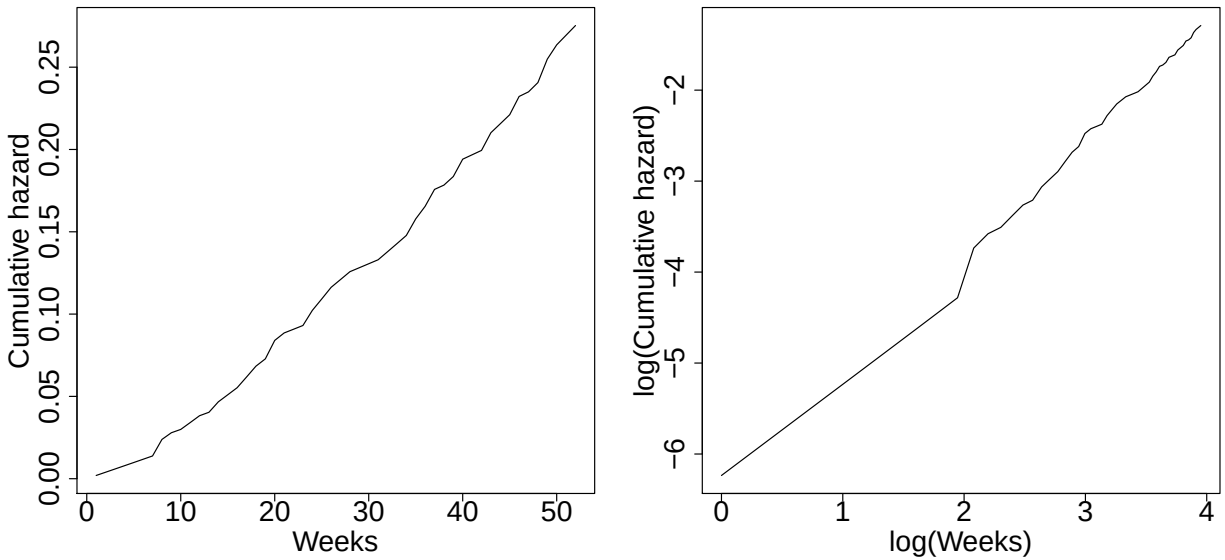


Figure: Estimated cumulative hazard function with Cox model. The covariates are fixed at the mean values.

The figure shows that Weibull model might be reasonable.

Cumulative hazard function

```
wei = survreg(Surv(week,arrest) ~ fin + age + race + wexp + mar
  + paro + prio,data=Rossi)
summary(wei)
```

```
               Value Std. Error      z        p
(Intercept)   3.9901     0.4191  9.521 1.72e-21
fin           0.2722     0.1380  1.973 4.85e-02
age           0.0407     0.0160  2.544 1.10e-02
race         -0.2248     0.2202 -1.021 3.07e-01
wexp          0.1066     0.1515  0.703 4.82e-01
mar           0.3113     0.2733  1.139 2.55e-01
paro          0.0588     0.1396  0.421 6.74e-01
prio         -0.0658     0.0209 -3.143 1.67e-03
Log(scale)   -0.3391     0.0890 -3.809 1.39e-04

Scale= 0.712
```

As in Cox model, the covariates 'age' and 'prio' have highly statistically significant coefficients, while the coefficient for 'fin' is marginally significant. The scale a is < 1.

Example: Recidivism

`predict, survfit{survival}=`

Plot the two fits (Weibull and Cox models) with 95% confidence interval.

```
p<-seq(0,0.99,length=100)
pred.wei<-predict(wei,newdata=Rossi.fin,type="quantile",
  p=1-p, se.fit=TRUE)
plot(c(1,100),c(0,1),type="n")
lines(pred.wei$fit[1,],p,lty=2,lwd=2,col="red")
lines(pred.wei$fit[1,]-1.96*pred.wei$se.fit[1,],p,lty=2,col="red")
lines(pred.wei$fit[1,]+1.96*pred.wei$se.fit[1,],p,lty=2,col="red")
lines(pred.wei$fit[2,]-1.96*pred.wei$se.fit[2,],p,lty=2,col="blue")
lines(pred.wei$fit[2,]+1.96*pred.wei$se.fit[2,],p,lty=2,col="blue")
lines(survfit(mod, newdata=Rossi.fin), conf.int=F, col=c("red","blue")
legend("bottomleft", legend=c('fin = 0 Cox', 'fin = 1 Cox',
  'fin = 0 Weibull', 'fin = 1 Weibull'),
  col=c("red","blue","red","blue"), lty=c(1,1,2,2))
```

Example: Recidivism

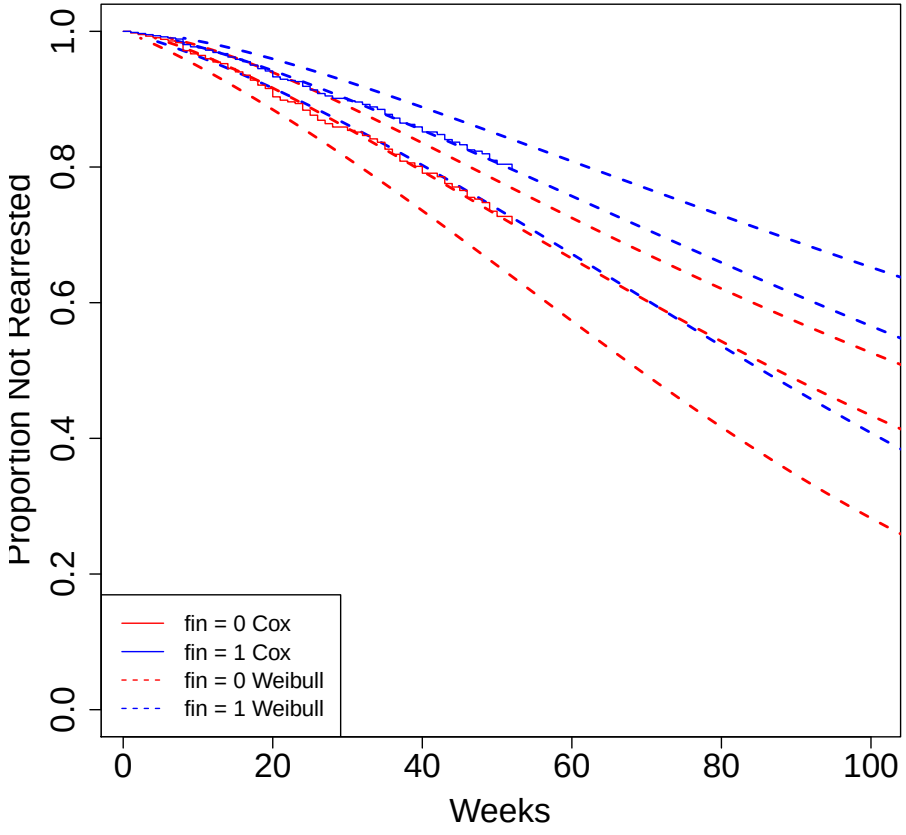


Figure: Estimated survival function with Cox and Weibull models. The covariates are fixed at the mean values.

Both models give similar results but Weibull regression model allows extrapolation.

Checking Proportional Hazards

- Model checking in regression models is based on the usual 'observed-predicted' residuals. However, there is no obvious analog in PH models because the value of the outcome is unknown.
- Schoenfeld (1982) proposed the first set of residuals for use with a fitted PH model; we refer to them as the **Schoenfeld residuals**.
- They are based on the individual contribution of each covariate β_k to the derivative of the log partial likelihood:

$$\frac{\partial l_p(\beta_k)}{\partial \beta} = \sum_{i=1}^{n} \delta_i \left\{ z_{ik} - \frac{\sum_{j \in \mathcal{R}(x_i)} z_{jk}\, e^{z_j \beta}}{\sum_{j \in \mathcal{R}(x_i)} e^{z_j \beta}} \right\} = \sum_{i=1}^{n} \delta_i (z_{ik} - \bar{z}_{x_i k}).$$

- The estimator of the Schoenfeld residuals for the ith subject on the kth covariate is

$$\hat{r}_{ik} = \delta_i (z_{ik} - \hat{\bar{z}}_{x_i k}),$$

  where $\hat{\bar{z}}_{x_i k}$ is given by the eq. above with β replaced by $\hat{\beta}$.
- For PH model, the residuals $\hat{r}_{ik}$ should exhibit a random (i.e. unsystematic) pattern at each failure time. Otherwise it suggests that as time passes, the covariate effect is changing.

Checking Proportional hazards

`cox.zph {survival}`

Scaled-down version of the Cox regression model fit to the recidivism data, eliminating the covariates whose coefficients were not statistically significant:

```
mod2 <- coxph(Surv(week, arrest) ~ fin + age + prio, data=Rossi)
summary(mod2)

coef exp(coef) se(coef)      z Pr(>|z|)
fin  -0.34695   0.70684  0.19025 -1.824 0.068197 .
age  -0.06711   0.93510  0.02085 -3.218 0.001289 **
prio  0.09689   1.10174  0.02725  3.555 0.000378 ***

cox.zph(mod2)

          rho   chisq      p
fin    -0.00657 0.00507 0.9433
age    -0.20976 6.54147 0.0105
prio   -0.08004 0.77288 0.3793
GLOBAL       NA 7.13046 0.0679
```

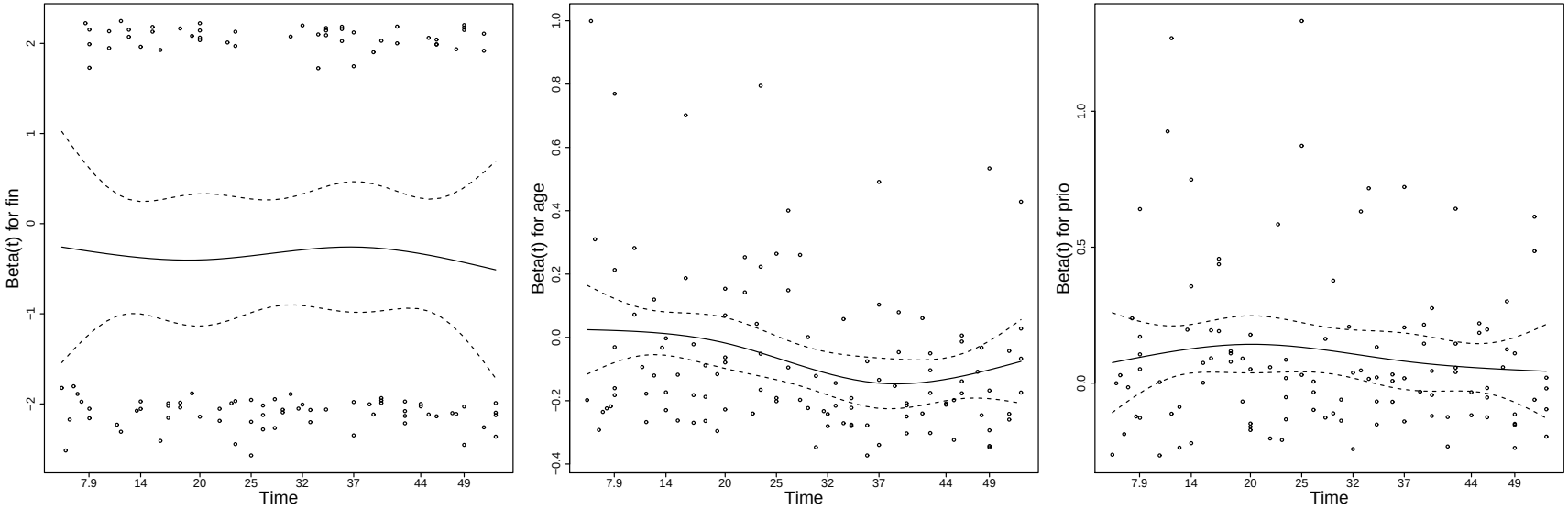
There is strong evidence of non-proportional hazards for age, while the global test (on 3 df) is not quite statistically significant.

Checking Proportional Hazards

`cox.zph {survival}`

Graphs of the scaled Schoenfeld residuals against time:

```
par(mfrow=c(2,2))
plot(cox.zph(mod2))
```



There appears to be a trend in the plot for age, with the age effect declining with time.

Stratified Cox PH model

`strata, coxph {survival}`

- We divide the age into *strata*.
- Each stratum is permitted to have a different baseline hazard function, while the coefficients β are assumed to be constant across strata.

```
Rossi$age.cat<-Rossi$age
Rossi$age.cat[Rossi$age<=19]<-1
Rossi$age.cat[Rossi$age>19 & Rossi$age<=25]<-2
Rossi$age.cat[Rossi$age>25 & Rossi$age<=30]<-3
Rossi$age.cat[Rossi$age>30]<-4

mod3<-coxph(Surv(week, arrest) ~ fin + strata(age.cat) + prio,
  data=Rossi)
summary(mod3)

n= 432, number of events= 114

        coef exp(coef) se(coef)      z Pr(>|z|)
fin  -0.34080   0.71120  0.19020 -1.792 0.073167 .
prio  0.09407   1.09864  0.02701  3.482 0.000497 ***
```

Stratified Cox PH model

`cox.zph {survival}`

```
cox.zph(mod3)

          rho  chisq     p
fin    -0.0183 0.0392 0.843
prio   -0.0771 0.6859 0.408
GLOBAL      NA 0.7299 0.694
```

There is no evidence of non-proportional hazards for the remaining covariates.

Cumulative hazard for the strata

```
bash<-basehaz(mod3)
id1<-which(bash["strata"]=="age.cat=1")
id2<-which(bash["strata"]=="age.cat=2")
id3<-which(bash["strata"]=="age.cat=3")
id4<-which(bash["strata"]=="age.cat=4")
plot(bash[,"time"],bash[,"hazard"],type="n")
lines(bash[id1,"time"],bash[id1,"hazard"],col=1)
lines(bash[id2,"time"],bash[id2,"hazard"],col=2)
lines(bash[id3,"time"],bash[id3,"hazard"],col=3)
lines(bash[id4,"time"],bash[id4,"hazard"],col=4)
legend("topleft",c("cat1","cat2","cat3","cat4"),col=1:4,lty=1)
```

Cumulative hazard function

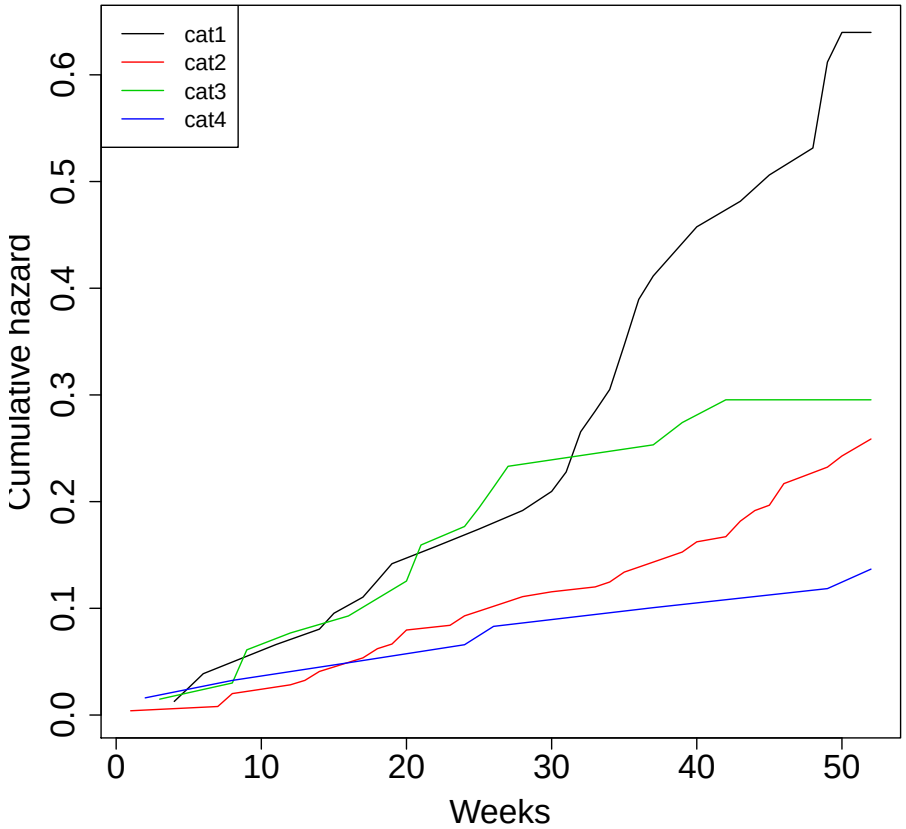


Figure: Estimated cumulative hazard function with Cox model, for the different age strata. The covariates are fixed at the mean values.

The figure shows that young people tend to be more frequently rearrested after 30-50 weeks.