Proportional hazards (Chapter 8)

The techniques in chapter 7 for comparing two groups are most useful when the two groups are similar (i.e., subjects from the same population but given different treatments).

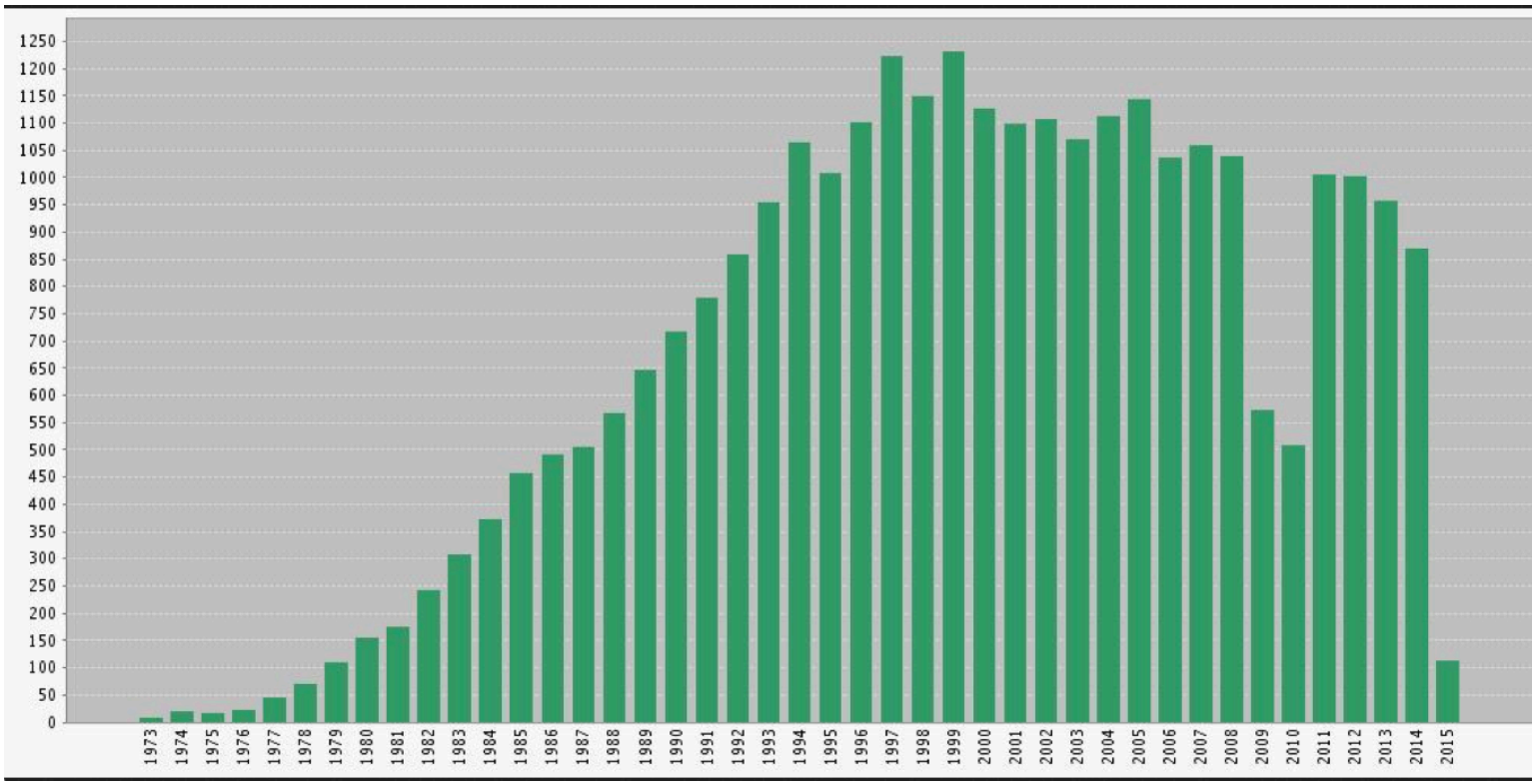
Often two groups differ also on other covariates, such as age, gender, level of education, blood pressure, alcohol use, etc., and we want to adjust for some of these covariates before comparing survival experiences.

Proportional hazards

In addition to comparing survival rates, we might wish to know which covariates count as risk factors. This is similar to regression problems where we want to know which factors (covariates) are significant in the model.

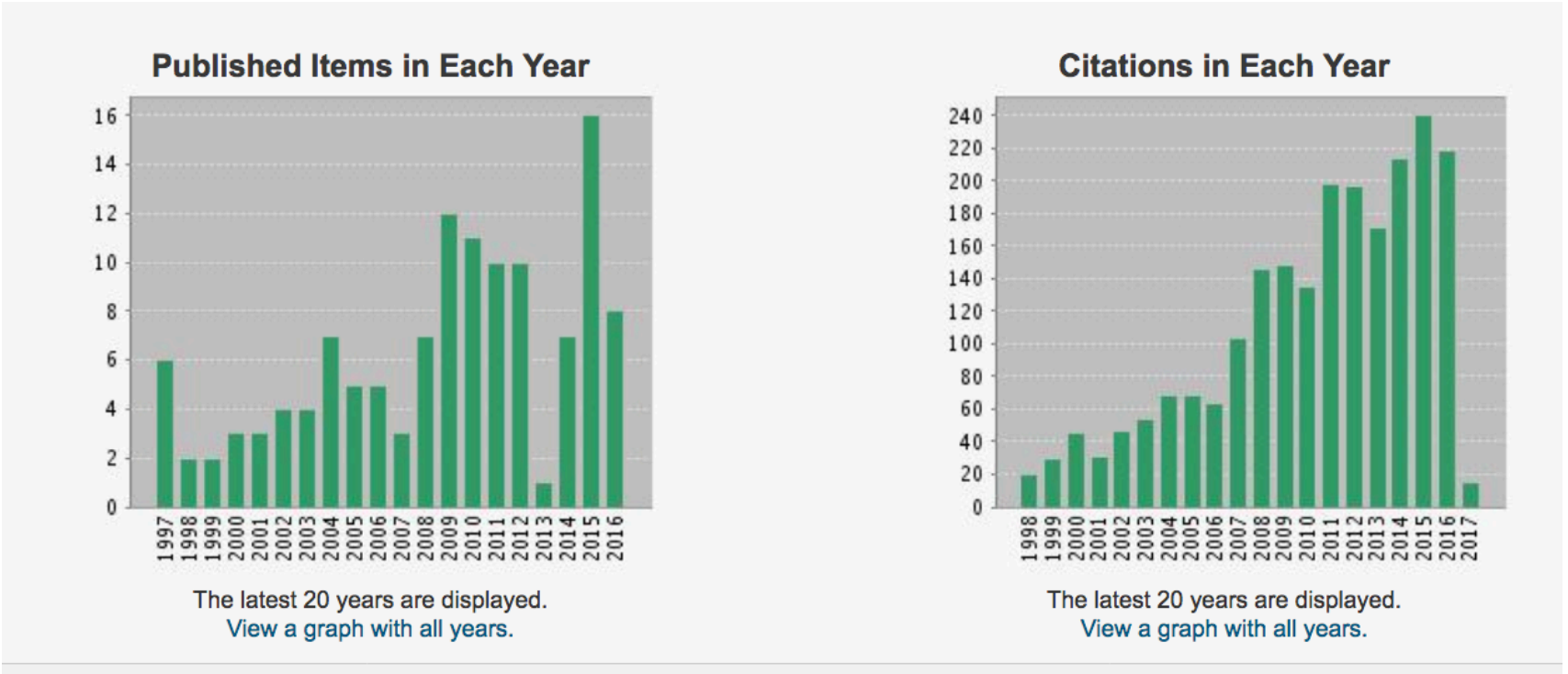
This is most frequently done using the Cox Proportional hazards technique from 1972, in a paper called "Regression models and life-tables" *The Journal of the Royal Statistical Society B*. This paper has been cited thousands of times and has been extremely influential.

Proportional hazards model: citations from Web of Knowledge



Bayesian nonparametric survival analysis: citations from Web of Knowledge

140 papers found with this topic.



Proportional hazards model

For the proportional hazards framework, the data consists of event times T_j, censoring information, δ_j, and a vector of covariates $\mathbf{Z}_j(t)$. Here j indexes the observation number so that $j = 1, \ldots, n$.

The vector of covariates has p variables, so we have $\mathbf{Z}_j(t) = (Z_{j1}(t), Z_{j2}(t), \ldots, Z_{jp}(t))'$ as a vector covariates associated with each observation, where the covariates might change over time. The covariates might include group membership as well as continuous variables such as age and blood pressure. Some covariates, for example sex, might not change over time. Generally, it is easier to have fixed covariates, so that $\mathbf{Z}_j(t) = \mathbf{Z}_j$, which we'll study first.

Proportional hazards model

The hazard rate depends on the covariates, so we have $h(t|\mathbf{Z})$ as the covariate-adjusted hazard rate at time t. The Cox Proportional hazards model treats the hazard adjusted for the covariates as being a multiple of the baseline hazard:

$$h(t|\mathbf{Z}) = h_0(t)c(\boldsymbol{\beta}'\mathbf{Z})$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is a parameter vector measuring the effects of each covariate (similar to the vector of coefficients in regression), and $c(\cdot)$ is a known (or assumed) function.

Proportional hazards model

This model is called semiparametric because the baseline hazard is treated nonparametrically and the covariates are treated parametrically. A common choice of $c(\cdot)$ is the exponential function, which guarantees that the covariate-adjusted hazard rates are positive:

$$c(\beta'\mathbf{Z}) = \exp(\beta'\mathbf{Z}) = \exp\left(\sum_{i=k}^{p} \beta_k Z_k\right)$$

Thus,

$$h(t|\mathbf{Z}) = h_0(t)\exp\left(\sum_{i=k}^{p} \beta_k Z_k\right)$$

Proportional hazards model

As a consequence,

$$\log\left(\frac{h(t|\mathbf{Z})}{h_0(t)}\right) = \sum_{i=k}^{p} \beta_k Z_k$$

which looks like a linear model. The usual techniques for linear models can be used. For example, if you have k groups to which an individual can belong, you can use dummy variables to code group membership, and interactions can be modeled using products of covariates.

Proportional hazards model

If you compare two individuals with covariates $\mathbf{Z}$ and $\mathbf{Z}^*$, then

$$\frac{h(t|\mathbf{Z})}{h(t|\mathbf{Z}^*)} = \frac{h_0(t)\exp\left(\sum_{i=k}^{p}\beta_k Z_k\right)}{h_0(t)\exp\left(\sum_{i=k}^{p}\beta_k Z_k^*\right)} = \exp\left[\sum_{k=1}^{p}\beta_k(Z_k - Z_k^*)\right]$$

which is constant in t. Thus, the hazard rates for two individuals are proportional (one is a constant multiple of the other over time). Thus, we can think of $\frac{h(t|\mathbf{Z})}{h(t|\mathbf{Z}^*)}$ as the relative risk of someone with risk factor $\mathbf{Z}$ having the event compared to someone with risk factor $\mathbf{Z}^*$.

Proportional hazards model

As a particular example, supposed we want to know the relative risk (also called hazard ratio) for the event for someone smoking versus someone who doesn't smoke, when all other covariates are the same. Then $\exp \sum_{k=1}^{p} \beta_k (Z_k - Z_k^*) = e^{\beta_p}$, where Z_p indicates whether or not a person smokes. The interpretation is similar if $Z_p = 1$ means that a patient received treatment A and $Z_p = 0$ means that the patient received treatment B. If you had two dummy variables, one indicator for each type of treatment, then you would introduce collinearity into the model which would cause problems.

Proportional hazards model

To construct dummy variables, you need $k-1$ 0-1 variables to code k categories. For three levels of a variable (e.g., placebo, treatment A, treatment B), you need to variables, e.g., $Z_1 = 1$ if treatment A is used; otherwise, $Z_1 = 0$, and $Z_2 = 1$ if treatment B is used; otherwise, $Z_2 = 0$. If $Z_1 = Z_2 = 0$, then a placebo was used.

Suppose we have survival data where the only covariate is the treatment (A, B, Placebo). The hazard rates for the three treatments would be

$$\text{hazard for treatment A: } h(t|Z_1 = 1, Z_2 = 0) = h_0(t)\exp(\beta_1)$$
$$\text{hazard for treatment B: } h(t|Z_1 = 0, Z_2 = 1) = h_0(t)\exp(\beta_2)$$
$$\text{hazard for placebo: } h(t|Z_1 = 0, Z_2 = 0) = h_0(t)$$

The relative risk for treatemnt A versus treatment B is

$$\frac{h_0(t)\exp(\beta_1)}{h_0(t)\exp(\beta_2)} = \exp(\beta_1 - \beta_2)$$

Proportional hazards model

Sometimes data are ordinal, and you must make a choice between coding it as quantitative versus categorical. An example is in different doses of a treatment (none=placebo, low, high), or stage of cancer (I, II, III, IV), or education level, (high school, some college, college, graduate degree). Coding these as a single variable and using integers to represent the ordinal value (e.g., 1,2,3,4) is similar to doing a test of trend in the previous chapter. However, for this coding, if the baseline risk is $h_0(t)e^{\beta_1}$ for category 1, then you are assuming that the risks for the other categories are $h_0(t)e^{2\beta_1}$, $h_0(t)e^{3\beta_1}$, and $h_0(t)e^{4\beta_1}$. Thus the relative risk of being in category i versus $i-1$ is $\exp(\beta_1)$.

Alternatively, you could separately estimate the risks for each category, ignoring the fact that the categories are related on an ordinal scale. In this case you use dummy variables for the different levels of risk and estimate them separately. This potentially loses some power to estimate but has the benefit of not assuming equally spaced relative risks.

Proportional hazards model

If you have two or more categorical variables, then you can either create dummy variables for each combination, treating combinations of categories as categories, or have separate dummy variables and then construct interactions.

An example is if you have both sex and race (ethnicity). For example if you have male and female patients and also black and white patients, you could treat the four combinations of sex and race as separate categories. In this case you could have

$$Z_1 = 1, \text{ if black, male}$$
$$Z_2 = 1, \text{ if white, male}$$
$$Z_3 = 1, \text{ if black, female}$$

If $Z_1 = Z_2 = Z_3 = 0$, then the patient is white and female, which would be used in this case for the baseline hazard.

Proportional hazards model

A different approach is to have

$$Z_1 = 1, \text{ if female}$$
$$Z_2 = 1, \text{ if black}$$
$$Z_3 = Z_1 Z_2$$

Therefore $Z_3 = 1$ if the patient is female and black. A white female would have $Z_1 = 1, Z_2 = 0, Z_3 = 0$, a black female would have $Z_1 = Z_2 = Z_3 = 1$, etc. Of course we could include more categories for race. The textbook was written at Ohio State about 25 years ago. In New Mexico, you would typically have a category for Hispanic at least. If you have three categories, you might have $Z_3 = 1$ if a patient is white, then $Z_4 = Z_1 Z_2$ and $Z_5 = Z_1 Z_3$ to get interactions between sex and whether or not a patient was black, and sex and whether or not a patient was white. The baseline case for this example is? ....

Proportional hazards model

.... male Hispanic. This is when $Z_1 = Z_2 = Z_3 = Z_4 = Z_5 = 0$.

Proportional hazards model

To do maximum likelihood, we use something called the partial likelihood. We assume that all death times are unique (there are no ties for the death times), and are $t_1 < \cdots < t_D$. The covariates associated with the ith ordered event time are denoted $Z_{(i)}$, and the kth covariate for this time is $Z_{(i)k}$. Let $R(t_i)$ be the set of individuals at risk at time t_i. Then

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\left[\sum_{k=1}^{p} \beta_k Z_{(i)k}\right]}{\sum_{j \in R(t_i)} \exp\left[\sum_{k=1}^{p} \beta_k Z_{jk}\right]}$$

The log-likelihood is $\ell(\boldsymbol{\beta}) = LL(\boldsymbol{\beta})$:

$$\ell(\boldsymbol{\beta}) = sum_{i=1}^{D} \sum_{k=1}^{p} \beta_k Z_{(i)k} - \sum_{i=1}^{D} \ln \left[\sum_{j \in R(t_i)} \exp\left(\sum_{k=1}^{p} \beta_k Z_{jk} \right) \right]$$

Proportional hazards model

The likelihood here is called a partial likelihood because instead of multiplying the probabilities of the observations, we multiply the probabilities of the observed death times conditional on there being a death at the time. That is, the quantity

$$\frac{\exp\left[\sum_{k=1}^{p}\beta_k Z_{(i)k}\right]}{\sum_{j\in R(t_i)}\exp[\sum_{k=1}^{p}\beta_k Z_{jk}]}$$

represents the probability that someone with covariates $Z_{(i)k}$ dies at time t_i given that at least one person dies at time t_i. We never actually use the probability of an event at time t_i. Instead we multiply these conditional probabilities and treat this as a likelihood.

The equation is related to the idea of competing exponentials. If I have three lightbulbs that burn out at rates 2, 3, and 4 (say 2 times per year, 3 times per year, and 4 times per year), then what is the probability that the rate 4 lightbulb burns out first?

Proportional hazards model

For the competing lightbulbs example, if we used first principles, we could let $f_i(t)$ be the density for the ith lightbulb. For the condition to hold, either the rate 4 lightbulb burns out first and the rate 3 lightbulb second, or the rate 4 lightbulb burns out first and the rate 2 lightbulb burns out second. The probability is therefore

$$\int_0^\infty \int_0^{y_2} \int_0^{y_3} f_2(y_2) f_3(y_3) f_4(y_4) dy_4 dy_3 dy_2 + \int_0^\infty \int_0^{y_3} \int_0^{y_2} f_2(y_2) f_3(y_3) f_4(y_4) dy_4 dy_2 dy_3$$

This is a general solution to the problem which works for any choice of densities $f_i(y)$. However, for exponentials, we get a much easier result. The probability that the rate i lightbulb goes first is its rate divided by the sum of the rates. Thus, the probability that the rate 4 lightbulb dies first is

$$\frac{4}{2+3+4} = \frac{4}{9}$$

Similarly, the probability that the rate 3 lightbulb dies first is $\frac{3}{9}$, and the probability that the rate 2 lightbulb dies first is $\frac{2}{9}$.

Proportional hazards model

This idea of competing exponentials is also useful in Markov chains where a state changes with different probabilities depending on the new state. Suppose someone has stage III cancer, and they can either die (with rate 1), or enter stage IV (rate 2), or their condition stays the same. What is the probability that their cancer develops to stage IV before they die?

Proportional hazards model

Maximum likelihood estimates for $\boldsymbol{\beta}$ can be found by numerical maximization, usually found by setting the partial derivatives with respect to β_k equal to 0. These are called score equations. The maximization can be done without knowing the baseline hazard $h_0(t)$.
Second derivatives are also useful for constructing the *Information matrix* $\mathbf{I}(\boldsymbol{\beta})$ for which the ijth element is

$$-\frac{\partial^2}{\partial\beta_i\partial\beta_j}\ell(\boldsymbol{\beta})$$

Based on large samples, the estimated coefficients, $\mathbf{b} = \widehat{\boldsymbol{\beta}}$ are approximately p-variate normal distributed with covariance matrix $\mathbf{I}^{-1}(\mathbf{b})$, meaning that we evaluate $\mathbf{I}(\boldsymbol{\beta})$ at $\boldsymbol{\beta} = \mathbf{b}$.

Proportional hazards model

Similar to regression models, a typical null hypothesis to test is that none of the risk factors change the hazard rate from the baseline (think of the baseline hazard as the overall mean). However, you can also test for a specific $\boldsymbol{\beta}$:

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$$

The Wald test is based on

$$(\mathbf{b} - \boldsymbol{\beta}_0)' \mathbf{I}(\mathbf{b})(\mathbf{b} - \boldsymbol{\beta}_0)$$

which has a χ^2 distribution with p degrees of freedom for large samples under the null hypothesis.

The likelihood ratio test is based on

$$X^2_{LR} = 2[\ell(\mathbf{b}) - \ell(\boldsymbol{\beta}_0)]$$

which is also approximately χ^2 with p degrees of freedom for large samples under the null hypothesis.

Proportional hazards model

A third test is the scores test with $\mathbf{U}(\boldsymbol{\beta}) = (U_1(\boldsymbol{\beta}), U_2(\boldsymbol{\beta}), \ldots, U_p(\boldsymbol{\beta}))'$ being the partial derivates with respect to β_k. Then

$$X^2_{SC} = \mathbf{U}(\boldsymbol{\beta}_0)'\mathbf{I}^{-1}(\boldsymbol{\beta}_0)\mathbf{U}(\boldsymbol{\beta}_0)$$

which is also approximately χ^2 with p degrees of freedom.

Proportional hazards model

In the case of one binary covariate and the null hypothesis that $\beta = \beta_1 = 0$, the score test simplifies considerably. Let

$$U(0) = \sum_{i=1}^{D} Z_{(i)} - \frac{Y_{1i}}{Y_{0i} + Y_{1i}}$$

where Y_{1i} is the number at risk in the $Z = 1$ risk group, Y_{0i} is the number at risk in the $Z = 0$ group.

$$I(0) = \sum_{i=1}^{D} \frac{Y_{0i} Y_{1i}}{(Y_{0i} + Y_{1i})^2}$$

Then

$$X^2_{SC} = U(0)/I(0)$$

Proportional hazards model with ties

There are different approaches for dealing with ties.

For the first approach, as usual, let $t_1 < \cdots < t_D$ be the distinct death times. Let d_i be the number of deaths at time t_i. Let $\mathbb{D}_i$ be the set of individuals who die at time t_i. Let

$$\mathbf{s} = \sum_{j \in \mathbb{D}_i} \mathbf{Z}_j$$

be the sum of the covariate vectors over all individuals who die at time t_i, and let R_i be the set of individuals at risk at time t_i (including those who die or are censored at time t_i).

Then the partial likelihood is defined as

$$L_1(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta' \mathbf{s}_i)}{\left[\sum_{j \in R_i} \exp(\beta' \mathbf{Z}_j)\right]^{d_i}}$$

Proportional hazards model with ties

Efron suggested an alternative partial likelihood as

$$L_2(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta' \mathbf{s}_i)}{\prod_{j=1}^{d_i} \left[\sum_{k \in R_i} \exp(\beta' \mathbf{Z}_k) - \frac{j-1}{d_i} \sum_{k \in \mathbb{D}_i} \exp(\beta' \mathbf{Z}_k) \right]}$$

which is similar to the previous one (particularly if the number of ties is small)

Proportional hazards model with ties

Cox suggested another approach based on a logistic model for the hazard rate. For this model, we assume

$$\frac{h(t|\mathbf{Z})}{1-h(t|\mathbf{Z})} = \frac{h_0(t)}{1-h_0(t)} \exp(\beta'\mathbf{Z})$$

Then let Q_i denote the set of all subsets of size d_i from the risk set R_i. Thus there are $\binom{R_i}{d_i}$ sets in Q_i. Then the partial likelihood is

$$L_3(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta'\mathbf{s}_i)}{\sum_{q\in Q_i} \exp(\beta'\mathbf{s}_q^*)}$$

where

$$s_q^* = \sum_{j=1}^{d_i} \mathbf{Z}_{qj}$$

If $\binom{R_i}{d_i}$ is large (which could happen if there is a large sample with a lot of ties), then $\mathbf{s}_q^*$ could be estimated by randomly sampling $q \in Q_i$

Proportional hazards model with ties

All three approximations to the partial likelihood reduce to the same partial likelihood when there are no ties.

Proportional hazards in R

To do proportional hazards in R, you can use `coxph()`, which depends on the `survival` library. In SAS, you can use PROC PHREG.

For dealing with the ties, the Efron method is the default. There is also an exact option which is more computationally intensive.

Example: recidivism data

We'll do an example with recidivism data, looking at the the amount of time until someone is arrested after being released from prison. Each person was followed for 52 weeks. Consequently, if the released person was not arrested within 52 weeks, then the observation is right-censored at 52 weeks. The variables in the study are:

Example: recidivism data

1. week: week of first arrest after release, or censoring time.
2. arrest: the event indicator, equal to 1 for those arrested during the period of the study and 0 for those who were not arrested.
3. fin: a dummy variable, equal to 1 if the individual received financial aid after release from prison, and 0 if he did not; financial aid was a randomly assigned factor manipulated by the researchers.
4. age: in years at the time of release.
5. race: a dummy variable coded 1 for blacks and 0 for others.
6. wexp: a dummy variable coded 1 if the individual had full-time work experience prior to incarceration and 0 if he did not.
7. mar: a dummy variable coded 1 if the individual was married at the time of release and 0 if he was not.
8. paro: a dummy variable coded 1 if the individual was released on parole and 0 if he was not.
9. prio: number of prior convictions.

Exmaple: recidivism data

10. educ: education, a categorical variable, with codes 2 (grade 6 or less), 3 (grades 6 through 9), 4 (grades 10 and 11), 5 (grade 12), or 6 (some post-secondary).

11. emp1 emp52: dummy variables coded 1 if the individual was employed in the corresponding week of the study and 0 otherwise.

Example: Recidivism data

week	arrest	fin	age	race	wexp	mar	paro	prio	educ	emp1	emp2	emp3	emp4	emp5	emp6	emp7	emp8	emp9	emp10	emp11
emp12	emp13	emp14	emp15	emp16	emp17	emp18	emp19	emp20	emp21	emp22	emp23	emp24	emp25	emp26	emp27	emp28	emp29	emp30	emp31	
emp32	emp33	emp34	emp35	emp36	emp37	emp38	emp39	emp40	emp41	emp42	emp43	emp44	emp45	emp46	emp47	emp48	emp49	emp50	emp51	
emp52																				
20	1	0	27	1	0	0	1	3	3	0	0	0	0	0	0	0	0	0	0	0
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA																				
17	1	0	18	1	0	0	1	8	4	0	0	0	0	0	0	0	0	1	1	1
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA																				
25	1	0	19	0	1	0	1	13	3	0	0	0	0	0	0	0	0	0	0	0
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA																				
52	0	1	23	1	1	1	1	1	5	0	0	0	0	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1																				
52	0	0	19	0	1	0	1	3	3	0	0	0	0	0	0	1	1	1	1	1
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0																				
52	0	0	24	1	1	0	0	2	4	0	0	0	0	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0																				
23	1	0	25	1	1	1	1	0	4	1	1	1	1	1	1	1	1	1	1	1
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA																				
52	0	1	21	1	1	0	1	4	3	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0																				
52	0	0	22	1	0	0	0	6	3	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1																				
52	0	0	20	1	1	0	0	0	5	0	1	1	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0																				
52	0	1	26	1	0	0	1	3	3	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
1																				
52	0	0	40	1	1	0	0	2	5	0	0	0	0	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1																				

Example: Recidivism data

To read in the data, make sure the dataset `Rossi.txt` is in your current working directory or use the url:

```
> data <- read.table("Rossi.txt'',header=T) #or the following
> data <- read.table("http://math.unm.edu/~james/Rossi.txt",he
> data[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
```

Note that some variables are constant over time, and some are not, in particular whether the person was working in each of the 52 weeks.

Example: Recidivism data

To fit the model, the syntax is similar as was used for Kaplan-Meier. Here we don't use time-dependent covariates:

```
> library(survival)
> attach(data)
> mod1 <- coxph(Surv(week, arrest)~fin+age+race+wexp+mar+paro+prio)
> mod1
coxph(formula = Surv(week, arrest) ~ fin + age + race + wexp +
    mar + paro + prio)

        coef exp(coef) se(coef)      z      p
fin  -0.3794     0.684   0.1914 -1.983 0.0470
age  -0.0574     0.944   0.0220 -2.611 0.0090
race  0.3139     1.369   0.3080  1.019 0.3100
wexp -0.1498     0.861   0.2122 -0.706 0.4800
mar  -0.4337     0.648   0.3819 -1.136 0.2600
paro -0.0849     0.919   0.1958 -0.434 0.6600
prio  0.0915     1.096   0.0286  3.194 0.0014
Likelihood ratio test=33.3  on 7 df, p=2.36e-05  n= 432,
number of events= 114
```

Example: recidivism data

The results suggest that for this model, age and number of prior arrests were the most significant risk factors, and there is some evidence that receiving financial aid also had an effect. A negative coefficient means that the hazard rate is lower for that variable, so receiving financial aid lowered the chance of recidvism in this data.

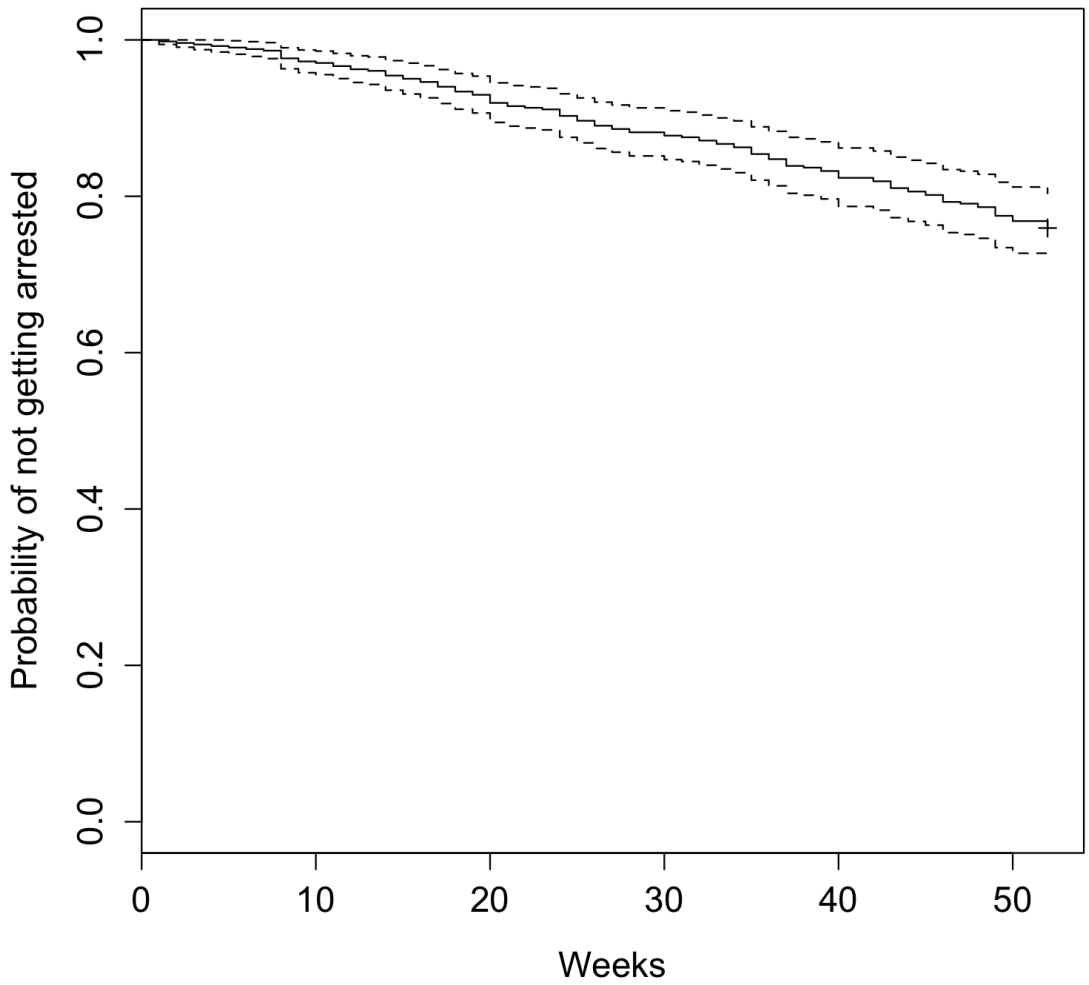
Example: recidivism data

To plot the estimated survival curve, you can do the following (assuming you've done `attach(data)`

```
> plot(survfit(mod1), xlab="Weeks", ylab="Probability of not getting arrested",cex.axis=1.3,cex.lab=1.3)
```

The plot is based on the estimated survival using the average covariate (for quantitative covariates) or baseline covariates (for qualitative) covariates. If you want to have more resolution for the curve, since the probabilities are all high, you could use an option in the plot statement such as `ylim=c(.5,1)`.

Example: Recidivism data



Example: Recidivism data

The main interest in this data was to see if financial aid after release from prison would improve recidivism rates, so it would make sense to plot the two survival curves separately for ex-prisoners receiving and not receiving financial aid.

```
> Rossi.fin <- data.frame(fin=c(0,1), age=rep(mean(age),2),
race=rep(mean(race),2),wexp=rep(mean(wexp),2),
mar=rep(mean(mar),2), paro=rep(mean(paro),2),
prio=rep(mean(prio),2))
> plot(survfit(mod1, newdata=Rossi.fin), conf.int=F,
lty=c(1,2), ylim=c(.6, 1))
```

Note that `lty=2` is given for the second group, which is `fin=1`, meaning the group that received financial aid. Thus, the dotted line corresponds to the group that received financial aid, and this survival curve appears to be higher (meaning a lower probability of getting arrested) than the no financial aid group. The p-value is close to 0.05, so there is some but not overwhelming evidence for the effectiveness of the financial aid.

Example: recidivism data

In practical terms, it might make more sense to think about how many fewer arrests would be predicted based on financial aid. The probabilities are obtained using

```
> summary(survfit(mod1,newdata=Rossi.fin))
...
 50     325         3     0.727      0.804
 52     322         4     0.717      0.796
```

Thus, there is roughly an 80% versus 72% estimated probability of no arrests. An 8% difference for a group of 432 individuals is about 35 individuals. Even if the effect seems marginally significant in a statistical sense, there is a pretty large practical significance to the difference, if it is real.

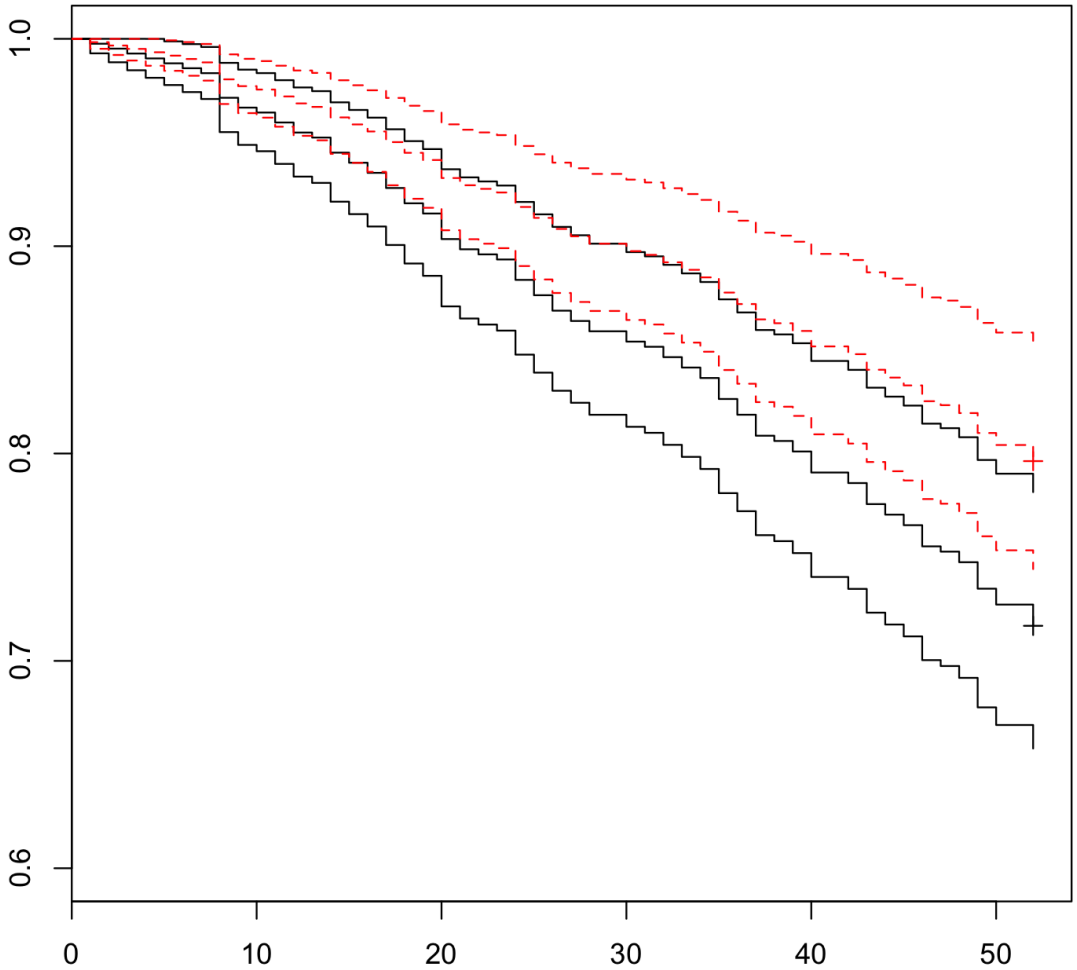
Example: recidivism data

To get confidence intervals and distinguish using color, use

```
 plot(survfit(mod1, newdata=Rossi.fin), conf.int=T,
lty=c(1,2), col=c("black","red"),ylim=c(.6, 1),
xlab="Weeks",ylab="Probability of no arrests",
cex.axis=1.3,cex.lab=1.3)
```

(I tried adding a legend, but encountered errors and didn't figure out the problem....)

Example: Recidivism data



Example: recidivism data

Note that the confidence intervals overlap. Overlapping confidence intervals is frequently used as an informal check for testing for a significant difference. If confidence intervals don't overlap, then it is usually safe to assume that the two groups are different (for other problems as well), but overlapping confidence intervals don't necessarily mean that there isn't a singificant difference between two groups. In this case, because the p-value is so close to 0.05, it isn't very surprising that there is substantial overlap in the confidence intervals.

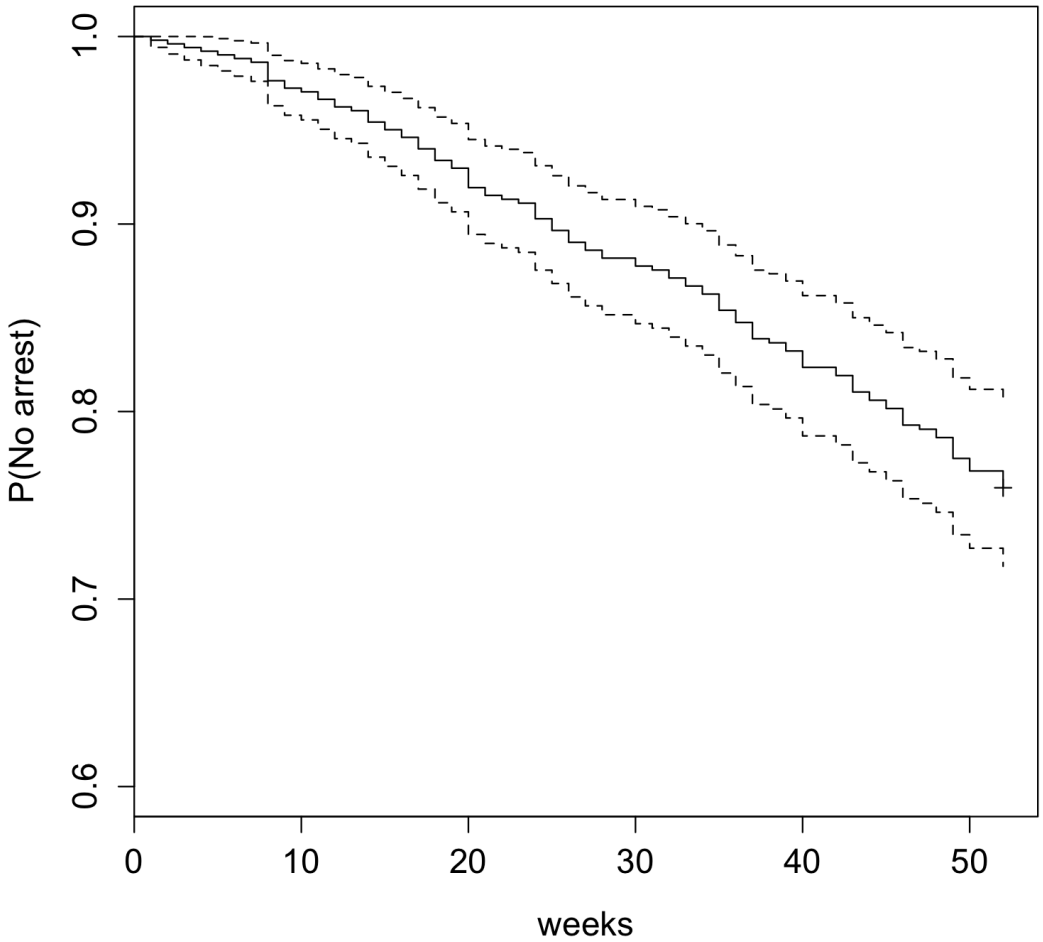
Example: recidivism data

For a comparison with the Kaplan-Meier estimate (not adjusting for covariates), we can do the following:

```
> library(survival)
> attach(data)
> mod2 <- survfit(Surv(week,arrest)~1)
> plot(survfit(mod1),cex.axis=1.3,cex.lab=1.3,ylim=c(.6,1))
> points(mod2$time,mod2$surv,type="s",col="red")
> legend(0,.8,legend=c("Cox","K-M"),col=c("black","red"),
lty=c(1,1),cex=1.5)
```

The Kaplan-Meier curve estimates slightly lower survival (i.e., higher arrest rates) compared to the covariate-adjusted average arrests, but overall, the curves are very similar.

Example: recidivism data



Getting the survival function from the hazard rate coefficients

The survival function is determined automatically in the software, but to get an idea of how this is estimated, we use the usual relationship that survival function is obtained by exponentiating the cumulative hazard:

$$\widehat{S}(t) = \exp[-\widehat{H}(t)]$$

where

$$\widehat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{W(t_i; \mathbf{b})}$$

and

$$W(t_i; \mathbf{b}) = \sum_{j \in R_i} \exp\left(\sum_{h=1}^{p} b_h Z_{jh}\right)$$

Getting the estimated survival for a new individual

If we want a new individual's estimated survival probability, using a set of covariates $\mathbf{Z}_0$ not necessarily in the data, then

$$\widehat{S}(t|\mathbf{Z} = \mathbf{Z}_0) = [\widehat{S}_0(t)]^{\exp(\mathbf{b}'\mathbf{Z}_0)}$$

where $\widehat{S}_0(t)$ is the estimated survival function assuming Z_0 This is useful for getting estimated survival rates for individuals who are not at just the average values of the covariates.

Getting the estimated survival for a new individual

As an example, for simplicity, I'll use a model with only `fin` (financial aid status), age and `prio` (number of prior convictions) as predictors.

```
> attach(data)
> mod1 <- coxph(Surv(week, arrest)~fin+age+prio)
> summary(mod1)
Call:
coxph(formula = Surv(week, arrest) ~ fin + age + prio)

  n= 432, number of events= 114

         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 ***
```

Model selection

Model selection, deciding which variables to include in the model, can be done in different ways, but is similar to regression and ANOVA procedures. You could, for example, use a forward selection to add variables one at a time, adding the most significant variable at each step (based on p-values, for example). Some variables might be desired to be included in the model even if they are not signficant. In this problem, the main question of interest is the role of the `fin` variable, so it would make sense to include it even if it doesn't pass a significance test.

Model selection

Another approach is to use the Aikake Information Criterion (AIC). This is typically done using

$$AIC = -2 \log L + 2p$$

where p is the number of parameters. A model with a lower AIC is preferred over one with higher AIC. For this example, to see whether a model with financial aid is better than one without, when `prio` and `age` are in the model, you can use

```
> mod1 <- coxph(Surv(week,arrest) ~ fin + age + prio)
> mod2 <- coxph(Surv(week,arrest) ~ age + prio)
> -2*mod1$loglik[2]+2*3
[1] 1327.714
> -2*mod2$loglik[2]+2*2
[1] 1329.085
```

Based on AIC, the model with financial aid is preferred using AIC even though financial aid did not have p-value $< .05$. Some statisticians and/or journals might prefer using AIC over p-values (or vice versa). Of course there is the danger of trying both and when they disagree, just picking the model that you prefer.

Getting the estimated survival for a new individual

Note that in this model, the financial aid model would not pass an $\alpha = .05$ test of significance. Although this variable is significant in one model and not in the other based on this criterion, the p-values of 0.047 and 0.068 are really not very different from each other. Both suggest moderate evidence against the hypothesis that financial aid had no effect.

Getting the estimated survival for a new individual

To get the estimated survival probabilities for a new individual, we can use the approach used to plot two curves used earlier to compare financial aid to no financial aid. Suppose we want to predict the survival curve for an individual who is 20 years old with 1 prior conviction and no financial aid. We can compare these values to the averages for the data using

```
> colMeans(data)
      week     arrest        fin        age       race          w
45.8541667  0.2638889  0.5000000 24.5972222  0.8773148  0.5717
      paro       prio       educ       emp1       emp2          e
 0.6180556  2.9837963  3.4768519  0.1388889         NA
```

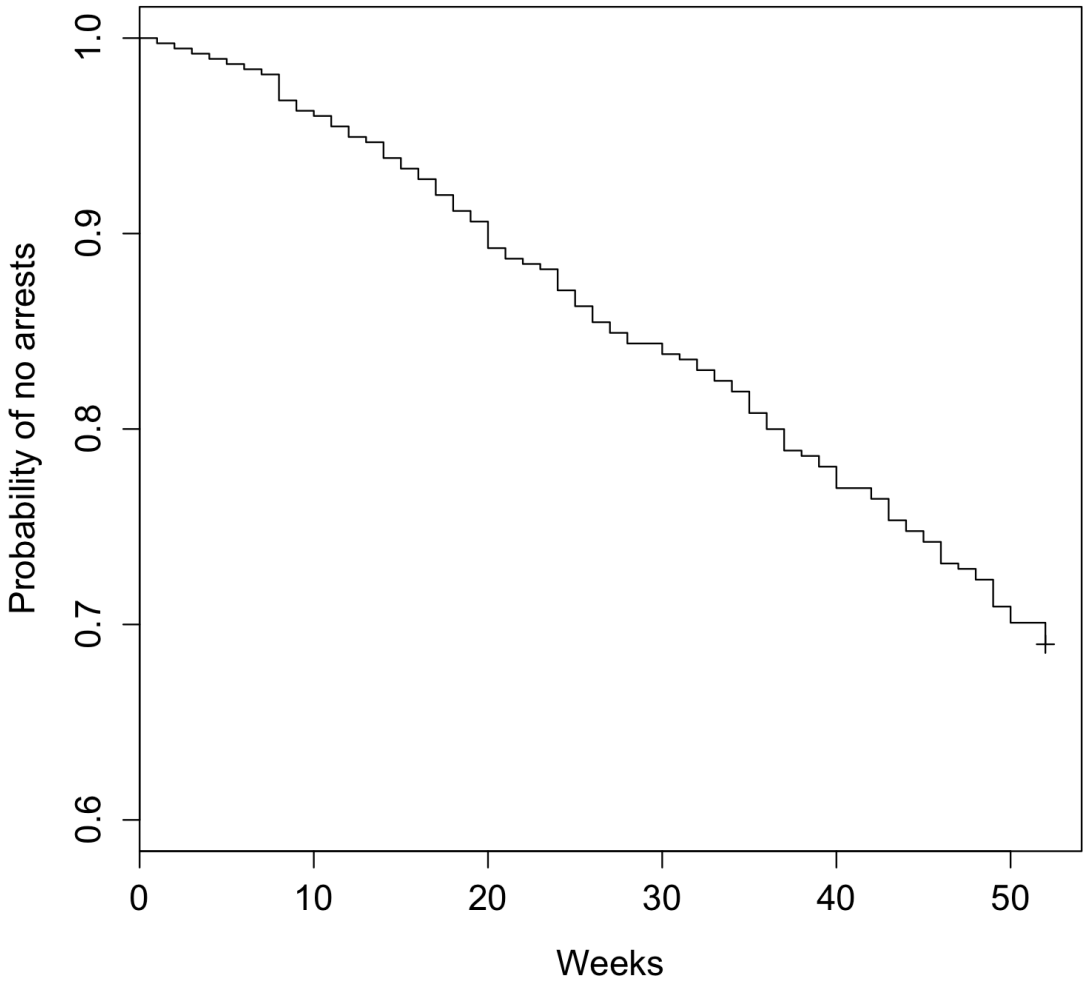
The averages show that 26% of the individuals were arrested within a year, that 50% received financial aid, and that the average age was 24.6 years. The average for `week` isn't very meaningful because this is a mixture of censored and non-censored times.

Getting the estimated survival for a new individual

To get the estimated survival curve for the individual aged 20 with 1 prior conviction and no financial aid, we use

```
> library(survival)
Loading required package: splines
> attach(data)
> mod1 <- coxph(Surv(week, arrest)~fin+age+prio)
> data1 <- data.frame(fin=0,age=20,prio=1)
> plot(survfit(mod1,newdata=data1),cex.axis=1.3,
cex.lab=1.3,conf.int=F,ylim=c(.6,1),ylab="Probability
of no arrests",xlab="Weeks")
```

Estimated survival for a new individual

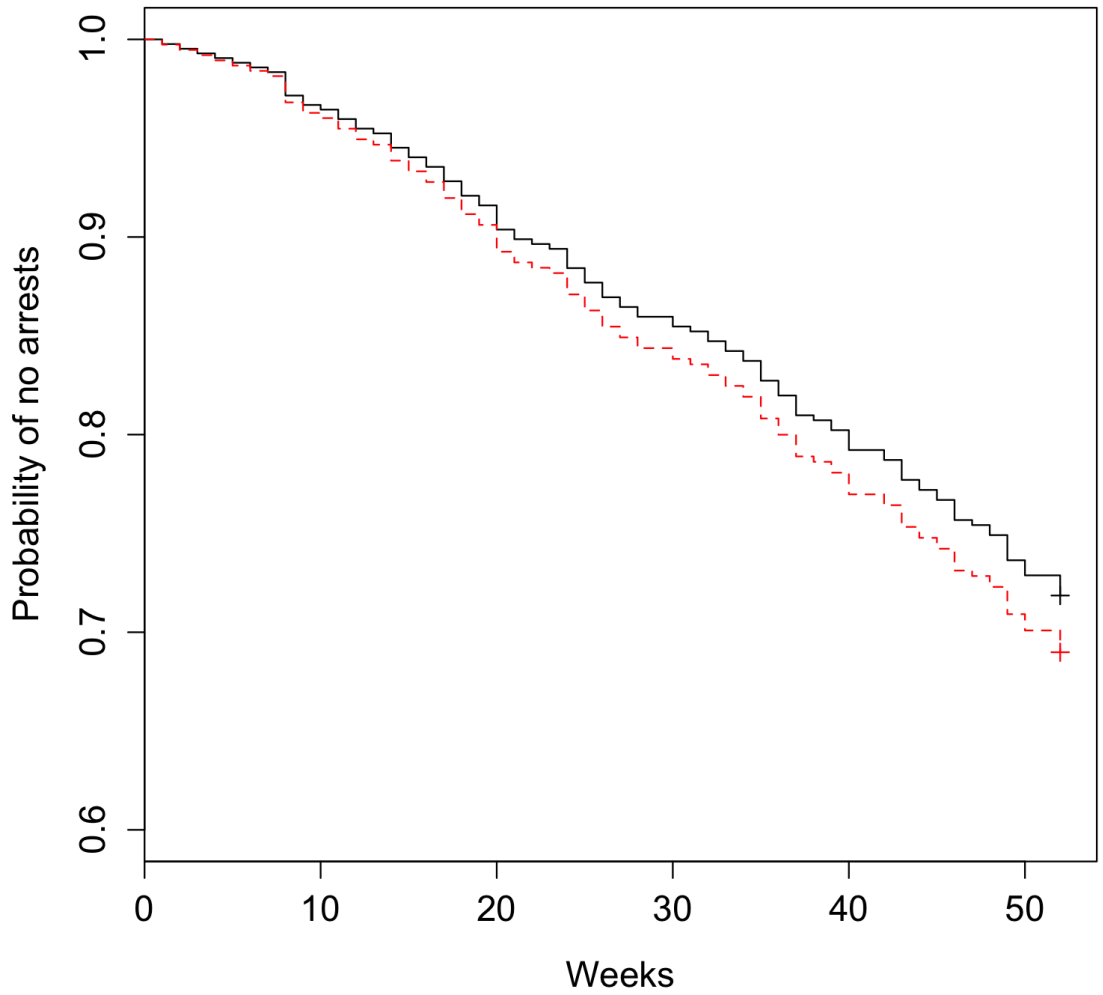


Estimated survival for a new individual

Instead of having the curve by itself, we might want to compare to an average survival curve. Here, I compare to the survival curve for the average age (close to 25), average number of prior convictions (close to 3), but no financial aid (so,not averaging over financial aid status). Again, other variables such as race, marital status, education, etc. were ignored, which is different from averaging over these values.

```
mod1 <- coxph(Surv(week, arrest)~fin+age+prio)
data1 <- data.frame(fin=c(0,0),age=c(mean(age),20),
prio=c(mean(prio),1))
plot(survfit(mod1,newdata=data1),cex.axis=1.3,
cex.lab=1.3,conf.int=F,ylim=c(.6,1),ylab="Probability
of no arrests",xlab="Weeks")
```

Estimated survival for a new individual (red, dotted curve)



Estimated survival for anew individual

So the curve suggests that the estimated probability for getting arrested within a year is slightly higher for a 20-year old with 1 prior conviction than for a 25-year old with 3 prior convictions gets arrested, although the differences are small.
Getting a confidence interval for the survival curve for the individual is a bit difficult, but can be done using Q_3 on the next slide as the variance estimate and using techniques from section 4.3 to generate the pointwise confidence intervals.

Estimated survival for a new individual (red, dotted curve)

$$\hat{V}[\hat{S}(t \mid \mathbf{Z} = \mathbf{Z}_0)] = [\hat{S}(t \mid \mathbf{Z} = \mathbf{Z}_0)]^2[Q_1(t) + Q_2(t; \mathbf{Z}_0)]. \tag{8.8.5}$$

Here,

$$Q_1(t) = \sum_{t_i \le t} \frac{d_i}{W(t_i, \mathbf{b})^2} \tag{8.8.6}$$

is an estimator of the variance of $\hat{H}_0(t)$ if $\mathbf{b}$ were the true value of $\boldsymbol{\beta}$. Here

$$Q_2(t; \mathbf{Z}_0) = \mathbf{Q}_3(t; \mathbf{Z}_0)^t \hat{V}(\mathbf{b}) \mathbf{Q}_3(t; \mathbf{Z}_0) \tag{8.8.7}$$

with $\mathbf{Q}_3$ the p-vector whose kth element is defined by

$$Q_3(t, \mathbf{Z}_0)_k = \sum_{t_i \le t} \left[\frac{W^{(k)}(t_i; \mathbf{b})}{W(t_i; \mathbf{b})} - Z_{0k} \right] \left[\frac{d_i}{W(t_i, \mathbf{b})} \right], \quad k = 1, \ldots, p \tag{8.8.8}$$

where

$$W^{(k)}(t_i; \mathbf{b}) = \sum_{j \in R(t_i)} Z_{jk} \exp(\mathbf{b}^t \mathbf{Z}_j)$$

Adding time-dependent covariates

A next step in the modeling process is to add time-dependent covariates. For the recidivism data, the only time dependent covariate is employment. One could imagine that marital status could have changed while in prison, so this could have been kept track of as well as a time-dependent covariate, but wasn't for this data.

To analyze the data in R, we want the data structured differently, so that there is a row for each non-missing week for each individual (i.e., weeks that are not missing for the employment variable). We want to create a single variable for employment instead of 52 such variables, but to do this, we need separate rows for the different weeks in which someone might or might not have been employed. For this data, we also want to distinguish between someone who is not employed but also not arrested versus someone is not employed and has been arrested, which is treated as an NA. This restructuring is a bit of a pain, and honestly could probably be more easily done in SAS using PROC TRANSPOSE (if you are good at SAS).

Adding time-dependent covariates

	start	stop	arrest.time	week	arrest	fin	age	race	wexp	mar	paro	prio	educ	employed
1.1	0	1	0	20	1	0	27	1	0	0	1	3	3	0
1.2	1	2	0	20	1	0	27	1	0	0	1	3	3	0
. . .														
1.19	18	19	0	20	1	0	27	1	0	0	1	3	3	0
1.20	19	20	1	20	1	0	27	1	0	0	1	3	3	0
2.1	0	1	0	17	1	0	18	1	0	0	1	8	4	0
2.2	1	2	0	17	1	0	18	1	0	0	1	8	4	0
. . .														
2.16	15	16	0	17	1	0	18	1	0	0	1	8	4	0
2.17	16	17	1	17	1	0	18	1	0	0	1	8	4	0
3.1	0	1	0	25	1	0	19	0	1	0	1	13	3	0
3.2	1	2	0	25	1	0	19	0	1	0	1	13	3	0
. . .														
3.13	12	13	0	25	1	0	19	0	1	0	1	13	3	0

Adding time-dependent covariates

```
> sum(!is.na(Rossi[,11:62])) # record count
[1] 19809
> Rossi.2 <- matrix(0, 19809, 14) # to hold new data set
> colnames(Rossi.2) <- c(start, stop, arrest.time, names(Rossi)[1:10
> row <- 0 # set record counter to 0
> for (i in 1:nrow(Rossi)) { # loop over individuals
+ for (j in 11:62) { # loop over 52 weeks
+ if (is.na(Rossi[i, j])) next # skip missing data
+ else {
+ row <- row + 1 # increment row counter
+ start <- j - 11 # start time (previous week)
+ stop <- start + 1 # stop time (current week)
+ arrest.time <- if (stop == Rossi[i, 1] && Rossi[i, 2] ==1) 1 else
+ # construct record:
+ Rossi.2[row,] <- c(start, stop, arrest.time, unlist(Rossi[i, c(1:1
+ }
+ }
+ }
```

Adding time-dependent covariates

```
> Rossi.2 <- as.data.frame(Rossi.2)
> remove(i, j, row, start, stop, arrest.time) # clean up
>
```

Adding time-dependent covariates

```
> mod2 <- coxph(Surv(start, stop, arrest.time) ~
+ fin + age + race + wexp + mar + paro + prio + employed,
+ data=Rossi.2)
> summary(mod2)
Call:
coxph(formula = Surv(start, stop, arrest.time) ~ fin + age +
race + wexp + mar + paro + prio + employed, data = Rossi.2)
n= 19809
coef exp(coef) se(coef) z p
fin -0.3567 0.700 0.1911 -1.866 6.2e-02
age -0.0463 0.955 0.0217 -2.132 3.3e-02
race 0.3387 1.403 0.3096 1.094 2.7e-01
wexp -0.0256 0.975 0.2114 -0.121 9.0e-01
mar -0.2937 0.745 0.3830 -0.767 4.4e-01
paro -0.0642 0.938 0.1947 -0.330 7.4e-01
prio 0.0851 1.089 0.0290 2.940 3.3e-03
employed -1.3282 0.265 0.2507 -5.298 1.2e-07
```

Adding time-dependent covariates

The results suggests that employment was highly significant for the probability of getting arrested, and including this in the model made other variables have weaker p-values. For example, with employment in the model, the p-value for `fin` becomes .062, and the p-values for `age` and `prio` are .033 and .0033, respectively.

Checking Proportional Hazards assumptions

An important assumption of the proportional hazards model is that the log of the hazard is linear in the covariates. This assumption can be tested using the `cox.zph()` function, which tests for linearity for each covariate separately.

```
> library(survival)
Loading required package: splines
> cox.zph(mod1)
Error in Surv(week, arrest) : object 'week' not found
> attach(data)
> cox.zph(mod1)
            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
> par(mfrow=c(2,2))
```

Checking proportional hazards assumptions

The global test (which helps correct for multiple testing issues) is not quite significant using $\alpha = .05$. If you look at individual variables, there is somewhat strong evidence that the age variable is violating the proportional hazards assumption. I.e., the hazard rate for age doesn't appear to be constant over time.

Plotting the output of `cox.zph()` gives diagnostic plots.

Adding time-dependent covariates