Chapter 11: Survival Analysis and Censored Data

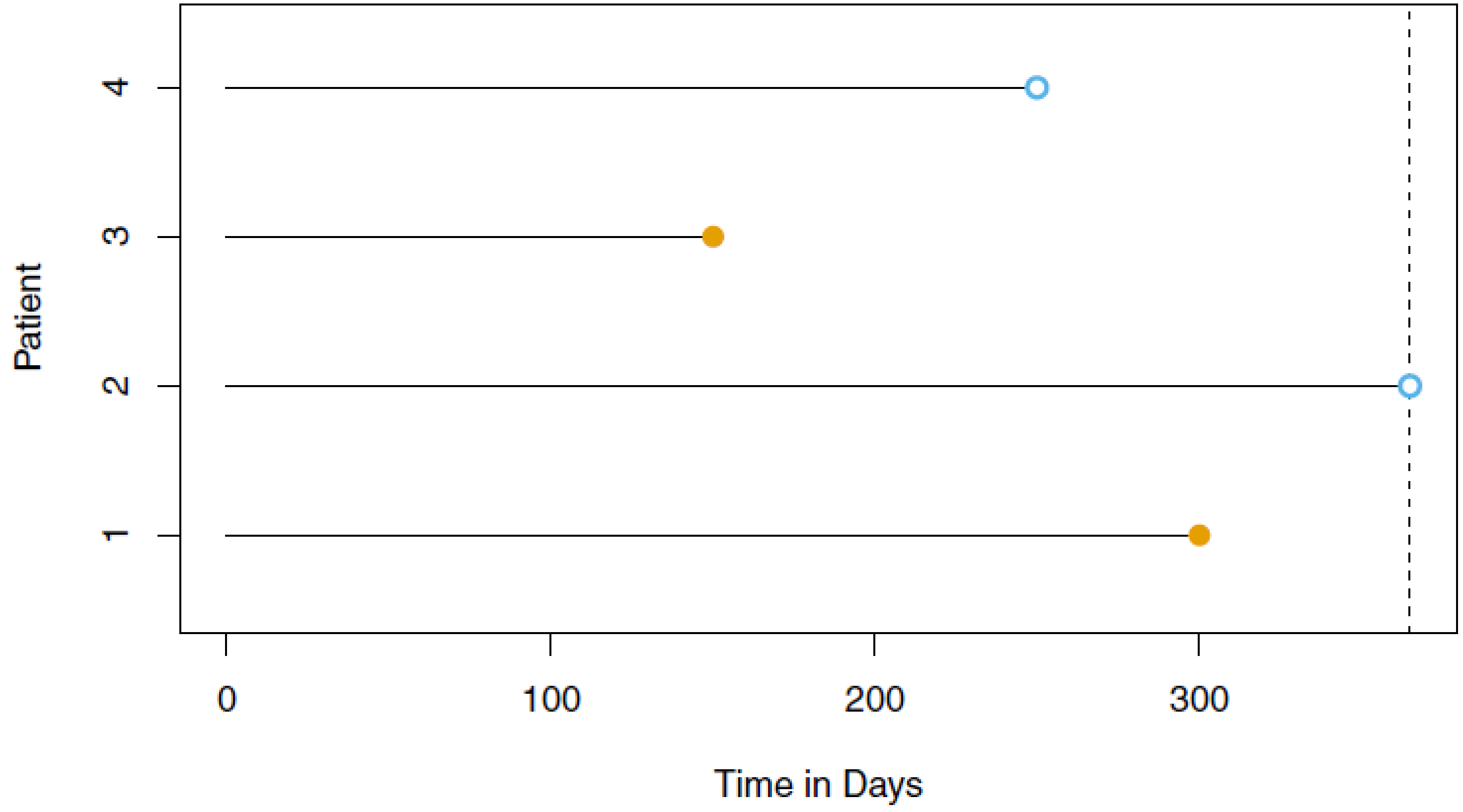
- Samples are followed until an event (death) happens or until the sample is censored.
- An event may be censored by removing it from the study or if the study terminated before the event happens.
- The survival time, T, is the time to the event while C is the time to a censoring event.
- For each individual we then have the random variable, $Y=\min(T,C)$.

Censored data

- In addition, we have an indicator random variable associated with the times, T and C,

$$\delta = \begin{cases} 1 \; if \; T \leq C \\ 0 \; if \; T > C \end{cases}$$

- In this figure $y_1=t_1$, $y_2=c_2$, $y_3=t_3$ and $y_4=c_4$
- $\delta_1=\delta_3=1$ and $\delta_2=\delta_4=0$



The Kaplan-Meier Survival Curve

- The survival function is the probability of surviving past a specific time, $S(t)=\Pr(T>t)$.
- We can't simply count up all the individuals that survived longer than t and divide by the total, since some individuals will have been censored before t.
- If we simply ignore all the individuals who were censored before t, then we are throwing out useful information.

The Kaplan-Meier Survival Curve

- Let $d_1 < d_2 < \cdots < d_K$ be the K unique times of death among the uncensored individuals.
- Let q_k be the total number of deaths at time d_k.
- Finally, let r_k be the total number alive just before time d_k. These "at risk" individuals can include individuals who will ultimately be censored.
- $\Pr(T > d_k) = \Pr(T > d_k | T > d_{k-1}) \Pr(T > d_{k-1}) + \Pr(T > d_k | T \leq d_{k-1}) \Pr(T \leq d_{k-1})$ [condition on all possible values of T]
- But $\Pr(T > d_k | T \leq d_{k-1})$ must be 0 since $d_{k-1} < d_k$
- $S(d_k) = \Pr(T > d_k) = \Pr(T > d_k | T > d_{k-1}) \Pr(T > d_{k-1}) = \Pr(T > d_k | T > d_{k-1}) S(d_{k-1})$ [continue by substituting for $S(d_{k-1})$ and so on]

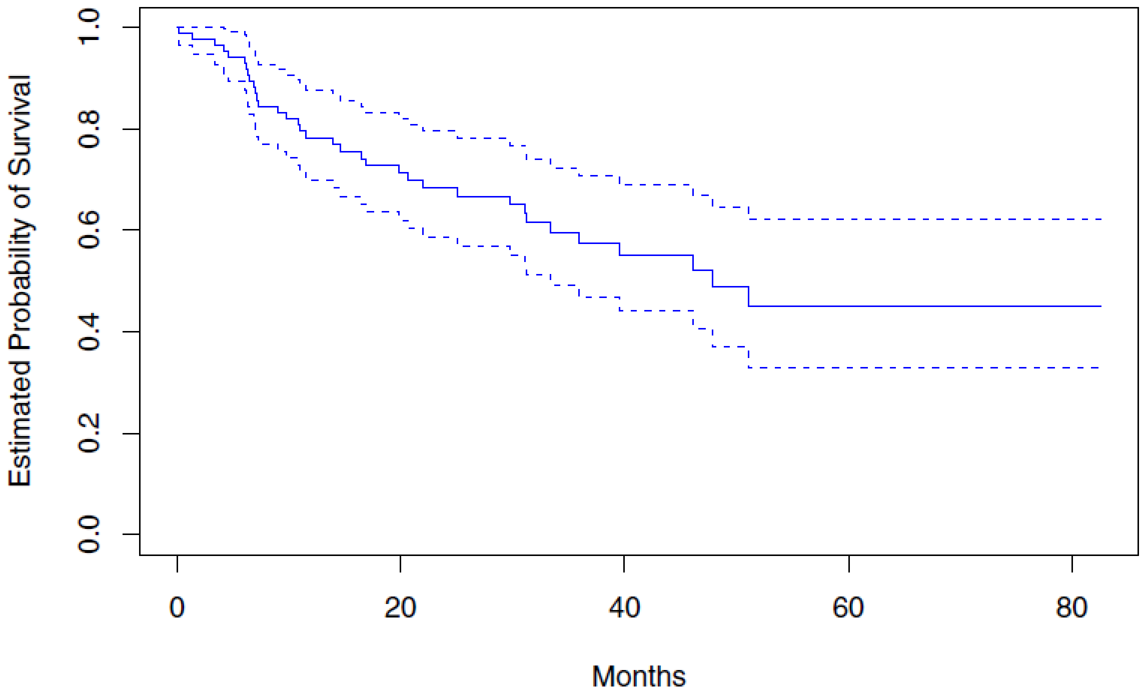
The Kaplan-Meier Survival Curve

- To estimate the conditional survival probability, we calculate the number of survivors among the at risk group, $r_j - q_j$, and the divide by the number of individuals at risk r_j,

$$\widehat{Pr}(T > d_j | T > d_{j-1}) = \frac{(r_j - q_j)}{r_j}$$

- Plugging this into the full equation,

$$\hat{S}(d_k) = \prod_{j=1}^{k} \left(\frac{r_j - q_j}{r_j} \right)$$



Log-Rank Test

- If we need to compare two survival curves to each other the log-rank test overcomes the difficulties of censored data. It will be a test over the entire length of the survival curve.
- The previous parameters are expanded to include subscripts "1" for group 1 and "2" for group 2.
- The null hypothesis will have the general construction, $W = \frac{X-E(X)}{\sqrt{Var(X)}}$
- For this problem $X = \sum_{k=1}^{K} q_{1k}$

	Group 1	Group 2	Total
Died	q_{1k}	q_{2k}	q_k
Survived	r_{1k} -q_{1k}	r_{2k} -q_{2k}	r_k -q_k
Total	r_{1k}	r_{2k}	r_k

Log-Rank Test

- Under the null hypothesis, $E(q_{1k}) = \frac{r_{1k}}{r_k} q_k$ since $q_{1k} = q_{2k} = q_k$
- Short aside: see problem 7. Consider an urn with r_k balls q_k/r_k are white. If we sample *without* replacement r_{1k} balls the probability of getting q_{1k} white balls in the sample follows a hypergeometric distribution which has a mean, $r_{1k}(q_k/r_k)$ and variance, $r_{1k} \frac{q_k}{r_k}\left(1 - \frac{q_k}{r_k}\right)\left(\frac{r_k - r_{1k}}{r_k - 1}\right) = r_{1k}\frac{q_k}{r_k}(1 - r_{1k}/r_k)(r_k - q_k) = Var(q_{1k})$
- Although various q_{ik} may be correlated the log-rank test uses the approximation, $Var\left(\sum_{k=1}^{K} q_{1k}\right) \approx \sum_{k=1}^{K} Var(q_{1k})$
- If the sample is large *p*-values can be derived assuming *W* has a normal distribution, otherwise a permutation test can used by randomly switching the labels “1” and “2”.

Hazard Function

- Can we develop a regression equation that can be used to predict the true survival time from the censored and uncensored survival data?
- The hazard function is set up to predict the probability of the event T in a small interval, $h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t < T \leq t + \Delta t | T > t)}{\Delta t}$
- As $\Delta t \to 0$, $h(t)$ is no longer a probability but is a conditional probability density function.

Hazard Function

- Recall that Pr(A|B)= Pr(A and B)/Pr(B)
- $h(t) = \lim_{\Delta t \to 0} \frac{\Pr\left(t<T\leq t+\Delta t \; and \; (T>t)\right)/\Delta t}{\Pr(T>t)} = \lim_{\Delta t \to 0} \frac{\Pr(t<T\leq t+\Delta t)/\Delta t}{\Pr(T>t)} = \frac{f(t)}{S(t)}$
- The function $f(t)$ is a probability density function or the instantaneous rate of death.
- We can use $f(t)$ and $S(t)$ to estimate likelihood, L_i, of sample observations.
- Thus, $L_i = \begin{cases} f(y_i) if \; the \; ith \; observation \; of \; not \; censored \\ S(y_i) \; if \; the \; ith \; observation \; is \; censored \end{cases}$
- By above, $L_i = f(y_i)^{\delta_i} S(y_i \;)^{1-\delta_i}$, for the entire sample, i=1,…,n the likelihood is $L = \prod_{i=1}^{n} f(y_i)^{\delta_i} S(y_i \;)^{1-\delta_i} = \prod_{i=1}^{n} h(y_i)^{\delta_i} S(y_i)$

Hazard Function

- What function might be used for $f(x)$? It could be the exponential function, $\lambda\exp(-\lambda t)$, or we could use covariates directly with the hazard function, $h(t|x_i) = exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)$.
- A popular function used to model aging in many organisms including humans is the Gompertz equation.
- The instantaneous mortality rate is, $\mu(t) = \frac{-1}{N}\frac{dN}{dt} = Aexp(\alpha t)$ (*)
- To derive how many survivors there will be at some future time, T, we can do the following to equation (*) above

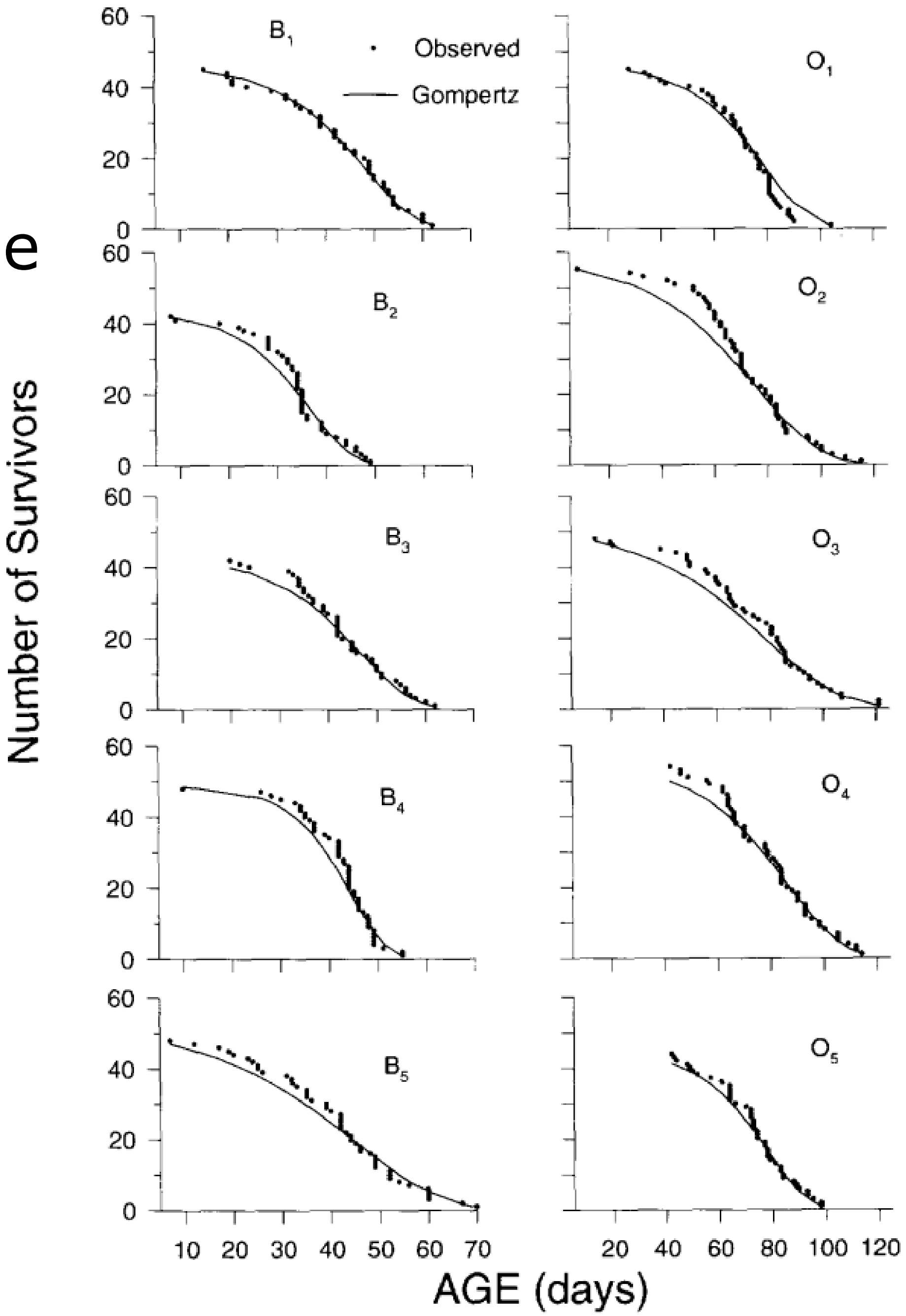
$$\int_{N_0}^{N_T} \frac{dN}{N} = -\int_0^T Aexp(\alpha t)dt \ \rightarrow \log(N)\,|_{N_0}^{N_T} = -\frac{A}{\alpha}\exp(\alpha t)\,|_0^T$$

$$log\left(\frac{N_T}{N_0}\right) = \frac{A}{\alpha}[1 - \exp(\alpha T)] \rightarrow N_T = N_0 exp\left(\frac{A(1 - \exp(\alpha T))}{\alpha}\right)$$

$$= N_0 Prob(surviving\ to\ T)$$

Gompertz Equation

- The distribution function of a random variable, $F(T)$, is the Prob($t<T$). For the Gompertz model that is $1 - exp\left(\frac{A(1-\exp(\alpha T))}{\alpha}\right)$
- Fact: The distribution function is related to the probability density function, $f(t)$, by $\frac{\partial F(t)}{\partial t} = f(t)$.
- Take some derivatives and do some algebra to get the density function for the Gompertz, $Aexp\left\{\frac{A(1-\exp(\alpha t))}{\alpha} + \alpha t\right\}$
- Female data on the right from Mueller et al., 1995, small sample size ~50 individuals per sex.



Gompertz Estimation

- Suppose we want to estimate A and α of the Gompertz from mortality that is observed over fixed time intervals, t_1, t_2,…,t_d. Suppose there are N_{tj} individuals alive at time t_j and d_{tj} deaths between then and time t_{j+1}. The empirical estimate of mortality is then, d_{tj}/ N_{tj}.
- This leads to the naïve estimate of instantaneous mortality as,

$$\mu(t_j) = \frac{d_{t_j}}{N_{t_j}} = Aexp(\alpha t_j)$$

- But the correct answer is $\frac{d_{t_j}}{N_{t_j}} = \int_{t_j}^{t_{j+1}} f(t)dt$. Another way to express value of the integral is, $F(t_{j+1}) - F(t_j)$.
- The naïve estimate produces biased estimates which get larger as the time interval gets larger. See Mueller et al., 1995. Exp. Geron. 30: 553-569.

Simulate Gompertz random variables

- With the Gompertz distribution function you can generate ages-at-death that follow the Gompertz equation using the inverse transform method.
- $\mathrm{F}(T) = \mathrm{U} = 1 - exp\left(\frac{A(1-\exp(\alpha T))}{\alpha}\right)$ Now solve this equation for T.
- After some algebra you get $T = \frac{ln\left[1-\frac{\alpha \ln(1-U)}{A}\right]}{\alpha}$
- Use a uniform (on (0,1)) random number generator to get U and then solve.

Proportional Hazards

- The likelihood function could be used to estimate the β parameters of covariates but would require a functional form for $f(t)$.
- Proportional hazards are more flexible.
- The model used is: $h(t|x_i) = h_0(t)exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)$, where $h_0(t)$ is the baseline hazard function which would apply if all x_i's were 0.
- The function $exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)$ is referred to as the relative risk since it reflects the changes to the hazard risk when the x_i's are not 0.

Proportional Hazards

- The only assumption that is implicit in the proportional hazards model is that a one unit increase in x_{ij} results in an increase in $h(t|x_i)$ by a factor $\exp(\beta_j)$.
- For the binary feature used in this figure the top two curves satisfy the proportional hazards model but the bottom two do not.

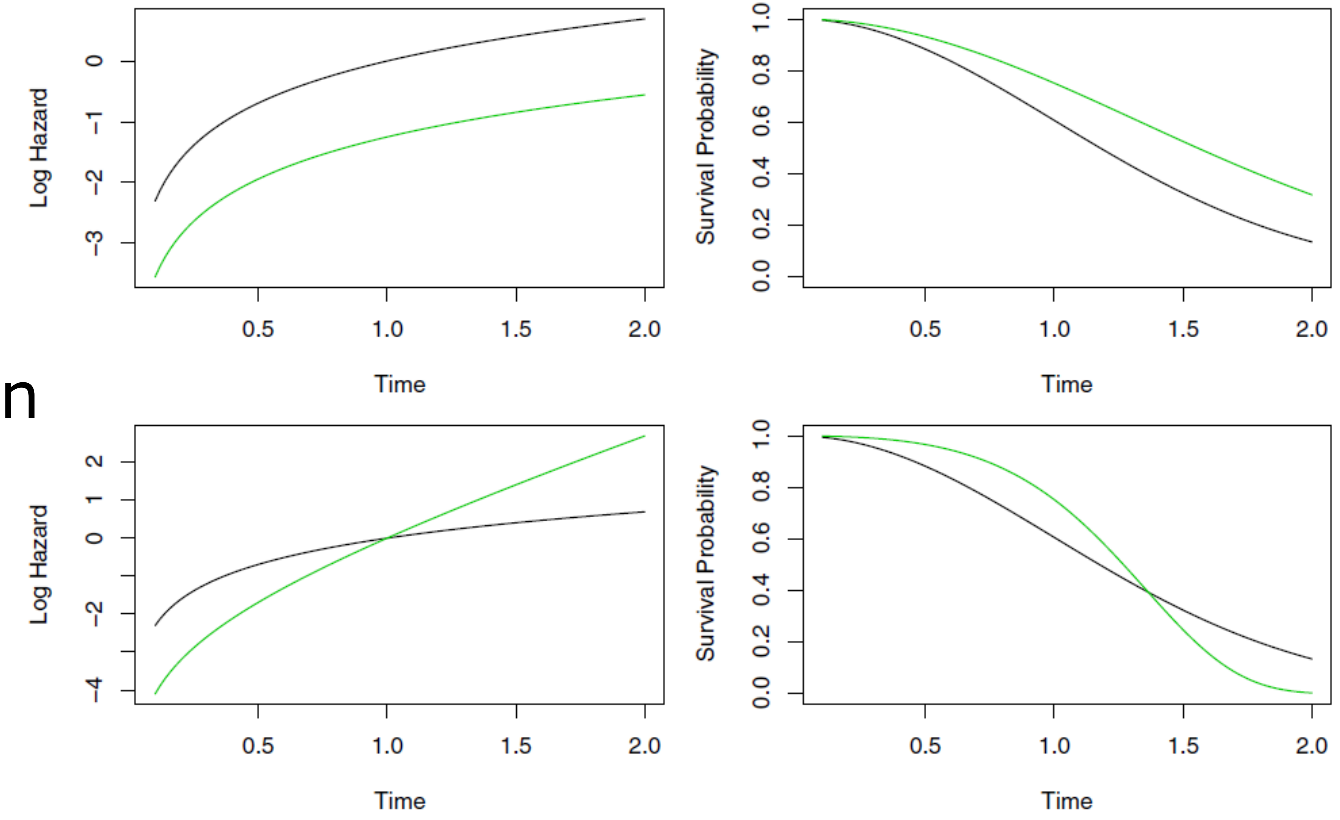


FIGURE 11.4. Top: *In a simple example with* $p = 1$ *and a binary covariate* $x_i \in \{0, 1\}$, *the log hazard and the survival function under the model* (11.14) *are shown (green for* $x_i = 0$ *and black for* $x_i = 1$*). Because of the proportional hazards assumption* (11.14), *the log hazard functions differ by a constant, and the survival functions do not cross.* Bottom: *Again we have a single binary covariate* $x_i \in \{0, 1\}$. *However, the proportional hazards assumption* (11.14) *does not hold. The log hazard functions cross, as do the survival functions.*

Cox's Proportional Hazards Model

- How do we estimate covariate parameters, β, without specifying a form for $h_0(t)$?
- Assume there are no ties for failure times, and that y_i is not censored but y_i is it's failure time. Then the hazard function for the *ith* observation is, $h(y_i|x_i) = h_0(y_i)exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)$, and the total hazard for the at risk observations is, $\sum_{i':y_{i'} \geq y_i} h_0(y_i)exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)$. Here the sum over *i'* includes individuals which may or may not be censored in the future.
- The probability that the *ith* observation will fail rather then any of the other at risk individuals is, $\frac{h_0(y_i)exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'} \geq y_i} h_0(y_i)exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)} = \frac{exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'} \geq y_i} exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}$
- The baseline function has cancelled out.
- The last ratio is called the partial likelihood and can be used to numerically estimate the model parameters, derive *p*-values and confidence intervals on parameter estimates.