Additive hazards (Chapter 9)

Additive hazard models are an alternative to proportional hazard models. Here the hazard model is more directly similar to a linear model:

$$h(t|\mathbf{Z}(\mathbf{t})) = \beta_0(t) + \sum_{k=1}^{p} \beta_k(t) Z_k(t)$$

where the covariates are allowed to vary over time. We can also allow the covariates to vary over time, which we didn't do for the proportional hazards model. Instead, you can also fit a model where the covariates don't vary over time, so that $\beta_k(t) = \beta_k$.

The model with varying coefficients was actually developed first, in 1989 by Aalen, and the model with fixed covariates was developed by Lin and Yang in 1995. Note that these were developed more recently than the Cox proportional hazard model of 1972.

Additive hazards

For Aalen's model, estimation of the coefficient functions can be done using least-squares instead of maximum likelihood or partial maximum likelihood (used for proportional hazards).

For terms in the model, we have t_j, the death times, δ_j, the censoring indicators, $Z_{jk}(t)$, $k = 1, \ldots, p$, the covariate functions. We also have $Y_j(t) = 1$ if the jth individual is at risk at time t; otherwise $Y_j(t) = 0$. Left-truncation is allowed so that if the jth individual is left-truncated before time t, then $Y_j(t') = 0$ for $t' < t$.

Additive hazards

For the jth individual, the model is

$$h(t|\mathbf{Z}_j(t)) = \beta_0 + \sum_{k=1}^{p} \beta_k(t) Z_{jk}(t)$$

Instead of directly estimating $\beta_k(t)$, their cumulative versions are estimated first

$$B_k(t) = \int_0^t \beta_k(u)\, du, \quad k = 1, \ldots, p$$

$\beta_k(t)$ can be estimated from the slope of $B_k(t)$, which is often done using kernel smoothing techniques.

Additive hazards

To set this up like a regression model, we first get a design matrix $\mathbf{X}$ which has a column of 1s followed by a matrix representing the covariates. Thus

$$\mathbf{X}(t) = \begin{pmatrix} 1 \cdot Y_1(t) & Y_1(t)Z_{11}(t) & Y_1(t)Z_{12}(t) & \cdots & Y_1(t)Z_{1p}(t) \\ 1 \cdot Y_2(t) & Y_2(t)Z_{21}(t) & Y_2(t)Z_{22}(t) & \cdots & Y_2(t)Z_{2p}(t) \\ \vdots & \vdots & & \vdots & \\ 1 \cdot Y_n(t) & Y_n(t)Z_{n1}(t) & Y_n(t)Z_{n2}(t) & \cdots & Y_n(t)Z_{np}(t) \end{pmatrix}$$

We also let $\mathbf{I}(t)$ be an $n \times 1$ vector where the ith element is equal to 1 if the ith subject dies at time t and is otherwise 0.

Additive hazards

Then

$$\widehat{\mathbf{B}}(t) = \sum_{t_i \leq t} [\mathbf{X}(t_l)' \mathbf{X}(t_i)]^{-1} \mathbf{X}(t_i)' \mathbf{I}(t_i)$$

and the covariance matrix is

$$Var(\widehat{\mathbf{B}}(t)) = \sum_{t_i \leq t} [\mathbf{X}(t_l)' \mathbf{X}(t_i)]^{-1} \mathbf{X}(t_i)' \mathbf{I}^D(t_i) \mathbf{X}_i(t_i) \left\{ [\mathbf{X}(t_l)' \mathbf{X}(t_i)]^{-1} \right\}$$

where $\mathbf{I}^D(t_i)$ is the diagonal matrix made from $\mathbf{I}(t_i)$. The main thing to get from this is that $\widehat{\mathbf{B}}(t)$ looks similar to the standard way to find β in a multiple regression problem, except that we are adding up over times to get a cumulative version. Estimates of $\beta_k(t)$ are obtained from the slopes of $B_k(t)$. The estimator $\mathbf{B}(t)$ is only defined for values of t for which $\mathbf{X}(t_l)' \mathbf{X}(t_i)$ is nonsingular.

Additive hazard: two sample problem

As an example, consider the case that we have two groups (e.g., male and female). This will be the only covariate, so it doesn't change over time. Then the design matrix is

$$\begin{pmatrix} Y_1(t) & Y_1(t)Z_{11} \\ \vdots & \vdots \\ Y_n(t) & Y_n(t)Z_{n1} \end{pmatrix}$$

This leads to

$$\widehat{B}_0(t) = \sum_{t_i \leq t} d_i \left\{ \frac{1}{N_2(t_i)} - \frac{Z_{i-1}}{N_2(t_i)} \right\}$$

$$\widehat{B}_1(t) = \sum_{t_i \leq t} d_i \left\{ \frac{-1}{N_2(t_i)} + Z_{i1} \left[\frac{1}{N_1(t_i)} + \frac{1}{N_2(t_i)} \right] \right\}$$

where $N_k(t_i)$ is number at risk in group k at time t_i.

Additive hazard in R

To fit an additive hazard model in R, you can use the `timereg` package.

```
> install.packages("timereg")
> library(timereg)
Loading required package: survival
Loading required package: splines
> a <- aalen(Surv(start,stop,arrest) ~ const(fin) + const(prio)
+ const(age))
> summary(a)

> summary(a)
Additive Aalen Model
Test for nonparametric terms
...
Parametric terms :
             Coef.     SE Robust SE      z P-val
const(fin)  -0.044 0.006     0.019 -2.37 0.018
const(prio)  0.013 0.001     0.004  3.40 0.001
const(age)  -0.009 0.000     0.001 -7.35 0.0000
```

Additive hazard in R

The use of `const()` told R that the covariate was constant over time. Adding the time-varying covariate `employed` yields as part of the output

```
Test for non-significant effects
             Supremum-test of significance p-value H_0: B(t)=0
(Intercept)                          12.10                    0
employed                              5.36                    0
```

suggesting that employment is not constant over time. This has a negligible change on the other p-values, with the p-value for `fin` going to 0.018.

Additive hazard in R

If we add more terms to the model, then we get

```
a <- aalen(Surv(start,stop,arrest) ~ const(fin) + const(prio) +
const(age) + const(educ) + const(mar) + const(race) + employed)
Test for non-significant effects
             Supremum-test of significance p-value H_0: B(t)=0
(Intercept)                           9.81                   0
employed                              5.36                   0

Test for time invariant effects
                  Kolmogorov-Smirnov test p-value H_0:constant effe
(Intercept)                          1.5                        0
employed                             1.5                        0
                    Cramer von Mises test p-value H_0:constant effe
(Intercept)                        14.5                     0.025
employed                           29.6                     0.004
```

Additive hazard in R

```
Parametric terms :
              Coef.     SE Robust SE       z P-val
const(fin)   -0.044 0.006     0.019 -2.380 0.017
const(prio)   0.010 0.001     0.004  2.530 0.011
const(age)   -0.008 0.000     0.001 -6.240 0.000
const(educ) -0.035 0.003     0.010 -3.690 0.000
const(mar)   -0.023 0.007     0.025 -0.933 0.351
const(race)   0.021 0.009     0.028  0.752 0.452
```

Parametric survival models (chapter 11)

The last several weeks have used nonparametric methods to estimate survival functions, not assuming anything about the distribution of the time to failure. Here we go back to an earlier part of the course where we used distributions such as the Weibull to model survival times.

The models considered here are called *accelerated failure time models*, and are linear models in the log of the time. Let $\boldsymbol{\theta}$ be the vector of covariate coefficients and $\mathbf{Z}$ the covariate vector. If x is the time to failure, then

$$S(x|\mathbf{Z}) = S_0[\exp(\boldsymbol{\theta}'\mathbf{Z})x]$$

where $\exp(\boldsymbol{\theta}'\mathbf{Z})$ is called the *acceleration factor*. The hazard is

$$h(x|\mathbf{Z}) = \exp(\boldsymbol{\theta}'\mathbf{Z})h_0[\exp(\boldsymbol{\theta}'\mathbf{Z})x]$$

Parametric survival models

Another representation is

$$Y = \ln X = \mu + \boldsymbol{\gamma}'\mathbf{Z} + \sigma W$$

where $\boldsymbol{\gamma} = -\boldsymbol{\theta}$ are regression coefficients and W is the error distribution. The error distriubtion is where the parametric modeling comes in. For the usual multiple regression problem, we assume that errors are normally distributed. Here, however, we might use other distributions such as the Weibull or log-logistic for the error. The Weibull is flexible in that it can accommodate either increasing, decreasing, or flat hazard rates, and is the only distribution allowing proportional hazard model and accelerated failure time models as a special cases.

Parametric survival models usually have parameters estimated by maximum likelihood.

Parametric survival models

To review the Weibull distribution, it has

$$S_X(x) = \exp(-\lambda x^{\alpha})$$

$$h_X(x) = \lambda\alpha x^{\alpha-1}$$

The survival function for $Y = \ln X$ is

$$S_Y(y) = \exp(\lambda e^{\alpha y})$$

Letting $\lambda = \exp(-\mu/\sigma)$ and $\sigma = 1/\alpha$,

$$Y = \ln X = \mu + \sigma W$$

(In this model there are no covariates.) Here W has the extreme value distribution with density

$$f_W(w) = \exp(w - e^w)$$

$$S_W(w) = \exp(-e^w)$$

Parametric survival models

This gives

$$f_Y(y) = (1/\sigma)\exp[(y-\mu)/\sigma - e^{[(y-\mu)/\sigma]}] \qquad (12.2.4)$$

and

$$S_Y(y) = \exp(-e^{[(y-\mu)/\sigma]}). \qquad (12.2.5)$$

$$L = \prod_{j=1}^{n} [f_Y(y_j)]^{\delta_j} [S_Y(y_j)]^{(1-\delta_j)}$$

$$= \prod_{j=1}^{n} \left[\frac{1}{\sigma} f_W\left(\frac{y_j-\mu}{\sigma}\right)\right]^{\delta_j} \left[S_w\left(\frac{y_j-\mu}{\sigma}\right)\right]^{(1-\delta_j)}$$

Parametric survival models

You can either get maximum likelihood estimates of λ and α or of μ and σ. From the invariance property of maximum likelihood estimates, $\widehat{f(\theta)} = f(\widehat{\theta})$, so you can get ML estimates for one set of parameters and plug in to get the ML estimates for the other set of parameters.

Parametric survival models

You can add covariates to the model using

$$Y = \mu + \boldsymbol{\gamma}'\mathbf{Z} + \sigma W$$

Parametric survival models

Alternatively, you can use a log logistic distribution. For this distribution, the hazard increases and then decreases

$$S_X(x) = \frac{1}{1 + \lambda x^{\alpha}} \tag{12.3.1}$$

and

$$H_X(x) = \ln(1 + \lambda x^{\alpha}). \tag{12.3.2}$$

Taking the log transform of time, the univariate survival function for $Y = \ln X$ is

$$S_Y(y) = \frac{1}{1 + \lambda e^{\alpha y}} \tag{12.3.3}$$

This log linear model with no covariates is, from (12.1.1),

$$Y = \ln X = \mu + \sigma W, \tag{12.3.4}$$

where W is the standard logistic distribution with probability density function,

$$f_W(w) = e^w/(1 + e^w)^2 \tag{12.3.5}$$

and survival function,

$$S_W(w) = 1/(1 + e^w) \tag{12.3.6}$$

Parametric survival models

Other parametric families of models can be used as well, such as the log normal, gamma, and generalized gamma. For all of these models, we assume W has the distribution and we model $Y = \log X = \mu + \gamma + \sigma W$. The generalized gamma includes Weibull, exponential, and log normal as special cases. For the generalized gamma, the density is

$$f(w) = \frac{|\theta|[\exp(\theta w)/\theta^2]^{(1/\theta^2)} \exp[-\exp(\theta w)/\theta^2]}{\Gamma(1\theta^2)}$$

When $\theta = 1$, this reduces to the Weibull model. If $\sigma = 1$ in addition, then this reduces to an exponential model. For $\theta = 0$, the generalized gamma reduces to the log normal model.

Parametric survival models

Because we have nested models, you could use likelihood ratio tests to test which model best fit the data. For example, to see if the data is reasonably modeled as Weibull rather than generalized gamma, you can test whether $-2 \log \Lambda$ is significant using χ^2_1. There is one degree of freedom for the test because the alternative hypothesis (generalized gamma) has one more parameter than the null hypothesis (Weibull). If the result is NOT significant, then it is reasonable to use the Weibull since using the more complex model (the generalized gamma) does not give a singificant improvement over using the less complex model (Weibull). Thus, the null hypothesis here is

$$H_0 : \theta = 1$$

and

$$H_A : \theta \neq 1$$

Parametric survival models

Similarly, you can test whether the data is exponential by using $-2 \log \Lambda$ testing the exponential against the generalized gamma by seeing if this is significant using χ^2_2 since there are two more parameters for the generalized gamma than the exponential.

Parametric survival models

When comparing two models that aren't nested, for example, the log normal versus Weibull, you can use AIC:

$$AIC = -2 \log L + 2(p + k)$$

Here p is the number of covariates and k is the number of parameters for the parametric distributions. Thus, $k = 1$ for the exponential model, $k = 2$ for the Weibull and log-normal models, and $k = 3$ for the generalized gamma. Since the Weibull and log normal models have the same number of parameters, comparing them is equivalent to comparing their likelihoods and going with whichever model has the higher likelihood.

Parametric survivial models

An example from the book is to find the best parametric model for the leukemia bone marrow transplant data. The idea is to fit the models several times using software assuming different parametric models and to use the likelihoods from the output to compare AIC values for the different models.

Parametric survival models

TABLE 12.5
Results of Fitting Parametric Models to the Transplant Data

		Allo Transplants	*Auto Transplants*
Exponential	Log likelihood	−81.203	−68.653
	AIC	164.406	139.306
Weibull	Log likelihood	−72.879	−68.420
	AIC	149.758	140.840
Log logistic	Log likelihood	−71.722	−67.146
	AIC	147.444	138.292
Log normal	Log likelihood	−71.187	−66.847
	AIC	146.374	137.694
Generalized gamma	Log likelihood	−70.892	−66.781
	AIC	147.784	139.562
	$\hat{\theta}$	−0.633	−0.261
	SE[$\hat{\theta}$]	0.826	0.725
	p-value for $H_o : \theta = 0$	0.443	0.719
	p-value for $H_o : \theta = 1$	0.048	0.082

Parametric survival models

From the table, the lowest AIC is for the log normal model for both the Allo and Auto patients, so it would make sense to use this model. The results are also roughly consistent with the likelihood ratio testing approaching based on the p-values at the bottom of the table. The likelihood ratios suggest that the generalized gamma fits significantly better than Weibull for the allo patients, only slightly better than for the auto patients ($0.05 < p < .1$), but not signficantly better than the log normal distribution for either type of patient.

The log normal distribution

It might be worth looking at some of the properties of the log normal distribution since we haven't done that before. For statisticians, if X is log normal, then $\log X$ is normally distributed. This might make the log normal distribution seem not very interesting, since in many cases, you can just transform the data with a logarithm and work with normally distributed (transformed) data.

In some cases though, people might prefer working with data on the original scale and working directly with the log normal distribution. The log normal distribution is used a lot in engineering as well as survival anslysis and is useful for dealing with right-skewed distributions (much like the gamma and Weibull distributions).

The log normal distribution

In particular, if X is lognormal, then $Y = \log(X)$ is normal. Suppose X is normal with mean μ and standard deviation σ. Then
Some important properties when X is lognormal with parameters λ and ζ.

1. Lettting $\zeta^2 = \ln[1 + (\sigma/\mu)^2]$, $\lambda = \ln \mu - \zeta^2/2$,
   $P(X \leq x) = \Phi(\frac{\log x - \lambda}{\zeta})$. (You can use λ and ζ^2 as the parameters instead of μ and σ.)
2. $E(X) = \exp(\lambda + \zeta^2/2) = x_m\sqrt{1 + (\sigma/\mu)^2}$, where x_m is the median.
3. $Var(X) = \mu^2(e^{\zeta^2} - 1)$.
4. If σ/μ is not large (< 0.3), then $\zeta \approx \delta_X$.
5. $\lambda = \ln x_m$, where x_m is the median.

The quantity $\delta = \sigma/\mu$ is also known as the coefficient of variation and is used frequently in quality control.

The log normal distribution

Example. The time between severe earthquakes at a given region follows a lognormal distribution with a coefficient of variation of 40%. The expected time between severe earthquakes is 80 years.

(a) Determine the parameters of this lognormally distributed random variable.

(b) Determine the probability that a severe earthquake will occur within 20 yr from the previous one.

(c) Suppose the last severe earthquake in the region took place 100 yr ago. What is the probability that a severe earthquake will occur over the next year?

The log normal distribution

Example: solution (a).
(a) We use

$$\zeta^2 = \ln[1 + \delta^2] = \ln[1 + 0.4^2] = \ln[1.16] = 0.14842.$$

Thus $\zeta = \sqrt{0.14842} = 0.3853$. Also
$\lambda = \ln \mu - \zeta^2/2 = \ln(80) - 0.14842/2 = 4.3078$.

The log normal distribution

Example: solution (b).

(b)

$$P(X \leq 20) = \Phi\left(\frac{\ln 20 - 4.3078}{0.3853}\right) = \Phi(-3.41) = 0.0003$$

.

The log normal distribution

(c) If an earthquake occurs in the next year, it will have occurred within 101 years of the last earthquake (i.e., $X < 101$), and we are given than $X > 100$. So we want

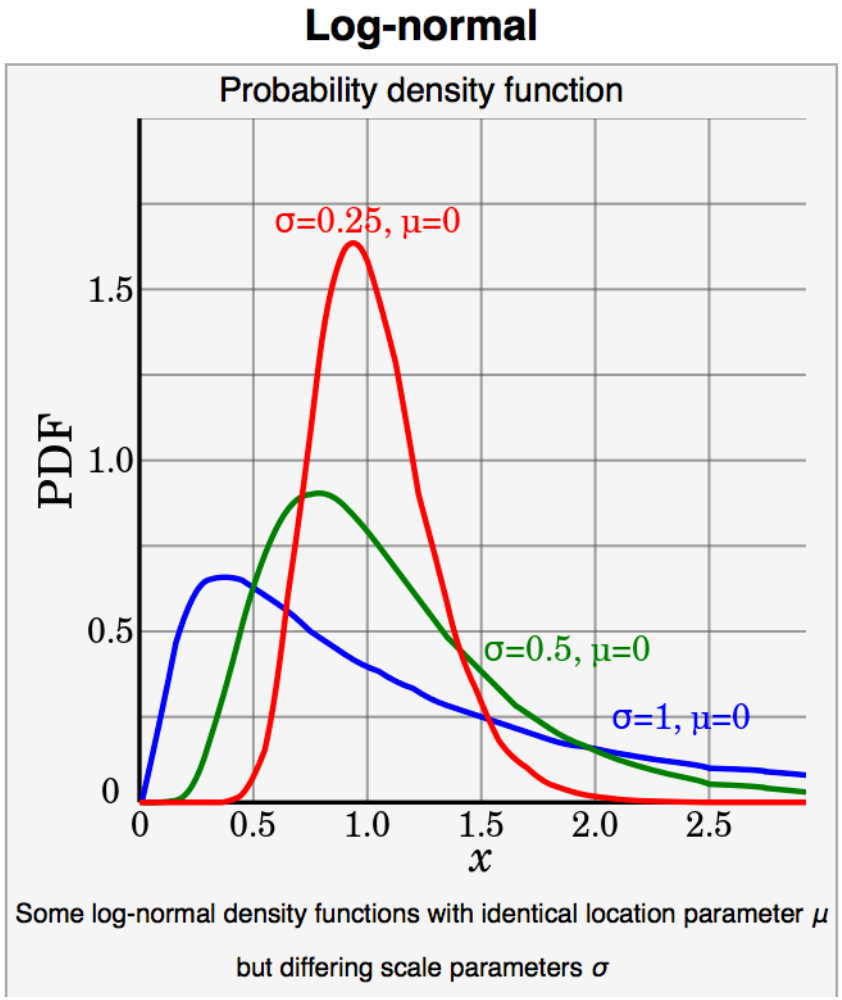
$$\begin{aligned} P(X < 101 \,|\, X > 100) &= \frac{P(100 < X < 101)}{P(X > 100)} \\ &= \frac{\Phi\left(\frac{\ln 101 - 4.3078}{0.3853}\right) - \Phi\left(\frac{\ln 100 - 4.3078}{0.3853}\right)}{1 - \Phi\left(\frac{\ln 100 - 4.3078}{0.3853}\right)} \\ &= \frac{\Phi(0.80) - \Phi(0.77)}{1 - \Phi(0.77)} \\ &= \frac{0.788 - 0.779}{1 - 0.779} \\ &= 0.04. \end{aligned}$$

The log normal distribution

Although it isn't terribly important to know the density (and it is easy to look up, anyway), it is good practice to be able to derive the density of a log normal distribution using transformation of variables techniques starting with the density for a normal distribution.

The book (Table 2.2, page 38) and Wikipedia gives the density assuming that the log normal distribution has parameters μ and σ which play the role of λ and ζ^2 as I have listed the properties. I think using new symbols for the parameters makes it easier to relate $Y = \log X$ to X where Y is normal with parameters μ and σ^2. If instead we say that X is log normal with parameters μ and σ, then $Y = \log X$ is normal but with weird parameters (in terms of μ and σ). I'd rather say that Y has mean μ and standard deviation σ and let X have different parameters.

Log-normal distribution from Wikipedia



Log-normal distribution from Wikipedia

Using a log normal regression model with a single covariate Z_1 equal to 1 if the patient received an auto transplant, we have the following regression model:

Parameter	*Estimate*	*Standard Error*	*Wald Chi Square*	*p-Value*
Intercept: μ	3.177	0.355	80.036	<0.0001
Type of Transplant: γ_1	0.054	0.463	0.0133	0.9080
Scale: σ	2.084	0.230	—	—

leukemia example

The results suggest that the type of transplant didn't signficantly affect the survival rates, which is consistent with results from nonparametric analyses in previous chapters.

Diagnostic tests

The book suggests using informal, visual diagnostics for appropriateness of models rather than formal tests. With graphical diagnostics, outliers and gross violations of models may become apparent, but small violations of the model probably will not be. The book makes the point that small samples may be under powered to look for violations of the model whereas large data sets tend to reject models very easily even when the model might lead to good predictions. This is related to the saying that "All models are wrong, but some are useful".

This is also increasingly a problem in that data sets are tending to get bigger, (e.g., big data), so that any model you can think of can typically be shown to be wrong. With a lot of data, if a model is close to being correct but is slightly wrong, then you can still have a lot of evidence (low p-value) that the model is wrong. Having a lot of evidence that a model is wrong is not the same as saying that the model is very wrong — you have to take into sample size and power when interpreting p-values.

Model diagnostics

The basic approach is to find a function of the cumulative hazard which is linear and check whether the same function of the cumulative hazard estimated from the data appears to be roughly linear. $\widehat{H}(x)$ can be estimated using the Nelson-Aalen estimator.

An example is that for the log logistic distribution,

$$H(x) = \ln(1 + \lambda x^{\alpha})$$

so

$$\ln\{\exp[H(x)] - 1\} = \ln \lambda + \alpha \ln x.$$

Thus,

$$\ln\{\exp[H(x)] - 1\}$$

plotted versus x should be approximatley linear.

Model diagnostics

For other distributions, different functions of the cumulative hazard are needed:

1. Exponential: $\widehat{H}(x)$ versus x
2. Weibull: $\ln \widehat{H}(x)$ versus $\ln x$
3. Log normal: $\Phi^{-1}(x)[1 - \exp(-\widehat{H}(x))]$ versus x

Diagnositc plots

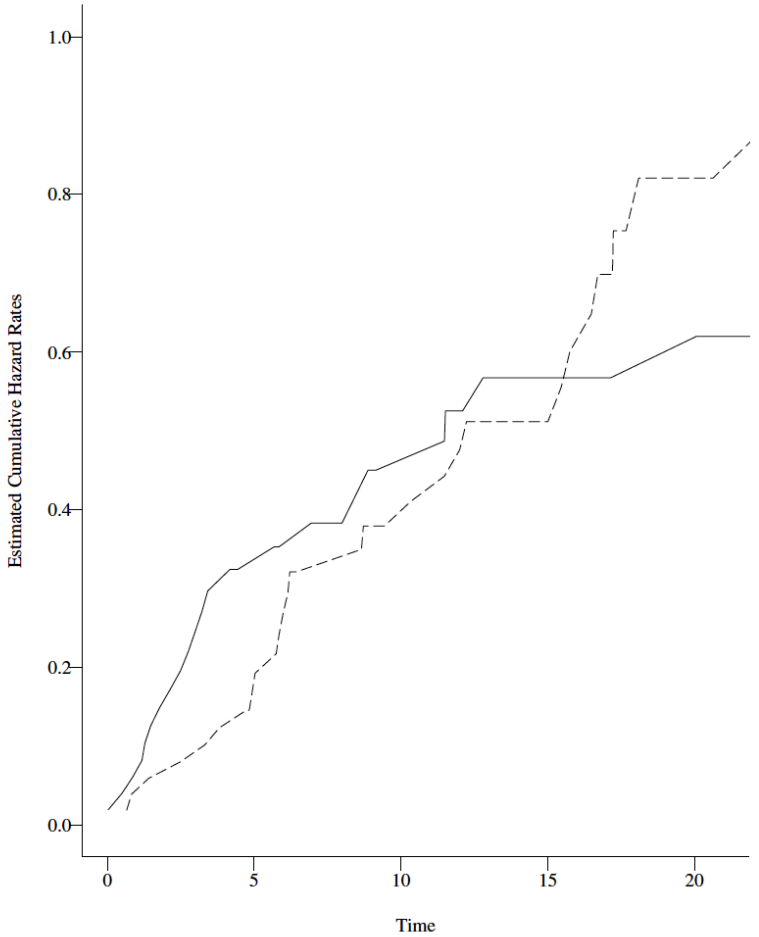


Figure 12.1 *Exponential hazard plot for the allo (solid line) and auto (dashed line) transplant groups.*

Diagnositc plots

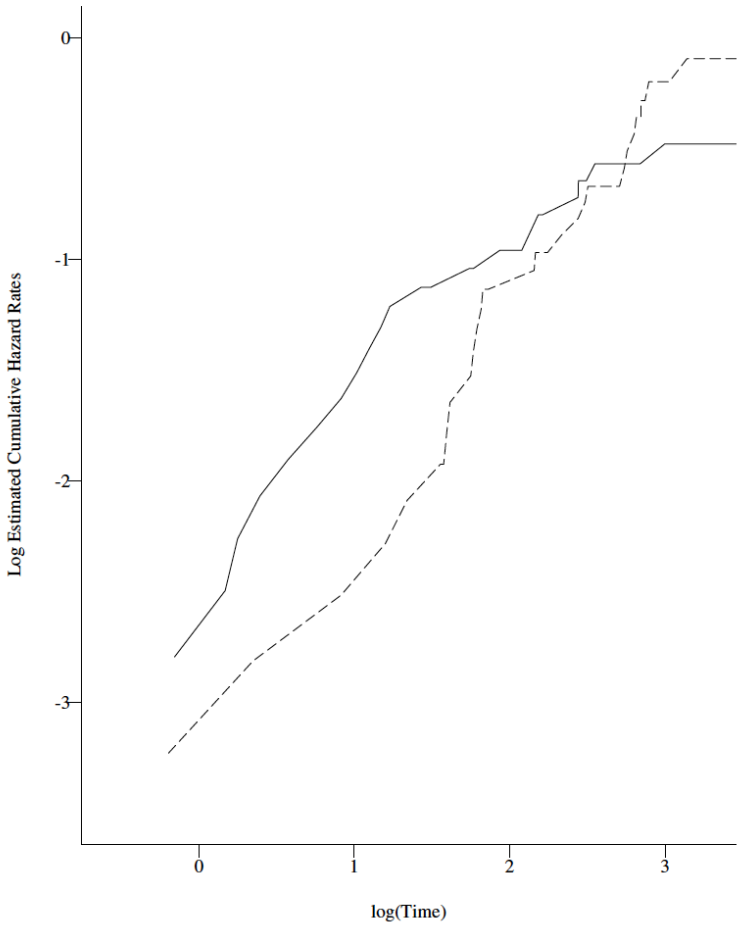


Figure 12.2 *Weibull hazard plot for the allo (solid line) and auto (dashed line) transplant groups.*

Diagnositc plots

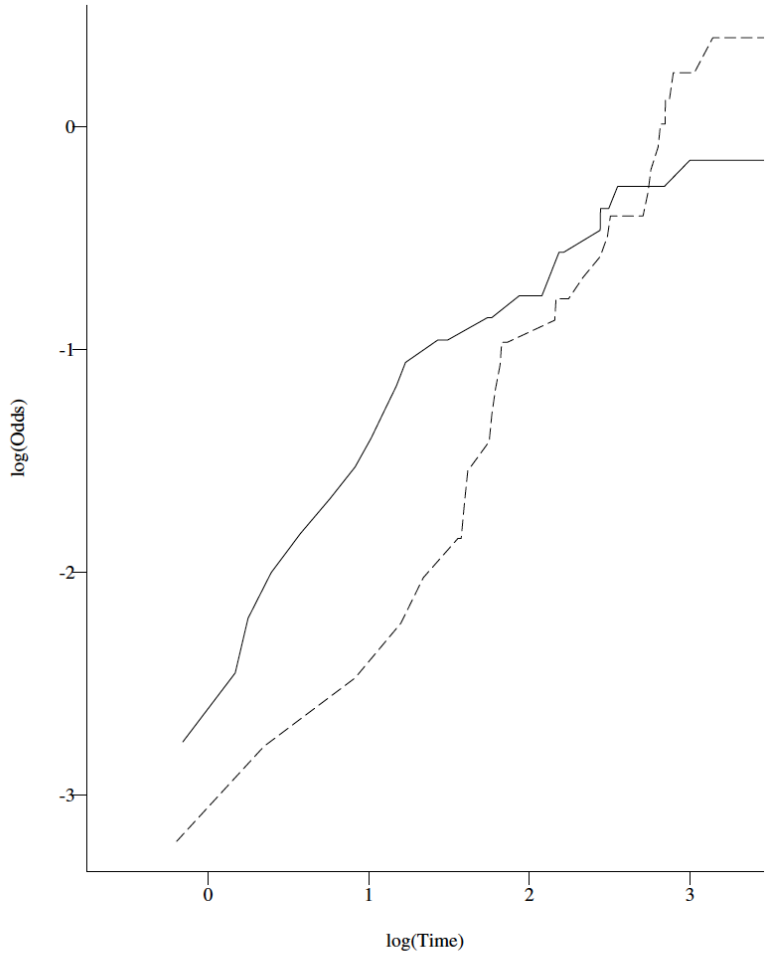


Figure 12.3 *Log logistic hazard plot for the allo (solid line) and auto (dashed line) transplant groups.*

Diagnositc plots

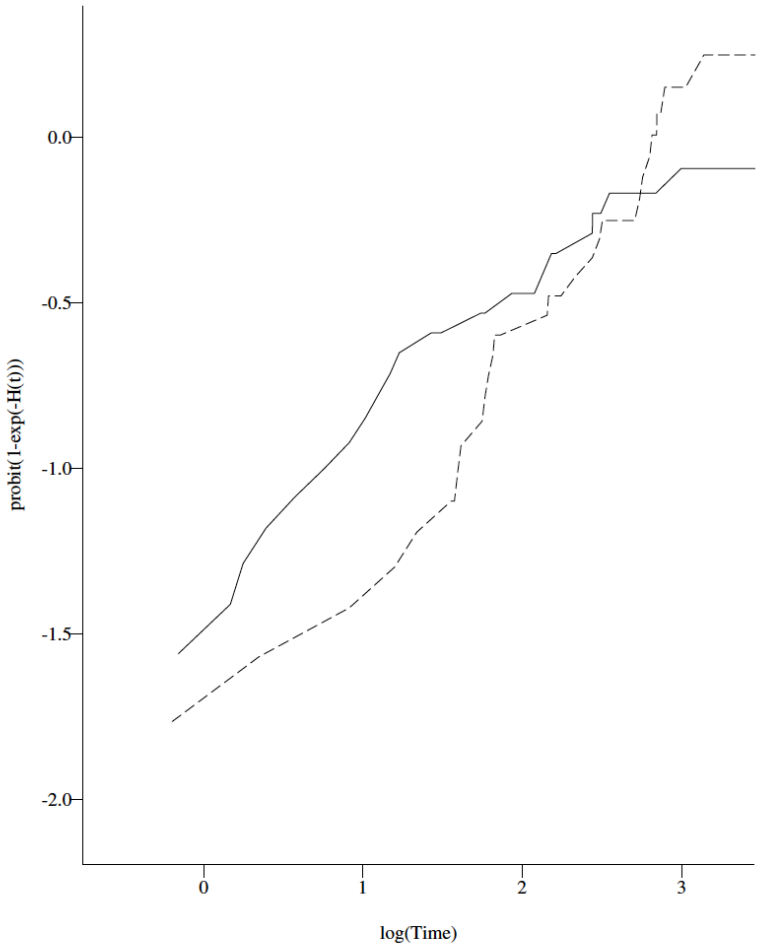


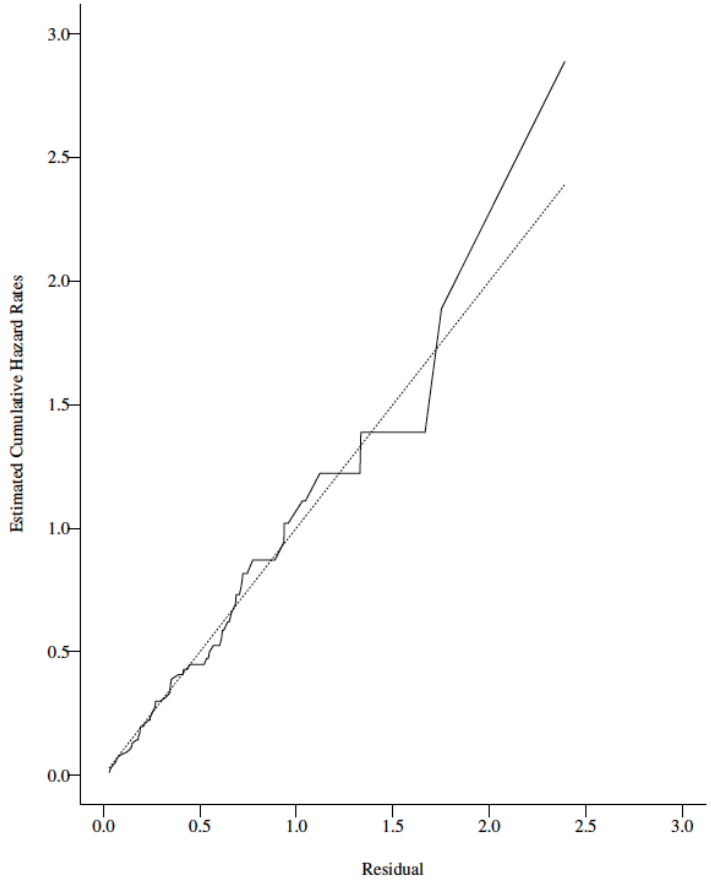
Figure 12.4 *Log normal hazard plot for the allo (solid line) and auto (dashed line) transplant groups.*

Diagnostic plots

Another approach is to use Cox-Snell residuals, r_j, which transform the cumulative hazard so that the residuals when plotted against the Nelson-Aalen estimator of the cumulative hazard should have slope of 1.

Exponential	$r_i = \hat{\lambda}\, t_i \exp\{\hat{\beta}^t \mathbf{Z}_i\},$
Weibull	$\hat{\lambda} \exp(\hat{\beta}^t \mathbf{Z}_i) t_1^{\hat{\alpha}},$
Log logistic	$\ln\left[\dfrac{1}{1 + \hat{\lambda} \exp(\hat{\beta}^t \mathbf{Z}_i) t_j^{\hat{\alpha}}}\right],$
Log normal	$\ln\left[1 - \Phi\left(\dfrac{\ln T_j - \hat{\mu} - \hat{\gamma}^t \mathbf{Z}_j}{\hat{\sigma}}\right)\right].$

Diagnositc plots



Multivariate survival data (Chapter 13)

Usually, we assume that individuals are independent, but this is not necessarily the case since siblings, cousins, and other relatives might end up in the same study. Also, if a study includes different populations or geographic regions, then there might be some correlation between people from similar populations or geographic regions due to environmental exposures incuding diet, lifestyle, pollution, etc. This might also be a concern in multi-center trials, where data is collected from different hospitals located in different cities and states, or even in multiple countries. Frailty models can also help with overdispersion — when there are unmeasured covariates that lead to greater than expected variation in survival times.

To model the associations within subgroups of individual survival times, frailty models can be used. A frailty is an unobserved random effect due to the subgroup. This is often modeled as a multiplicative effect on the hazard that is shared by all members of a group. .

Frailty models

To allow for subgroups, the cox proportional hazard model can be extended as

$$h_{ij}(t) = h_0(t)\exp(\sigma w_i + \beta' \mathbf{Z}_{ij}), \quad i = 1, \ldots, G, \quad j = 1, \ldots, n_i$$

where G is the number of groups, and n_i is the number of individuals in group i. The values $w_1, \ldots, w_G$ are called the *frailties* (the name suggesting that some groups are more frail or vulnerable than others). The w_i values are assumed to be from a distribution with mean 0 and variance 1. This makes this a random effects model as opposed to treating the subpopulation as a covariate.

Different distributions have been tried in the literature for w_i, including gamma, inverse Gaussian, uniform, and other distributions. The presence of a random effect can also be tested by testing the hypothesis that $\sigma = 0$ (if $\sigma = 0$, then there is no random effect.

Frailty models

A score test can be used to test for associations between subgroups or for overdispersion. Here the null hypothesis is that $\sigma = 0$ and the alternative is that $\sigma \neq 0$ in the model presented two slides ago. We now have t_{ij}, δ_{ij} and $\mathbf{Z}_{ij}$ for survival times, censoring, and covariates, where i indexes the group and j indexes the individual within the group. $Y_{ij}(t)$ is the number at risk at time t for the jth individual in subgroup i.

The idea of the test is to compute a test statistic under the assumption of no associations. Thus, a Cox proportional hazard model can be fit ignoring the subgroup information, and a test stiatistic is computed that is approximately standard normal. Rejecting the null hypothesis suggests that random effects for grouping should be used.

Frailty models

For the test statistic, you compute

$$S^{(0)}(t) = \sum_{i=1}^{G} \sum_{j=1}^{n_i} Y_{ij}(t) \exp(\mathbf{b}'\mathbf{Z}_{ij})$$

where $\mathbf{b}$ was estimated using the proportional hazards model without group membership information. Then residuals

$$M_{ij} = \delta_{ij} - \widehat{H}_0(t)(t_{ij}) \exp(\mathbf{b}'\mathbf{Z}_{ij})$$

are computed, and the test statistic is

$$T = \sum_{i=1}^{G} \left\{ \sum_{j=1}^{n_i} M_{ij} \right\}^2 - D + C$$

where D is the total number of deaths, and C is a correction factor. T is the test statistic.

Frailty models

The correction factor is

$$C = \sum_{i=1}^{G} \sum_{j=1}^{n_i} \frac{\delta ij}{S^{(0)}(t_{ij})^2} \sum_{b=1}^{G} \left[\sum_{k=1}^{n_i} Y_{bk}(t_{ij}) \exp(\mathbf{b}'\mathbf{Z}_{bk}) \right]^2$$

The test statistic T can also be written

$$T = \sum_{i=1}^{G} \sum_{j=1}^{n_i} \sum_{k=1}^{n_j} M_{ij} M_{ik} I(k \neq j) + \left(\sum_{i=1}^{G} \sum_{j=1}^{n_i} M_{ij}^2 - N \right) + C$$

where $N = \sum n_i$ is the total sample size. The first term is a function of the correlations between individuals; the second is a measure of overdispersion, and the correction term C tends to 0 as N increases. The quantity $T/\sqrt{V}$ where V is the estimated variance is approximately standard normal and can be used for testing. The formulas needed for V are not very enlightening, and are in equations 13.2.6–13.2.9 in the book.

Frailty models

As an example where there are many clusters, the book gives a study on mice where litters are used. From each litter, three female mice were selected, two as controls and one for treatment, which was exposure to a drug which caused tumors. The time measured is the time to development of a tumor.

TABLE 13.1
Data On 50 Litters of Rats

Group	*Treated Rat*	*Control Rats*	*Group*	*Treated Rat*	*Control Rats*
1	101^+	104^+, 49	26	89^+	104^+, 104^+
2	104^+	104^+, 102^+	27	78^+	104^+, 104^+
3	104^+	104^+, 104^+	28	104^+	81, 64
4	77^+	97^+, 79^+	29	86	94^+, 55
5	89^+	104^+, 104^+	30	34	104^+, 54
6	88	104^+, 96	31	76^+	87^+, 74^+
7	104	94^+, 77	32	103	84, 73
8	96	104^+, 104^+	33	102	104^+, 80^+
9	82^+	104^+, 77^+	34	80	104^+, 73^+
10	70	104^+, 77^+	35	45	104^+, 79^+
11	89	91^+, 90^+	36	94	104^+, 104^+
12	91^+	92^+, 70^+	37	104^+	104^+, 104^+
13	39	50, 45^+	38	104^+	101, 94^+
14	103	91^+, 69^+	39	76^+	84, 78
15	93^+	104^+, 103^+	40	80	80, 76^+
16	85^+	104^+, 72^+	41	72	104^+, 95^+
17	104^+	104^+, 63^+	42	73	104^+, 66
18	104^+	104^+, 74^+	43	92	104^+, 102
19	81^+	104^+, 69^+	44	104^+	98^+, 78^+
20	67	104^+, 68	45	55^+	104^+, 104^+
21	104^+	104^+, 104^+	46	49^+	83^+, 77^+
22	104^+	104^+, 104^+	47	89	104^+, 104^+
23	104^+	83^+, 40	48	88^+	99^+, 79^+
24	87^+	104^+, 104^+	49	103	104^+, 91^+
25	104^+	104^+, 104^+	50	104^+	104^+, 79

$^+$ Censored observation

Mice example

For the frailty model, the hypothesis of interest is whether the litter has an effect. If the litter has no effect, then we can just analyze the data ignoring the litter and just using placebo versus treatment as the only covariate. If the litter has an effect, then we could use a frailty model for the random effect of the litter.

A third alternative is to treat the litter as a covariate. In this case, you have to estimate a separate parameter for each litter, which means that you have to estimate 50 parameters. This is a lot more parameters than the frailty model where we treat the frailties as random variables rather than parameters. This allows better estimation of the other parameters of interest in the model. Note that there are 150 observations in the data, but only 40 of these are not censored.

For the mice example, the value of $T/\sqrt{V}$ is 1.33, which, as a z-score, is not significant and gives a two-tailed p-value of 0.184.

Gamma frailty model

If you decide to have a random effect in your model, one possibility is the Gamma frailty model. This is a variation on the first frailty model presented with

$$h_{ij}(t) = h_0(t) u_i \exp(\beta' \mathbf{Z})$$

where u_i is the random effect and is assumed to have a gamma distribution with mean 1 and variance θ. Thus

$$f(u) = \frac{u^{1/\theta - 1} \exp(-u/\theta)}{\Gamma(1/\theta)\theta^{1/\theta}}$$

Large values of θ suggest stronger correlations within groups, and therefore more separation between groups. If a parametric form is given to $h_0(t)$, the baseline hazard, then estimation of parameters can be done with maximum likelihood.

Gamma frailty model

The book fits the gamma frailty model to the mice data (even though the litter effect wasn't significant), and gets

Model	Treatment	Frailty
Cox with frailties	$b = 0.904, SE = 0.323$	$\widehat{\theta} = 0.472, SE(\widehat{\theta}) = 0.462$
Cox indpendence	$b = 0.897, SE = 0.317$	

Adding the frailties increased the estimated effect size of the treatment but also increased its standard error. The variability in the estimate of θ is such that it could plausibly be 0, resulting in the independence model.

Gamma frailty model

If the baseline hazard is not parametric, then instead of using maximum likelihood, the book suggests using the EM algorithm. This algorithm is used in numerous areas of statistics, genetics, and other disciplines and is an iterative approach.
The idea roughly is that if you don't have the complete likelihood due to unknown information, you fill in the data using expectations (the E-step). Then you maximize this likelihood (M-step). This gives you an estimate of the model. Using this estimate, you can make a better guess at the missing information (E-step again), so you fill this in. This better guess gives you an improved likelihood calculation (M-step again), which you can then use to get better estimates of the expected missing information. The process is iterated until convergence is achieved. There is a famous paper dealing with the EM algorithm written by statisticians in 1977 by Dempster, Laird, and Rubin. Rubin is famous for dealing with missing data problems. In addition to this paper, Nan Laird is known for work in longitudinal data analysis and statistical genetics.

EM algorithm

The EM algorithm comes up a lot in genetics where you have observed data (phenotypes such as blood type) with unobserved variables such as genotypes (OA versus AA genotypes both result in type A phenotype).

EM algorithm

The example in the book for doing the EM algorithm is rather difficult, and was not available in software at the time that the authors wrote the book, but they implemented a SAS macro to implement it. Variations on this EM algorithm have since resulted in research papers, some of which are still fairlly recent.

Rather than jumping into the details of the particular EM algorithm, we'll look at a simpler example to get the idea of how it works

EM algorithm

We'll do an example with coin flipping. You are given two coins, A and B, where the probability of heads for each coin is θ_A and θ_B, respectively. You wish to estimate $\theta = (\theta_A, \theta_B)$.

In the first part of the experiment, you randomly select a coin, then flip it 10 times and keep track of the number of heads. You repeat this procedure five times, so there are a total of 50 coin flips. You could represent the data by $(x_1, \ldots, x_5)$ and $(z_1, \ldots, z_5)$ where x_i is the number of heads on the ith set of 10 flips, and z_i is an indicator of which coin was chosen.

EM algorithm

Assuming that z_1 is not constant (you picked both coins at least once), then you can estimate θ by

$$\widehat{\theta}_A = \frac{\text{number of heads with } A}{\text{number of flips with } A}$$

$$\widehat{\theta}_B = \frac{\text{number of heads with } B}{\text{number of flips with } B}$$

This is the intuitive estimator and is also the maximum likelihood estimator.

EM algorithm

Equivalently, we could write

$$\widehat{\theta}_A = \frac{\sum_{i=1}^{5} x_i I(z_i = A)}{10 \cdot \sum_{i=1}^{5} I(z_i = A)}$$

$$\widehat{\theta}_B = \frac{\sum_{i=1}^{5} x_i I(z_i = B)}{10 \cdot \sum_{i=1}^{5} I(z_i = B)}$$

Assuming that the denominators are non-zero.

EM algorithm

Where things get more difficult is if you randomly pick a coin each time but you don't know which coin you picked. What you observe are the counts $x = (x_1, \ldots, x_5)$, but you don't observe $z = (z_1, \ldots, z_5)$. If the coins have very different biases, such as $\theta_A = 0.95$ and $\theta_B = 0.1$, then you would have a good guess at what the z values are.

If you observe $x = (9, 2, 10, 0, 0)$, then you could feel fairly confident that you picked coin A on the first and third trials, and coin B on the second, fourth, and fifth trials. However, if $\theta = (0.5, 0.6)$, then it will be very difficult to tell which trials were from which coin.

EM algorithm

For the case where z is unobserved, the variables in z are often called hidden variables or latent variables. (This is analogous to the random coefficients in the frailty model.)

This is where the EM algorithm comes in. A simplification of the algorithm is the following. Start with an initial guess, $\theta^{(0)} = (\theta_A^{(0)}, \theta_B^{(0)})$. Given this initial guess for θ, you could then take a best guess for $z = (z_1, \ldots, z_5)$. Given this value for z, you can then get a likelihood estimate for the data. The likelihood of the data is

$$L(\theta_A, \theta_B) = \prod_{i=1}^{5} \binom{10}{x_i} \theta_i^{x_i} (1 - \theta_i)^{10 - x_i}$$

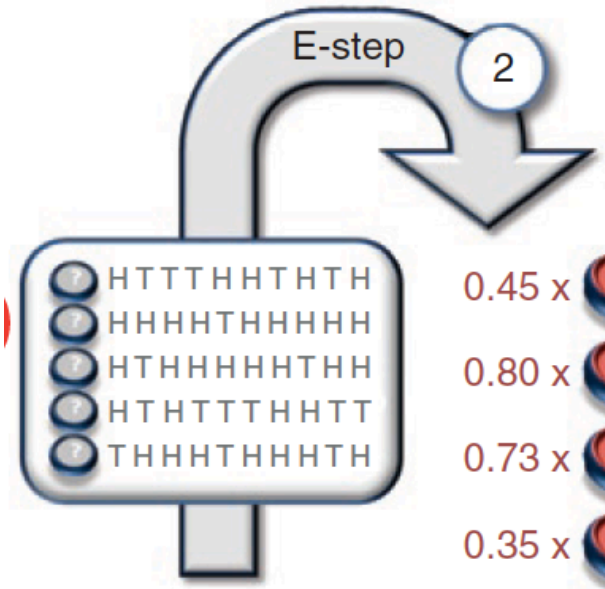
where $\theta_i = \theta_A$ if $z_i = A$ and $\theta_i = \theta_B$ if $z_i = B$. Given the likelihood function, you get an updated estimate of θ which you can call $\theta^{(1)} = (\theta_A^{(1)}, \theta_B^{(1)})$.

EM algorithm

Now that you have an updated estimate for θ, this will result in a possibly new best guess for z, which result in a new estimate of the likelihood, and so on. If the best guess for z doesn't change from one iteration to the next, then the algorithm stops.

For the actual EM algorithm, instead of using the best guess for z, we instead use the probability that the coin picked was A versus B given the data and our current guess of θ. This requires Bayes theorem, since we get $P(A|\text{data})$ from $P(\text{data}|A)$.

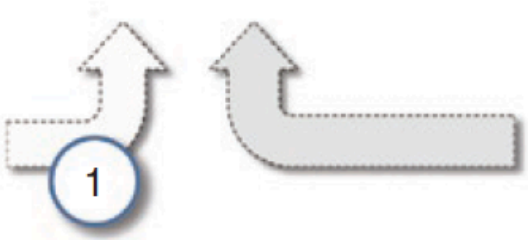
EM example



$\hat{\theta}_A^{(0)} = 0.60$

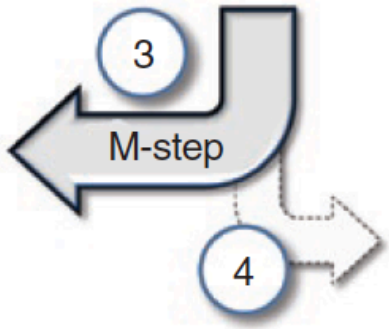
$\hat{\theta}_B^{(0)} = 0.50$

	Coin A	Coin B
0.45 x A, 0.55 x B	≈ 2.2 H, 2.2 T	≈ 2.8 H, 2.8 T
0.80 x A, 0.20 x B	≈ 7.2 H, 0.8 T	≈ 1.8 H, 0.2 T
0.73 x A, 0.27 x B	≈ 5.9 H, 1.5 T	≈ 2.1 H, 0.5 T
0.35 x A, 0.65 x B	≈ 1.4 H, 2.1 T	≈ 2.6 H, 3.9 T
0.65 x A, 0.35x B	≈ 4.5 H, 1.9 T	≈ 2.5 H, 1.1 T
	≈ 21.3 H, 8.6 T	≈ 11.7 H, 8.4 T



$$\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3 + 8.6} \approx 0.71$$

$$\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7 + 8.4} \approx 0.58$$



$\hat{\theta}_A^{(10)} \approx 0.80$

$\hat{\theta}_B^{(10)} \approx 0.52$

EM example

Here we use Bayes Theorem for each of the five samples to estimate the probabilities that each set of 10 coin flips came from that particular coin. This gives a distribution on the possible coins that were flipped. In reality only one coin was flipped for each set of 10, but we use a distribution to reflect our uncertainty on which coin was flipped.

For example, the probability that the first coin flipped was A given that 5 out of 10 flips were heads and that our initial guess is $\theta_A = 0.6$ and $\theta_B = 0.5$ is roughly 0.45, and $P(B|5) = 0.55$.

EM example

We can think of the likelihood now as

$$P(5,5) \times P(9,1) \times P(8,2) \times P(4,6) \times P(7,3)$$

where $P(9,1)$ means the probability of observing 9 heads and 1 tail. To get one of these probabilities, we can condition on the coin being used

$$\begin{aligned} P(9,1) &= P(9,1|A)P(A) + P(9,1|B)P(B) \\ &= \binom{10}{9}\theta_A^9(1-\theta_A)^1 \times 0.45 + \binom{10}{9}\theta_A^9(1-\theta_A)^1 \times 0.55 \end{aligned}$$

Taking a product over all of the observations gives us a new likelihood for θ_A and θ_B, and we can maximize this new likelihood to get updated estimates $\theta_A^{(1)}$ and $\theta_B^{(1)}$. These, in turn will imply slightly adjusted probabilities that observations 1 through 5 came from A versus B, which lead to a new likelihood, and so on.

EM example

An easier way of doing this is with the idea of membership weights. Here, let the 5 heads in the first toss belong to A with weight $5(.45) = 2.25$ and the 5 heads belong to B with weight $5(.55) = 2.75$. There are four membership weights which add up to 10 for each observation: heads belonging to A, heads belonging to B, tails belonging to A, tails belonging to B.

Then we update $\widehat{\theta}_A^{(0)}$ using the probability of belonging to A given heads, which is

$$\theta_A^{(1)} = \frac{21.3}{21.3 + 8.6} = 0.71$$

and

$$\theta_B^{(1)} = \frac{11.7}{11.7 + 8.4} = 0.58$$

The 10th iteration should be

$$\theta_A^{(10)} = 0.80, \quad \theta_B^{(10)} = 0.52$$