Proportional hazards regression

Patrick Breheny

October 3

Introduction

- Today we will begin discussing regression models for time-to-event data
- There are a number of ways one could think about modeling the dependency between the time to an event and the factors that might affect it
- The two most common approaches are known as *proportional hazards models* and *accelerated failure time models*

Proportional hazards

- We'll start with proportional hazards models
- As the name implies, the idea here is to model the hazard function directly:

$$\lambda_i(t) = \lambda(t) \exp(\mathbf{x}_i^T \boldsymbol{\beta})$$

- Here, the covariates act in a multiplicative manner upon the hazard function; note that the exponential function ensures that $\lambda_i(t)$ is always positive
- In this model, the hazard function for the ith subject always has the same general shape $\lambda(t)$, but can be, say, doubled or halved depending on a patient's risk factors

Exponential regression

- In general, any hazard function can be used; today, we'll restrict attention to the constant hazard for the sake of simplicity
- Thus, the *exponential regression* model is:

$$\lambda_i(t) = \lambda \exp(\mathbf{x}_i^T \boldsymbol{\beta})$$

- Note that if $\mathbf{x}_i$ contains an intercept term, we will have a problem with identifiability – there is no way to distinguish β_0 and λ

Identifiability

- For a variety of reasons (convenience, simplicity, numerical stability, accuracy of approximate inferential procedures), it is preferable to estimate β_0 rather than λ, so this is the parameterization we will use
- Of course, having estimated β_0, one can easily obtain estimates and confidence intervals for λ through the transformation $\lambda = \exp(\beta_0)$
- In today's lecture notes, we will discuss how to estimate the regression coefficients and carry out inference concerning them, and then illustrate these results using the `pbc` data

Solving a nonlinear system of equations

- Maximum likelihood estimation of β is complicated in exponential regression by the need to solve a nonlinear system of equations
- This cannot be done in closed form; some sort of iterative procedure is required
- The basic idea is to construct a linear approximation to the nonlinear system of equations, solve for $\widehat{\beta}$, re-approximate, and so on until convergence (this is known as *Newton's method*)
- We will begin by working out the score and Hessian with respect to the *linear predictor*, $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$

Log-likelihood, score, and Hessian

- Under independent censoring and assuming $\widetilde{T}_i|\mathbf{x}_i \sim \mathrm{Exp}(\lambda_i)$, the log-likelihood contribution of the ith subject in exponential regression is

$$\ell_i(\eta_i) = d_i\eta_i - t_i e^{\eta_i}$$

- The first and second derivatives with respect to the linear predictors are therefore

$$\frac{\partial \ell}{\partial \eta_i} = d_i - t_i e^{\eta_i}$$
$$\frac{\partial^2 \ell}{\partial \eta_i^2} = -t_i e^{\eta_i}$$

Vector/matrix versions

- Letting $\boldsymbol{\mu}$ denote the vector with ith element $t_i e^{\eta_i}$ and $\mathbf{W}$ denote the diagonal matrix with ith diagonal element $t_i e^{\eta_i}$, we can express the system of derivatives as

$$\begin{aligned} \nabla_{\boldsymbol{\eta}} \ell &= \mathbf{d} - \boldsymbol{\mu} \\ \nabla^2_{\boldsymbol{\eta}} \ell &= -\mathbf{W} \end{aligned}$$

- As we remarked earlier, solving for the values of β that satisfy the score equations is complicated because $\boldsymbol{\mu}$ is nonlinear; thus, consider the Taylor series approximation about $\tilde{\boldsymbol{\eta}}$

$$\begin{aligned} \nabla_{\boldsymbol{\eta}} \ell(\boldsymbol{\eta}) &\approx \nabla_{\boldsymbol{\eta}} \ell(\tilde{\boldsymbol{\eta}}) + \nabla^2_{\boldsymbol{\eta}} \ell(\tilde{\boldsymbol{\eta}})(\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}) \\ &= \mathbf{d} - \boldsymbol{\mu} + \mathbf{W}(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}) \end{aligned}$$

where $\boldsymbol{\mu}$ and $\mathbf{W}$ are fixed at $\tilde{\boldsymbol{\eta}}$

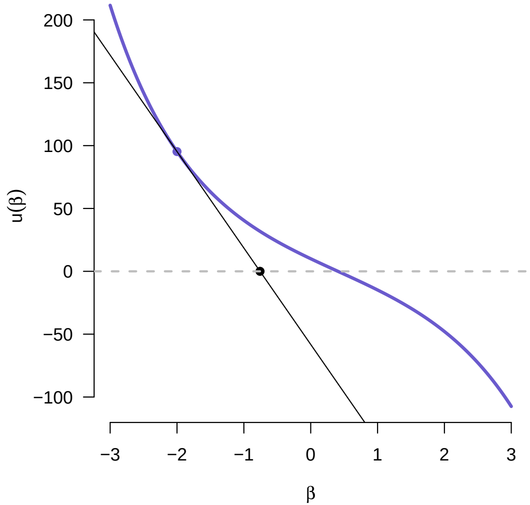
Solving for β

- All the preceding is only a means to an end, however – we're actually estimating $\boldsymbol{\beta}$, not $\boldsymbol{\eta}$
- Substituting this expression into the previous equation and solving for $\boldsymbol{\beta}$, we obtain

$$\widehat{\boldsymbol{\beta}} \leftarrow \widetilde{\boldsymbol{\beta}} + (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{d} - \boldsymbol{\mu})$$

- Again, this is an iterative process, which means that this is not an exact solution for $\widehat{\boldsymbol{\beta}}$; rather, we must solve for $\widehat{\boldsymbol{\beta}}$, recompute $\boldsymbol{\mu}$ and $\mathbf{W}$, re-solve for $\widehat{\boldsymbol{\beta}}$, and so on
- Newton's method will converge to the MLE (although this is not absolutely guaranteed) provided that the likelihood is log-concave and coercive, both of which (typically) hold for exponential regression

Newton's method: animation



Crude R code

- Below is some crude R code providing an implementation of this algorithm

```
b <- rep(0, ncol(X))
for (i in 1:20) {
  eta <- as.numeric(X%*%b)
  mu <- t*exp(eta)
  W <- diag(t*exp(eta))
  b <- b + solve(t(X) %*% W %*% X) %*% t(X) %*% (d-mu)
}
```

- This is crude in the sense that it isn't as efficient as it could be and in that it assumes convergence will occur in 20 iterations; a better algorithm would check for convergence by examining whether $\widehat{\beta}$ has stopped changing

Wald approach

- Since $\widehat{\beta}$ is the MLE, our derivation of the Wald results from earlier means that

$$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathrm{N}\left(\boldsymbol{\beta}, \mathbf{I}^{-1}\right);$$

  we just have to work out the information matrix with respect to $\boldsymbol{\beta}$

- Applying the chain rule, we have

$$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathrm{N}\left(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\right)$$

- It is very easy, therefore, to construct confidence intervals for β_j with $\widehat{\beta}_j \pm z_{1-\alpha/2}\mathrm{SE}_j$, where $\mathrm{SE}_j = \sqrt{(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}_{jj}}$

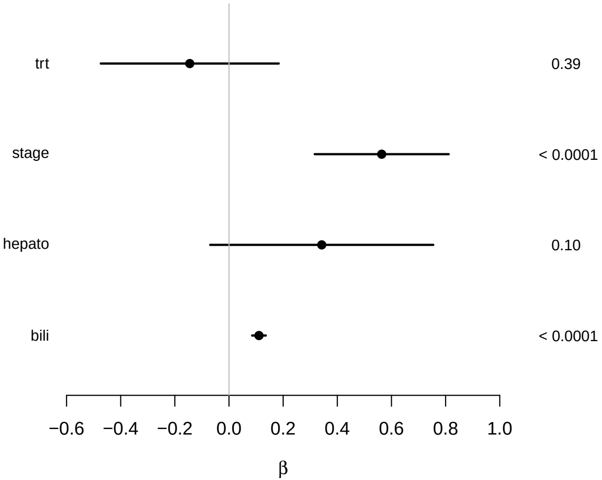
Likelihood ratio approach

- One could also, in principle, construct likelihood ratio confidence intervals
- As we remarked last time, this would involve profiling; i.e., calculating the profile likelihood $L(\beta_j, \widehat{\boldsymbol{\beta}}_{-j}(\beta_j))$ over a range of values for β_j
- Unfortunately, you would need to write your own software to do this; the `survival` package does not offer this as an option

pbc data: Setup

- To illustrate, let's fit an exponential regression model to the `pbc` data, and include the following four factors as predictors:
  - `trt`: Treatment (D-penicillamine, placebo)
  - `stage`: Histologic stage of disease (1, 2, 3, 4)
  - `hepato`: Presence of hepatomegaly (enlarged liver)
  - `bili`: Serum bilirubin (mg/dl)
- We can fit this model using our crude `R` code (the `survival` package does have a function for exponential regression, but its setup doesn't exactly match ours today, so I'm postponing coverage of the function to next week)

Results


trt
stage
hepato
bili
0.39
< 0.0001
0.10
< 0.0001
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
β

Interpretation of coefficients

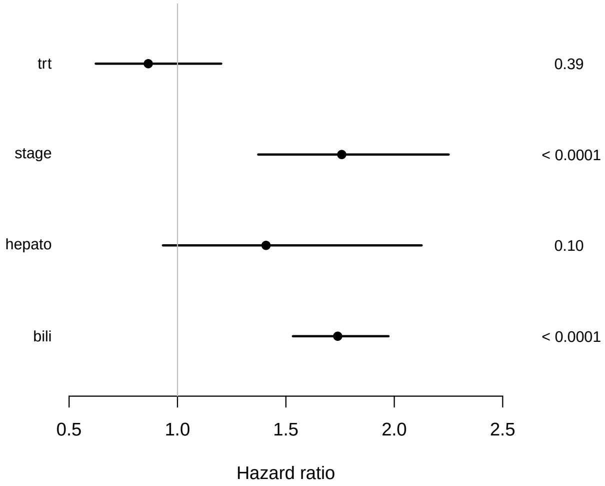
- As in other regression models, the interpretation of the regression coefficients involves the effect of changing one factor while all others remain the same
- Consider a hypothetical comparison between two individuals whose explanatory variables are the same, except for variable j, where it differs by $\delta_j = x_{1j} - x_{2j}$:

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \exp(\delta_j \beta_j)$$

Hazard ratios

- Note that for any proportional hazard model, $\lambda_1(t)/\lambda_2(t)$ is a constant with respect to time
- This constant is known as the *hazard ratio*, and typically abbreviated HR, although some authors refer to it as the *relative risk* (RR)
- Thus, the interpretation of a regression coefficient in a proportional hazards model is that $e^{\delta\beta}$ is the hazard ratio for a δ-unit change in that covariate
- In particular, $\mathrm{HR} = e^{\beta}$ for a one-unit change
- So, for stage in our pbc example, $\mathrm{HR} = e^{0.564} = 1.76$; a one-unit change in stage increases a patient's hazard by 76%

Results (hazard ratios; $\delta_{\text{bili}} = 5$)

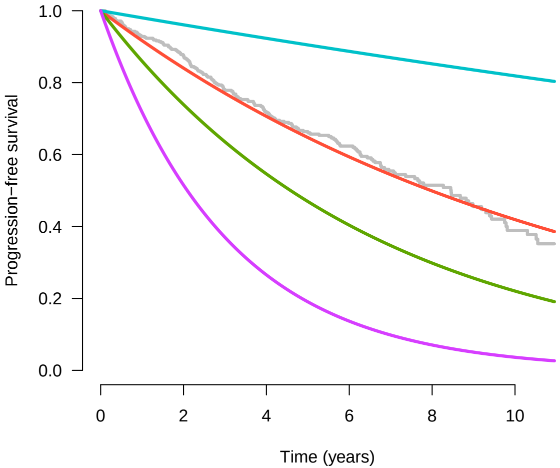


Wald, Score, and Likelihood ratio intervals

- As in the previous lecture, note that the Wald CIs account for the uncertainty with respect to the other parameters:
  - Wald SE is $\sqrt{(\mathbf{I}^{-1})_{jj}} = 0.126$
  - Naïve SE is $\sqrt{(\mathbf{I}_{jj})^{-1}} = 0.024$
- Score and LR confidence intervals require profiling; our next homework assignment asks you to calculate these intervals and compare them to the Wald interval

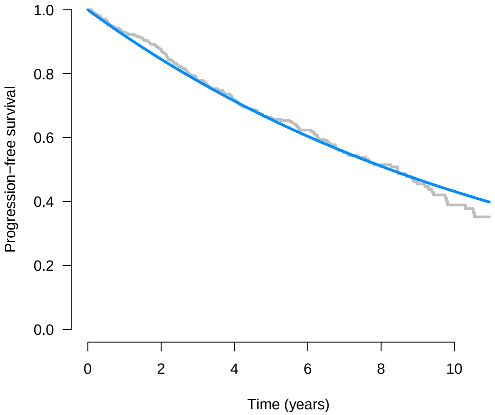
Predicted survival: Some examples

We can also predict survival curves at the individual level



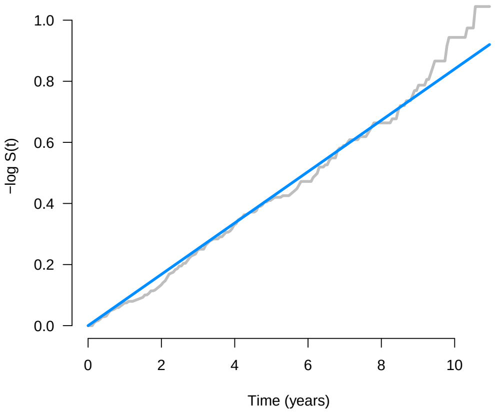
Diagnostic plot (original scale)

As a crude diagnostic plot to check whether the exponential distribution seems reasonable, we can plot the Kaplan-Meier estimate against the best exponential fit:



Diagnostic plot (linear)

Alternatively, since the exponential model implies $-\log S(t) = \lambda t$, we can obtain a linear version of the diagnostic plot:

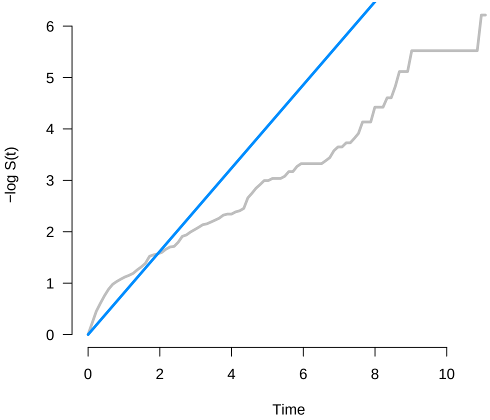


Limitations

- These diagnostic plots, although useful for identifying gross lack of fit, have some clear limitations
- The main limitation is that our model does not assume $\widetilde{T}_i \sim \mathrm{Exp}(\lambda)$, but rather that $\widetilde{T}_i | \mathbf{x}_i \sim \mathrm{Exp}(\lambda_i)$
- Thus, we may see a departure from linearity in the plot on the previous page, but it doesn't necessarily imply a violation of model assumptions

Diagnostic plot (simulated)

For example, consider this simulated diagnostic plot for two groups, each independently following an exponential distribution, but with different rate parameters:



Comments

- Nevertheless, these diagnostic plots are useful when that the covariates do not have an overwhelming effect on survival (covariates do not "dominate")
- If any covariates do have overwhelming effects, one may considering stratifying the diagnostic plots
- For example, we may wish to construct separate diagnostic plots for each stage in our pbc example

Residuals?

- In linear regression, of course, we don't face these issues because we can directly examine residuals
- In survival analysis, however, residuals are more complicated in that some of them will be censored
- There are ways of dealing with this, and of obtaining residuals for time-to-event regression models, but we will postpone this discussion for a later lecture