Hypothesis Tests in Survival Analysis

David M. Rocke

October 14, 2025

Maximum Partial Likelihood

If subject $j(i)$ is the one who fails at time t_i, then the *partial likelihood* is

$$L(\beta|T) = \prod_i \frac{\theta_{j(i)}}{\sum_{k \in R(t_i)} \theta_k}$$

where T stands for all the data including times, censoring, and covariate values, while β is the vector of coefficients.

The *partial log likelihood* is

$$\ell\ell(\beta|T) = \sum_i \left[\ln[\theta_{j(i)}] - \ln\left(\sum_{k \in R(t_i)} \theta_k \right) \right]$$

and with $\theta_k = \exp(\eta_k)$, let

$$\begin{aligned} \theta_k^m &= \frac{\partial}{\partial \beta_m} \theta_k \\ &= \theta_k \frac{\partial}{\partial \beta_m} \eta_k \\ &= \theta_k x_{mk} \end{aligned}$$

Then

$$\ell\ell(\beta|T) = \sum_i \left[\ln[\theta_{j(i)}] - \ln\left(\sum_{k \in R(t_i)} \theta_k \right) \right]$$

$$\frac{\partial}{\partial \beta_m} \ell\ell(\beta|T) = \sum_i [\theta_{j(i)}]^{-1} \theta^m_{j(i)} - \left[\sum_{k \in R(t_i)} \theta_k \right]^{-1} \sum_{k \in R(t_i)} \theta^m_k$$

which is the gradient vector, AKA the score statistic, and similarly we can derive the Hessian, whose negative inverse is the Fisher information. This can be used with a variety of optimization techniques such as Newton's method to find the MLE. Similar calculations can be used with tied event times.

In logistic regression, if all the values of a covariate for cases are larger than all the values for controls, or the reverse, then the covariate is very predictive, but the coefficient will diverge in estimation to $\pm\infty$. The same happens in Cox regression if the covariate values for the individuals with events are increasing (or decreasing) as the event times go from smallest to largest.
Paradoxically, this makes the numerical optimization invalid while strongly indicating the covariate is related to risk.

Wald Tests

We use the maximum partial likelihood estimates $\hat{\beta}$ of the parameter vector β which has estimated covariance matrix V from the Fisher information. The diagonal entries of V are the squares of the standard errors which we can use for tests and confidence intervals for single coefficients (these are given in the output). A linear combination $c^\top \hat{\beta}$ of coefficients has covariance matrix $c^\top V c$. The hypothesis $H_0 : \hat{\beta} = \beta_0$ can be tested with $(\hat{\beta} - \beta_0)^\top V (\hat{\beta} - \beta_0)$ which is asymptotically $\chi^2(p)$ under the null, where p is the number of parameters.

Likelihood Ratio Tests

Asymptotically, the log likelihood ratio $2[\ell\ell(\hat{\beta}) - \ell\ell(\beta_0)]$ is $\chi^2(p)$. This test, as well as the Wald test, can be used with partial specification by the null hypothesis of the coefficients in which case the dimension of the χ^2 statistic is the number of linear constraints. For example, if one coefficient is specified to be zero, this is equivalent to leaving that variable out and re-running the optimization. That is a test of that one coefficient and has dimension 1. Note that the Wald test of one coefficient uses the previous coefficient estimates, whereas the LR test re-estimates all the coefficients.

Score Tests

The score statistic AKA the gradient is 0 at the MLE. Let $G(\beta)$ be the score statistic, so that $G(\hat{\beta}) = 0$. The statistic

$$G(\beta_0)^\top V G(\beta_0)$$

has asymptotically a $\chi^2(p)$ distribution as well. If there are no ties, then the score test and the log rank test (as used in `survdiff`) are the same. In many cases, the LR test has faster convergence than the Wald test, though the book indicates that they are similar. The score test is generally less accurate.

Coding and Transforming Predictors

- A factor is a categorical covariate. If it has two levels, and those are coded as 0 and 1 in a numerical variable, then the coefficient is the predicted difference in the two levels (such as male/female).
- If there are more than two levels, then one can represent the factor with one fewer predictors than the number of levels. For example if the coding is group $= 1, 2, 3$, then we could define $x_1 = 1$ iff group $=2$ and $x_2 = 1$ iff group $= 3$.

Coding and Transforming Predictors

- This is rather old fashioned, though may sometimes be useful. Instead

  `dtype = factor(dtype,labels = c("NHL","HOD"))`

  redefines the variable to be a *factor*, which is inherently categorical.
- The coefficients though are by default comparisons with the first level.

Coding and Transforming Predictors

- Numerical variables may be transformed linearly so that the coefficient is in interpretable units.
- It also may improve the model to use the log, square root, or inverse of the original variable.
- The urge to categorize numeric variables should be resisted unless there is strong evidence that it helps.
- Hemoglobin A1C test (from CDC web page):

$$\begin{cases} \text{Normal} & \text{A1C} < 5.7\% \\ \text{Prediabetes} & 5.7\% \leq \text{A1C} \leq 6.4\% \\ \text{Diabetes} & \text{A1C} \geq 6.5\% \end{cases}$$

Use of Tests in Model Building

Generally in a survival analysis we have chosen a response, such as progression-free survival, and perhaps a main predictor, such as drug vs. standard of care. We may have other covariates of interest which are thought possibly to influence survival, or even perhaps the efficacy of the drug (the latter would imply an interaction term). If a covariate is not statistically significant, does that mean we should remove it from the model? Well, not necessarily.

We could compare the models with a measure of predictive performance such as the Akaiki Information Criterion (AIC) or the Bayesian Information Criterion (BIC). We might keep a predictor in the model because it is useful or because other studies have used it.

For clinical trials, the analysis must be prespecified and usually consists of a simple comparison of treatment and standard of care (remembering that these trials usually are randomized). Secondary analysis, also usually prespecified, can encompass covariates.

ANOVA

The analysis of variance is in linear regression the division of the sums of squares into parts assigned to covariates and interactions as well as the total and the error term. More generally, this describes a comparison of two models in which one is derived from the other by omitting terms. This can be done for many types of regression models including the ones in `coxph`.

OED: (post-classical Latin) *analysis*: act of resolving (something) into its elements (13th cent. in British and continental sources).

KMsurv hodg data set

Data on lymphoma: Hodgins disease and non-Hodkins lymphoma:

```
gtype =   Graft type (1=allogenic, 2=autologous)
dtype =   Disease type (1=Non Hodgkin lymphoma, 2=Hodgkins disease)
time =    Time to death or relapse, days
delta =   Death/relapse indicator (0=alive, 1=dead)
score =   Karnofsky score
wtime =   Waiting time to transplant in months
```

Karnovsky score indicates general functionality of the individual. There are 43 patients, 20 with Hodgkin's disease and and 23 with non-Hodkins lymphoma.

```
> hodg.cox1 <- coxph(hodg.surv ~ gtype * dtype + score + wtime, data = hodg2)
> anova(hodg.cox1)
Analysis of Deviance Table
 Cox model: response is hodg.surv
Terms added sequentially (first to last)

             loglik   Chisq Df Pr(>|Chi|)
NULL        -87.258
gtype       -87.194  0.1285  1    0.71996
dtype       -86.995  0.3973  1    0.52848
score       -74.445 25.1003  1  5.442e-07 ***
wtime       -73.899  1.0920  1    0.29604
gtype:dtype -71.181  5.4357  1    0.01973 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

This sequential addition of variables is not the most useful. The test indicates the usefulness of the given variable in a model already including variables that proceed it in the list.

```
> drop1(hodg.cox1)
Single term deletions

Model:
hodg.surv ~ gtype * dtype + score + wtime
            Df    AIC
<none>         152.36                    Any drop increases the AIC (bad)
score        1 167.60
wtime        1 153.64                    Can't drop gtype or dtype
gtype:dtype  1 155.80                      if gtype:dtype is in the model

> drop1(hodg.cox1,test="Chisq")
Single term deletions

Model:
hodg.surv ~ gtype * dtype + score + wtime
            Df    AIC     LRT Pr(>Chi)
<none>         152.36
score        1 167.60 17.2365  3.3e-05 ***        score and gtype:dtype
wtime        1 153.64  3.2792  0.07016 .            are significant by LR test
gtype:dtype  1 155.80  5.4357  0.01973 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The command `drop1` respects hierarchy, meaning that if an interaction is in the model then none of the subsidiary terms can be dropped. A LR test can be added but the AIC is always given.

$$\text{AIC} = -2\ell\ell + 2p$$

where p is the effective number of parameters. When terms are added to a model, the $\ell\ell$ cannot drop, but when penalized by the dimension of the predictors it may.

The penalty term is $2p$ by default but may be set to be kp. If $k = \ln(n)$ this is called the Bayesian Information Criterion (BIC) which favors smaller models than the AIC. Significance testing usually results in smaller models than the AIC or else the same model.

If there are missing values, then the models may be fit to different data sets which makes the inference invalid.

There is also an `add1` command which requires a term indicating the largest possible model that can be considered. An example is on the next slide.

```
> add1(hodg.cox1,scope = ~ gtype * dtype * score * wtime,test="Chisq")
Single term additions

Model:
hodg.surv ~ gtype * dtype + score + wtime
            Df    AIC    LRT Pr(>Chi)
<none>         152.36
gtype:score  1 149.03 5.3339  0.02091 *
dtype:score  1 153.13 1.2355  0.26634
gtype:wtime  1 153.13 1.2321  0.26699
dtype:wtime  1 154.27 0.0884  0.76619
score:wtime  1 154.30 0.0630  0.80188
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```