Notes of "Regression Methods in Biostatistics"

Notes of "Regression Methods in Biostatistics"

Di Zhen

Mar, 2020

Ch 3.5 Basic Method for Survival Analysis

3.5.1 Right Censoring

- A survival time is said to be **right-censored** at time t if it is only known to be greater than t.
  - e.g. a patient had only been under observation for 6 weeks, so we only know the relapse time is longer than that. "right" because on a graph the relapse time would lie somewhere to the right of the censoring time of 6 weeks.
- Survival analysis makes it possible to analysis right-censored data like these without bias or losing information contained in the length of the follow-up times.
- To deal with right-censoring ,we have to assume **independent censoring**:
  - After adjustment for covariates, future event risk for a censored subject does not differ from the risk among other subjects who remain in follow-up and have the same covariate values.
  - Under this assumption, subjects are censored independent of their future risk.

3.5.2 Kaplan-Meier Estimator of the Survival Function

- The **survival function** at time t, denoted $S(t)$, is the probability of being event-free at t; equivalently, the probability that the survival time is greater than t.
  - The survival function estimate given by the chain of conditional probabilities is equal to the sample proportion still in remission.
- **Kaplan-Meier Estimator**: The patient censored during certain week has disappeared from the denominator and does not contribute to the calculations for any subsequent week.
  - This method requires modification in the case of **competing risks data** where **cumulative incidence functions** define the probability of failure in the presence of other causes of failure.

3.5.3 Interpretation of Kaplan-Meier Curves

- Survival curve descend rapidly: high risk of relapse, short time in remission
- Survival curve descend slowly: low risk of relapse, long time in remission

3.5.4 Median Survival

- **Median survival time** is the time at which half the relevant population has experienced the outcome event.
    - In the absence of censoring, with every survival time observed exactly, the median survival time could be simply estimated by the sample median of survival times, i.e. the earliest time at which half the study participants have experienced the event.
    - In the presence of censoring, we need to use Kaplan-Meier estimate $\hat{S}(t)$ to estimate the median. The median survival time is estimated by the earliest time at which Kaplan-Meier curve dips below 0.5
- Other quantiles of distribution of survival times can be obtained from Kaplan-Meier estimate $\hat{S}(t)$.
    - The pth quantile is estimated as the earliest time at which the Kaplan-Meier curve drops below $1-p$.
    - Limitation: when the curve does not reach $1-p$, the pth percentile cannot be estimated.
- We are unable to estimate the mean of distribution in the typical case.
- Graphs are useful for giving overall impressions of the survival function, but it is difficult to read quantities from them.

3.5.5 Cumulative Event Function

- The **cumulative event function** at time t, denoted $F(t)$, is the probability that the event has occurred by time t, or equivalently, the probability that the survival time is less than or equal to t. Note that $F(t) = 1 - S(t)$.
- Parametric methods resting on explicit assumptions about the form of these distributions are somewhat less robust than Kaplan-Meier approach.

3.5.6 Comparing Groups Using the Log-rank Test

- **Log-rank test** is used for comparison of the survival experience of two or more groups.
    - Null hypothesis for this test is that the survival distributions being compared are equal at all follow-up times.
    - Log-rank test also accommodates right-censoring.
    - It works by comparing observed numbers of events in each group to the number expected if the survival functions were the same. The comparison accounts for differences in length of follow-up in calculating the expected numbers of events.
- The log-rank test is easily generalized to the comparison of more than two groups.
    - The log-rank test statistic for $K > 2$ groups follows an approximate χ^2 distribution with $K-1$ degrees of freedom.
    - Null hypothesis: $H_0 : S_1 = \ldots = S_k(t)$ for all t. Where $S_k(t)$ is the survival function for the kth group at time t.
    - Alternative hypothesis: some or all of the survival curves differ at one or more points in time.
- When the null hypothesis is rejected, we can know the differences by:
    - Kaplan-Meier plots

    - pairwise log-rank tests
- If there are more than 2 groups which are defined by ordered categories (e.g. disease stages) or categories based on a numerical variable (e.g. number of positive nodes), then a trend test based on the log-rank is available.
    - The log-rank test only uses information about the **ranks** of the survival times rather than their actual values.

Ch 6.1 Survival Data

6.1.2 Hazard Function

- The **hazard function** $h(t)$ is the short-term event rate for subjects who have not yet experienced the outcome event.
    - The hazard function is systematically related to both the survival and cumulative event functions.

6.1.3 Hazard Ratio

- The hazard ratio:

$$HR(t) = h_c(t)/h_1(t)$$

where $h_c(t)$ is the hazard function in the cadaveric group, $h_1(t)$ is the corresponding hazard function in the reference group.

6.1.4 Proportional Hazards Assumption

- Under the **proportional hazards assumption**, the hazard ratio does not vary with time. That is, $HR(t) \equiv HR$.
    - This is not necessary for the Cox proportional hazards model. We can generalize the model by including an interaction between the predictor and time; this allows the hazard ratio for that predictor to change with time. This is implemented using **time dependent covariates (TDCs)**.

Ch 6.2 Cox Proportional Hazards Model

6.2.1 Proportional Hazards Models

- In the linear model for continuous outcomes. the linear predictor $\beta_1 x_1 + \ldots + \beta_p x_p$ is linked directly to the conditional mean of the outcome, $E[y|x]$:

$$E[y|x] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

- In the logistic model for binary outcomes, the linear predictor is linked to the conditional mean through the logit transformation:

$$log\frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

  where $p(x) = E[y|x]$ is the probability of the outcome event for an observation with predictor values $x = (x_1, \ldots, x_p)$.
- In proportional hazards regression models, the linear predictor is linked through the log-transformation to the hazard ratio. If the hazard ratio obeys the proportional hazards assumption:

$$log[HR(x)] = log\frac{h(t|x)}{h_0(t)} = \beta_1 x_1 + \ldots + \beta_p x_p$$

where $h(t|x)$ is the hazard at time t for an observation with covariate value x, and $h_0(t)$ is the **baseline hazard function**, defined as the hazard at time t for observations with all predictors equal to zero.

- Baseline hazard does not apply to any possible observation, and argues for centering continuous predictors.
- For $h(t|x)$ gives:

$$h(t|x) = h_0(t)exp(\beta_1 x_1 + \ldots + \beta_p x_p) = h_0(t)HR(x)$$

  - This defines a **multiplicative model**, in the sense that the predictor effects act to multiply the baseline hazard.
  - Thus $HR(x)$ cannot be negative.
- Taking the log of both sides:

$$log[h(t|x)] = log[h_0(t)] + \beta_1 x_1 + \ldots + \beta_p x_p.$$

  - This defines a **log-linear model**, which implies that the log of the hazard is assumed to change linearly with any continuous predictors.
    - This shows log baseline hazard plays the role of the intercept in other regression models, though in this case it can change over time.

6.2.2 Parametric Versus Semi-parametric Models

- Two options in dealing with the baseline hazard $h_0(t)$:
  - Model it with a parametric function:
    - Examples: exponential survival model specifies that the hazard is a constant, while the Weibull regression model has a hazard which is a polynomial in time.
    - The baseline $h_0(t)$ is specified by a small number of additional parameters
    - However, the adequacy of the model for the baseline hazard has to be checked.
  - Semi-parametric model:
    - Example: Cox model, or Cox proportional hazards model does not require us to specify a parametric form for the baseline hazard $h_0(t)$.
    - Estimation of the regression parameters $\beta_1, \beta_2, \ldots, \beta_p$ is done without having to estimate the baseline hazard function.
    - The Cox model is more robust than parametric proportional hazards models because it is not vulnerable to misspecification of the baseline hazard.

6.2.3 Hazard Ratios, Risk, and Survival Times

- Interpretation of Cox model:
  - For predictors with hazard ratios less than 1 ($\beta < 0$), increasing values of the predictors are associated with lower risk and longer survival times. When hazard ratios are greater than 1 ($\beta > 0$), increasing values of the predictor are associated with increased risk and shorter survival times.
  - e.g. The hazard ratio for treatment, 0.82, means that estimated short-term mortality risk among patients assigned to this remedy was 82% of the risk in the placebo group. This ratio is assumed to be constant over the 10 years of follow-up.
  - e.g. The hazard ratio for medicine levels,1.16, means that for each quantity increase in medicine, short-term risk is increased by a factor of 1.16.

- A model with predictors $x_1, x_2, \ldots, x_p$, coefficient β_j is the increase in the log-hazard ratio for a one-unit increase in predictor x_j, holding the values of the other predictors constant. It follows that $exp(\beta_j)$ is the hazard ratio for a one-unit increase in x_j.

$$log[HR(x)] = log\frac{h(t|x)}{h_0(t)} = \beta_1 x_1 + \ldots + \beta_p x_p$$

- Note: The definition of the hazard is a short-term rate, and risk in this sense is from cumulative risk over a defined follow-up period.

6.2.4 Hypothesis Tests and Confidence Intervals

- In the Cox model, the estimated coefficients have an approximate **normal distribution** when there are adequate numbers of events in the sample. The normal approximation is better for the **coefficient estimates** than for the hazard ratios, so hypothesis tests and confidence intervals are based on calculations involving the coefficients and their standard errors.

  (If there are fewer than 15-25 events, the normal approximation is suspect and bootstrap CIs may work better)

  - For each predictor in the model, **Wald Z-tests** can be used to test the null hypothesis $H_0 : \beta = 0$, or the hazard ratio equals 1. (There is no association between the predictor and the risk.)

    $Z \sim N(0, 1)$.

  - A 95% CI for each β : $\hat{\beta} \pm 1.96SE(\hat{\beta})$, while CIs for hazard ratios are $exp(\hat{\beta} \pm 1.96SE(\hat{\beta}))$. The two intervals are usually very similar.

  - Interpretation:

    e.g. the data are consistent with risk reductions as large as 43%, but also with risk increases of 17% (95%CI for the hazard ratio from 0.57 to 1.17)

- Since Cox regression is a **likelihood-based** method, tests for predictors can also be obtained using the **likelihood ratio (LR) tests** for the logistic regression model.
  - Compare twice the difference in log-likelihoods for nested models to a χ^2 distribution with degrees of freedom equal to the between-model difference in the number of parameters.
- The Wald and LR results are similar in most situations but not exactly the same. They will be close when the sample size is large or the estimated hazard ratio is near 1.

6.2.5 Binary Predictors

- 0/1 coding recommended.

6.2.6 Multilevel Categorical Predictors

- One group serves as the reference category. We obtain estimated hazard ratios for other groups with respect to the reference group.
- Or pairwise comparisons between any two categories.

6.2.6.1 Categories with No Events

- Problem: Hazard ratios with respect to a reference category with no events are infinite, and the accompanying hypothesis tests and CIs are hard to interpret.
- Solution: select an alternative reference group.

6.2.6.2 Global Hypothesis Tests

- Global hypothesis tests for the overall effect of a multilevel categorical predictor can be conducted using **Wald** or **likelihood ratio (LR)** χ^2 **tests**, with degrees of freedom equal to the number of categories minus 1.
- A **log-rank test** will show different groups do not have equal survival.
- The statistical significance of **pairwise comparisons** should be interpreted with caution.
- With a large number of categories, multiple comparisons can lead to inflation of the **familywise type-I error rate (FER)**.
- Some comparisons may lack power due to small numbers in either of the categories being compared.

6.2.6.3 Ordinal Predictors and Tests for Trend

- Question: if the histology score is ordinal, we can ask: does the log mortality hazard increase linearly with higher histology ratings?
- Solution: **Tests for tread across categories**. For Cox model, these linear trend tests assess log-linearity of the hazard ratios.
    - check: whether the linear trend adequately captures the pattern of the coefficients across categories, or whether there are also important departures from this trends.
    - Use both categorical and log-linear terms
    - A Wald test for the **joint effect of the categorical terms** can be used to assess the departure from log-linearity. Large p-value means a linear trend across categories is an adequate description of the association between histology score and mortality risk.

*4.3.5 Testing for Trend Across Categories

- A **contrast** is a weighted sum of the regression coefficients of the form $\alpha_1\beta_1 + \alpha_2\beta_2 + \ldots + \alpha_p\beta_p$ in which the weights, or **contrast coefficients**, sum to zero: that is, $\alpha_1 + \alpha_2 + \ldots + \alpha_p = 0$
- Patterns of resulting contrast coefficients:
    - **Integer-valued**

      Using integers is just a convenience.
    - **Evenly spaced**

      Underlying the even spacing is the assumption that the "distance" between adjacent categories are all the same.
    - **Symmetric about zero**

      This implies that they also sum to zero, as required.
- If the number of levels is odd, the contrast coefficients are sequential integers with spacing of 1, and the middle category has coefficient 0 and drops out. If the number of level is even, a spacing of 2 is the smallest that gives integer-valued contrast coefficients, none of them are omitted.
- Little p-value and trend of coefficient demonstrate there is a relationship between predictors and outcome.
- Some details:
    - β_0 does not figure in the contrasts, so does effect of any adjustment variables held constant by the model.
    - drop reference category from contrasts.
    - Using the categorical version of the model in conjunction with contrasts to test for the trend can be more efficient when there is both trend and departure from it.

Number of categories	Linear contrast
3	$\beta_3 = 0$
4	$-\beta_2 + \beta_3 + 3\beta_4 = 0$
5	$-\beta_2 + \beta_4 + 2\beta_5 = 0$
6	$-3\beta_2 - \beta_3 + \beta_4 + 3\beta_5 + 5\beta_6 = 0$

- To test both a linear trend and a departure from trend, one method is to use a model in which the categorical variable is treated both as continuous and categorical.
    - The continuous version accounts for the trend, while the categorical version captures departure from it.
    - In F-test, large p-value of "Joint" (all joint categories) shows little evidence for departures from a linear trend.
    - This model is only useful for testing from departure from trend. Neither the coefficient nor the t-test for the effect of categorical predictors as continuous is interpretable.

6.2.7 Continuous Predictors

- The hazard ratio for continuous predictors is affected by the scale of measurement.
    - e.g. if the ages range from 26 to 78, a 1-year difference in age is small compared to the range of values. A 5-year increase in age might provide a more clinically interpretable result.
- The ratio of the hazards for any two patients who differ in age by k years - that is, for a patient at age $x + k$ compared with another at age x:

$$\begin{aligned} &\frac{h_0(t)exp(\beta(x+k))}{h_0(t)exp(\beta(x))} \\ &= \frac{exp(\beta(x+k))}{exp(\beta(x))} \\ &= exp(\beta(x+k) - \beta x) \\ &= exp(\beta k). \end{aligned}$$

    - Thus a k-unit change in a predictor multiplies the hazard by $exp(\beta k)$.
    - $\hat{\beta} = log(HR)$, for 5 years: $exp(\hat{\beta}5) = [exp(\beta)]^5$
- Changes in the scale of a continuous variable do not affect Wald and LR test.
- Hazard ratio can be interpreted in terms of percent changes in risk. A k-unit increase in the predictor implies a $100(exp\hat{\beta}k - 1)\%$ change in risk.
    - e.g. If HR = 1.04, we can say that estimated mortality risk among patients increases about 4% for every year increase in age, 22% for the increase in mortality risk associated with a 5-year increase in age.

6.2.8 Confounding

- The **regression adjustment** controls confounding in the Cox model. The attenuation of the unadjusted hazard ratio for one predictor in the adjusted model is typical of **confounding**.
    - e.g. we examine the association between bilirubin levels and survival among patients in the DPCA trial. Patients with higher bilirubin may also be more likely to have edema or spiders, other signs of liver damage which are correlated with elevated bilirubin levels but not mediators of its effects, and all associated with higher mortality risk.

    - The adjusted HR represents the effect of a one-unit change in bilirubin while holding edema and spiders constant

***4.4 Confounding**

- The underlying premise of multi-predictor regression analysis of observational data is that with a sufficiently refined model (and good enough data), we can estimate causal effect, free or almost free of confounding.
- To assess confounding we first need a **hypothesized causal framework**: The potential confounder should be plausible as a cause of both the predictor of interest and the outcome, or as a proxy for such as cause. Within this framework, the data provide support for confounding if :
    - The potential confounder is associated with the predictor of interest, and also independently associated with the outcome.
    - The coefficient for the effect of the primary predictor on the outcome changes when we add the potential confounder to the model.
- Confounders often explain some of the association of a predictor of interest with the outcome, so that the adjusted effect, which may have a causal interpretation, is often weaker than the unadjusted effect.
    - In some extreme, the effect of a factor of interest may be completely confounded by a second variable.
    - At the other extreme, we may find little or no association in unadjusted analysis, because it is **masked** or **negatively confounded** by another predictor.
        - Negative confounding can occur under the following circumstances:
            - The predictors are inversely correlated, but have regression coefficients with the same sign
            - The two predictors are positively correlated, but have regression coefficients with the opposite sign.
        - In the adjusted model, the addition of negative confounder increases the coefficient of associated predictor.
- Confounding is difficult to rule out. For the multi-predictor linear model to control confounding successfully and estimate causal effects without bias, all potential confounders must have been:
    - Recognized and assessed by design in the study (hard to achieve)
    - Measured without error (hard to achieve)
    - Accurately represented in the systematic part of the model (cannot be taken for granted)
- Uncontrolled confounding induces bias in unadjusted (or inadequately adjusted) estimates of the causal effects, while adjusted estimate induces les bias.

6.2.9 Mediation

- We establish mediation by requiring an association between the predictor of interest with the mediator and the outcome. The statistical test to establish mediation requires that we test each of these associations at the 0.05 level. Both null hypotheses are rejected with $p < 0.01$.

***4.5 Mediation**

- If the primary predictor is a cause of one of the covariates, which in turn affects the outcome, this would be an instance of **mediation**.

    - e.g. statin drugs reduce low-density LDL cholesterol level, which in turn appear to reduce risk of heart attack; in this model, reductions in LDL mediate the protective effect of statins.
- A potential mediator must make sense in terms of a **hypothetical causal framework**. Within this framework, the data support mediation if we find that:
    - The potential mediator is associated with predictor of interest and with the outcome, controlling for the predictor of interest.
    - The coefficient for the effect of the primary predictor on the outcome changes when we add the potential mediator to the model.
- Mediators behave like confounders in regression models, and can only be distinguished by the hypothesized causal framework.
- Indirect effects via the mediator
    - The effect of the primary predictor on the mediator, and of the mediator on the outcome, together comprise the hypothesized **indirect causal pathway** via the mediator.
    - If the model used to estimate these effects adequately control confounding of both relationships, then the two effects may together have a causal interpretation as the **indirect effect** of the primary predictor.
    - The overall null hypothesis of no indirect effect is rejected only if both underlying null hypotheses are rejected at the nominal α level.
- If the indirect pathway exists, and confounding has been controlled, then the coefficient for the primary predictor before adjustment for the mediator has a causal interpretation as the **overall effect** of the primary predictor on the outcome.
- The coefficient adjusted for the mediator is interpretable as the **direct effect** of the primary predictor via other pathways that do not involve the mediator.
- The **difference** between overall and direct effects of the primary predictor is interpretable as the **indirect effect**.
- Note: Tests for the difference between the overall and direct effects can give false-positive results, because of the collapsibility. In these models the coefficient for the primary predictor will generally change if a powerful predictor is added to the model, even if the new covariate is not associated with the primary predictor(no mediating role).

6.2.10 Interaction

- Example of interaction between two binary variables in the PBC data:

Group	rx	hepatom	$h(t\|X)$
1	Placebo	No	$h_0(t)$
2	DPCA	No	$h_0(t)exp(\beta_1)$
3	Placebo	Yes	$h_0(t)exp(\beta_2)$
4	DPCA	Yes	$h_0(t)exp(\beta_1+\beta_2+\beta_3)$ $=h_0(t)exp(\beta_1)exp(\beta_2)exp(\beta_3)$

    - The interaction hazard ratio $exp(\beta_3)$ gives the ratio of DPCA treatment effects among patients with and without hepatomegaly.
    - The ratio of the hazard for group 2 to the hazard for group 1 gives the effect of DPCA is the absence of hepatomegaly: $exp(\beta_1)$.

- The ratio of the hazard for group 4 to the hazard for group 3, gives the effect of DPCA in the presence of hepatomegaly:

$$\frac{h_0(t)exp(\beta_1)exp(\beta_2)exp(\beta_3)}{h_0(t)exp(\beta_2)} = exp(\beta_1)exp(\beta_3)$$

- Similar methods can be used to obtain estimates of the effect of hepatomegaly stratified by treatment assignment: that is, b comparing groups 3 and 1, then 4 and 2.

***3.7 Interpretation of Negative Findings**

- **Negative finding** - data do not enable us to reject a null hypothesis of interest
- A negative result worth discussing is best interpreted in terms of the **point estimate and CI**.

***4.6 Interaction**

- The assumption - the causal effect of the primary predictor was the same within strata defined by the covariates - may not always hold. We have to use regression to model the resulting interaction, so that we can estimate causal effects that differ according to the level of a covariate.
- **Interaction** is also referred to as **effect modification** or **moderation**, and must be distinguished from both confounding and mediation.

6.2.11 Model Building

- avoid including too many predictors for the number of observed events.
- at least ten events per predictor.

6.2.12 Adjusted Survival Curves for Comparing Groups

- Example: to examine the survival from the donor type(x_1), but there are 2 potentially important confounders of this effect: x_1 and x_2. After adjustment for these two factors, the HR for donor type is reduced.
  - Criteria: The adjusted curves for the two groups should differ only by x_1, with the other covariates being held constant.
  - Curve can be obtained using the coefficient estimates from the Cox model and an estimate of the baseline survival function, $\hat{S}_0(t)$, based on the Breslow baseline hazard estimate. The survival function by raising the baseline survival to the $exp(\hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p)$ power:

$$\{\hat{S}_0(t)\}^{exp(\hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p)}$$

  - It is conventional to use values for the adjustment variables which are close to the "center" of the data.
  - The differences between the survival curves are narrower after adjustment.

6.2.13 Predicted Survival for Specific Covariate Patterns

- $\{\hat{S}_0(t)\}^{exp(\hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p)}$ is useful for making predictions for specific covariate patterns.

Ch 6.3 Extension to the Cox Model

6.3.1 Time-Dependent Covariates

- A **time-dependent covariate(TDC)** in a Cox model is a predictor whose values may vary with time.

- **Immortal time bias** can be avoided (violation of **proportional hazards assumption** can be addressed) by using TDC.
- Example: a study of the effect of lung transplantation on survival in children, where the natural time origin in this study is the time of listing for transplantation, not transplantation itself.
    - lung transplantation has to be treated as a TDC.
    - Immortal time bias:
        - The survival time measured from listing forward will on average be longer in the transplanted group even if transplantation has no protective effect. Those children are selected for better prognosis, thus randomization assumption does not hold.
        - In the period of time before transplantation, they appear to be protected by a procedure that has not taken place. We can get into trouble by using information from the future to estimate current risk
    - **Indicator** of transplantation $X(t)$. In an unadjusted model, the hazard at time t:

$$\begin{aligned} h(t|x) &= h_0(t)exp\{\beta X(t)\} \\ &= \begin{cases} h_0(t) & \text{before transplantation} \\ h_0(t)exp(\beta) & \text{at or after transplantation} \end{cases} \end{aligned}$$

- Some more complicated cases:
    - Predictors will only be measured occasionally, but we need a value at each event time.
    - $X(t)$ often should be evaluated using all available information up until t.
    - Mediation can be evaluated using TDCs.
    - TDC confounds and mediates the effects of a time-dependent exposure or treatment.

6.3.2 Stratified Cox Model

- If the proportional hazards assumption does not hold for the binary or categorical predictors, we can accommodate the violation by fitting a **stratified Cox model** in which a separate baseline hazard is used.
- Example of stratification variable with two levels and a model with covariate x:

$$\begin{aligned} h(t|x=0) &= h_{00}(t) \\ h(t|x=1) &= h_{01}(t) \end{aligned}$$

- Generally to a stratification variable with two or more levels, and to a model with covariates:

$$h_{0j}(t)exp(\beta_1 x_1 + \ldots + \beta_p x_p)$$

    - we assume that the effect of each of the covariates is the same across strata

- No estimates, CIs, or p-values are obtained for the stratification variable, this approach is less useful for any predictor of direct interest.
- Because the resulting fit to the stratification variable is unrestricted, it can be used to rule out confounding of a predictor of interest by a covariate that violates the proportional hazards assumption.
- Stratification is also useful in the analysis of stratified randomized trials.

6.3.2.1 Number of Strata

- Stratification is a flexible approach to adjustment for a categorical variable even when it has a large number of levels.
- When the number of strata gets large, there can be some loss of efficiency in estimation of the treatment or other covariate effects.

- 5 or 6 strata generally suffice
- at least 5-7 events per stratum

6.3.2.2 Interaction Between Stratum and a Predictor of Interest

- Include a product term between the treatment and stratum indicators.
- Note that in the stratified model only the product term and the treatment indicator term are entered as predictors.

6.3.2.3 Stratified and Adjusted Survival Curves

- To obtain adjusted survival curves according to the levels of a stratification factor:
  - Example: compare the survival curves according to edema, adjusting for age. We have to specify a value for age - centering age on its mean of 50. Under the stratified Cox model, the survival function for a PBC subject with centered $agec$ is:

    $$[S_{0j}(t)]^{exp(\beta agec)}$$

    where edema (j = 1) and no edema (j = 0), therefore, $S_{01}(t)$ and $S_{00}(t)$.

Ch 6.4 Checking Model Assumptions and Fit

- Log-linearity
- Proportional hazards

6.4.1 Log-linearity of the Hazard Function

- Hazard ratio is log-linear in continuous predictors.
- Diagnostics for violations of log-linearity: attempt more general models and examine improvements in fit.
  - generalized by adding polynomial terms for the predictor in question
  - check effect sizes and p-values
- Categorize continuous predictor using well-chosen cut-point
  - check log-linearity by assessing both trend and departure from trend in ordinal predictors
  - limitation: susceptible to outliers for polynomial models, sensitive to the number and placement of the cut-points for categorization.
- Restricted cubic splines
  - A series of "knots" along the values of the predictor. The choice of the number of knots is a balance between the sample size and the degree of flexibility desired.
  - fit a polynomial curve
  - flexible: not greatly sensitive to number and placement of knots
  - A p-value alone should not used for model selection

6.4.2 Proportional Hazards

6.4.2.1 Log-Minus-Log Survival Plots

- An approach to diagnose the violation of proportional hazards assumption:

$$S_1(t) = [S_0(t)]^{exp(\beta)}$$

where $S_0(t)$ is the survival function for placebo patients and $S_1(t)$ is the corresponding survival function for the treated patients. When proportional hazards hold, the **log-minus-log transformation** gives:

$$log\{-log[S_1(t)]\} = \beta + log\{-log[S_0(t)]\}$$

where β is the log of the hazard ratio for treatment.

- This approach assumes a categorical variable but can be adapted to a continuous variable by, for example, factoring a continuous variable into quartiles.
- In log-minus-log plot, nonproportional hazards show patterns:
    - converging curves: difference between groups decreases with time
    - diverging curves: difference increases with time
    - crossing: may be harmful early but protective later

6.4.2.2 Smoothing the Hazard Ratio

- Address the violation: a nonparametric, smoothed estimate of the hazard ratio against time.
    - Works by smoothing a specialized type of residual, **scaled Schoenfeld residuals**.
    - A nearly constant smoothed estimate of the HR means assumption of proportional hazards is approximately satisfied.
    - Advantages:
        - Works for both categorical and continuous variables
        - influential points are identifiable from the plots of the Schoenfeld residuals
    - DFBETA statistics, a measure of how much coefficients are changed by the deletion of individual observations.

6.4.2.3 Schoenfeld Test

- A test for violation of proportional hazards, related to diagnostic plot using smooths of scaled Schoenfeld residuals. It assesses the correlation between the scaled Schoenfeld residuals and time.
    - If the HR is constant, the correlation should be zero.
    - Positive values of the correlation ρ suggest that the log-hazard ratio increases with time and vice versa.
    - Sensitive in cases where the log-hazard ratio is linearly increasing or decreasing with time. Sensitive to a few large residual values.

6.4.2.4 Graphical Diagnostics Versus Testing

- Limitation of Schoenfeld test:
    - may lack power to detect important violations in small samples
    - in large samples they may find statistically significant evidence of model violations which do not meaningfully change the conclusions
- **Graphical methods** give extra information about the magnitude and nature of model violation, and should be the first-line approach in examining the fit of model.

6.4.2.5 Stratification

See Ch 6.3.

- No assumption is made about the relationships between the stratified hazard functions specific to the different levels of the predictor.

6.4.2.6 Modeling Interactions with Time

See Ch 6.3.

- Example: Because the proportional hazards does not hold, the hazard ratio:

$$HR(t) = \frac{h_1(t)}{h_0(t)}$$

is a function of t. To address this, we define $\beta(t) = log\{HR(t)\}$ as a coefficient for edema which changes with time. The hazard function:

$$h(t|edema) = h_0(t)exp\{\beta(t)edema\}$$

where $edema$ is a 0/1 indicator.

  - model the log-hazard ratio for edema as a linear function of time using a main effect $edema$ and an interaction term $edemat$ (TDC).

$$\begin{aligned}\beta(t)edema &= (\beta_0 + \beta_1 t)edema \\ &= \beta_0\ edema + \beta_1 t\ edema \\ &= \beta_0\ edema + \beta_1\ edemat\end{aligned}$$

Or model the log-hazard ratio as linear in log time, defining the product term with $log(t)$ in place of t.

  - split follow-up time into sequential periods and model the log-hazard ratio for edema as a step function with a different value in each period.
    - Example: two TDCs

$$\beta(t)\ edema = \beta_1\ edema_1 + \beta_2\ edema_2$$

      - analogous to categorizing a continuous predictor to model nonlinear effects

Ch 6.5 Competing Risks Data

6.5.1 What Are Competing Risks Data?

- Previously: project forward the experience of a **censored observation** by representing their experience with those followed longer.
  - An assumption: **independent censoring**, that the future risk of those whose follow-up has ended can be represented by those who are followed longer.
  - An objective: to make an extrapolation to a setting in which the source of incomplete follow-up is eliminated.
  - Objectives and approaches to incomplete follow-up may differ depending on whether it is due to death or the inability of a study to retain or follow participants.
- **Competing risks data** arise when multiple events can occur and follow-up can end due to occurrence of one or more of those types of events, precluding observation of at least one of the other event types.
- In analysis of competing risks data, two major approaches can be taken:
  - **elimination**: seeks to extrapolate to a scenario in which a type of event is not possible
  - **accommodation**: acknowledges and allows for the competing risks in the analysis

6.5.2 Notation for Competing Risks Data

- Denote competing risks outcome data by two variable:
  - denotes the time of the first event:

    Y: be the time of the first observed event of any type.

- denotes the type of event:

  $\Delta = k$ if the kth event type occurs first.

  where each of K possible types of failure are denoted by a numerical code.

- This notation is standard for ordinary survival analysis where there are two possible events: censoring and failure.

6.5.3 Summaries for Competing Risk Data

- The extension of the hazard function to competing risks data is given by the **cause-specific hazard function**.
- The extension of the survival function is given by **cumulative incidence function**.

6.5.3.1 Cause-specific Hazard Functions

- The **cause-specific hazard function** for event type k, $h_k(t)$, is the short-term rate at which subjects experience the onset of the kth event among those who have not yet experienced the event of interest(e.g. fracture) or a competing event (e.g. death) prior to t.
- Estimating and modeling cause-specific hazard functions:
  - set up the data as ordinary survival data with the kth failure type as the only type of failure and treat competing causes even death as "censored".
  - does not distinguish those are accommodated with those are attempt to eliminate.

6.5.3.2 Cumulative Incidence Functions

- The **cumulative incidence function** for cause type k at time t, $F_k(t)$, is the proportion who have developed the kth event prior to t.
  - A measure of prevalence of a particular event at each time t for a population which started with none.
  - The cumulative incidence function for the kth event at time t_j, $F_k(t_j)$, equals:

    $$F_k(t_k) = F_k(t_{j-1}) + \tilde{S}(t_{j-1})h_k(t_j)$$

    where $\tilde{S}(t_{j-1})$ is the chance of being free of events at time t_{j-1}. The hazard $h_k(t_j)$ denotes the cause-specific hazard for the kth event at time t_j.
- Example: count the number of people who at 1 year were alive without fracture (n_0), had experienced a fracture(n_1), or had died without experiencing a fracture(n_2).
  - The cumulative incidence of fracture at 1 year would be $\frac{n_1}{n_0+n_1+n_2}$. The cumulative incidence function is the sum over all time precious.
  - The Kaplan-Meier method: $\frac{n_1+n_2}{n_0+n_1+n_2}$ (combining death and fracture into a single event).

6.5.3.3 Cox Model for Cause-Specific Hazard Functions

- One approach to allowing predictors to affect the onset of competing risks is to model the cause-specific hazard function using **proportional hazard formulation**
  - $h_k(t)$ is the kth cause-specific hazard and let $x_1, \ldots, x_p$ be a set of predictors. The model:

    $$h_k(t|x) = h_{0k}(t)exp(\beta_1 x_1 + \ldots + \beta_p x_p)$$

    where $h_{0k}(t)$ is a baseline hazard function.
  - Only difference: hazard ratios apply to the kth cause-specific hazard.

6.5.3.4 Fine-Gray model for Cause-Specific Hazard Functions

- The **Fine and Gray model** adapts the spirit of a proportional hazards model to the cumulative incidence formulation. It models a different kind of rate function. The kth cause-specific hazard function is the rate of event among the who have experienced no event.
    - The new rate function for the kth type of failure as $f_k(t)$.
    $$f_k(t|x) = f_{0k}(t)exp(\beta_1 x_1 + \ldots + \beta_p x_p)$$
    where $f_{0k}(t)$ takes the place of the usual baseline (or cause-specific) hazard function.

Ch 6.6 Some details

6.6.1 Bootstrap Confidence Intervals

6.6.2 Prediction

6.6.3 Adjusting for Non-confounding Covariates

6.6.4 Independent Censoring

See 3.5.1.

6.6.5 Interval Censoring

- The timing of many events is not observed exactly with precision. The actual time of an incident is only known up to an interval of possible values - **interval-censored**.
    - **interval-censored**: between the last visit at which the participant tested negative and the first at which the result was positive.
    - In settings where intervals arise because of the study follow-up schedule and are regularly spaced, **pooled logistic regression** can be used to handle the interval censoring.
    - More complicated: time between intervals is unequal.

6.6.6 Left-Truncation

- The setting where some survival times are not observed because the sampling scheme tends to miss short survival times is known as **left-truncation**.
    - To avoid bias, we need to consider the length of the period between the time origin and entry into the cohort.
    - We denote the **truncation time** by V.
    - Truncation time is the key feature, while staggered entry into a cohort does not imply left-truncation.
- **Right-truncation** can also arise if a study recruits based on an endpoint and people with large event times (or who never had the event) are not recruited.
- Survival analysis of risk factors can be conducted on the natural time-scale with origin at HIV infection under an assumption of **independent truncation**:
    - The assumption of **independent truncation**: the time of delayed entry and subsequent survival are independent and it is satisfied when the incidence of a disease and survival post-diagnosis are independent.
    - Under independent truncation, the analysis uses the truncation time V along with the time and censoring indicators (X, Δ), where X is the follow-up time relative to diagnosis.

- By ignoring the truncation, the estimator estimates higher post-diagnosis survival probabilities, because the analysis fails to account for under-sampling of short survival times and thus overestimates survival.
    - The effect of ignoring truncation on hazard ratios in a Cox model is less predictable but can often attenuate them.

Ch 6.7 Sample Size, Power, and Detectable Effects

- To compute the sample size that will provide power of $gamma$ in two sided tests with type I error of α to reject the null hypothesis $\beta_j = 0$ for the effect of a predictor X_j, accounting for the loss of precision due to adjustment for covariates:

$$n = \frac{(z_{1-\alpha/2} + z_\gamma)^2}{(\beta_j^\alpha \sigma_{x_j})^2 \psi (1-\rho_j^2)}$$

  where ψ is the probability that an observation is uncensored, so that the expected number of events $d = n\psi$.
- The variance inflation factor: $1/(1-\rho_j^2)$.
- For problems with fixed values of n and ψ, power is given by:

$$\gamma = 1 - \phi[z_{1-\alpha/2} - |\beta_j^\alpha| \sigma_{x_j} \sqrt{n\psi(1-\rho_j^2)}]$$

- minimum detectable effect (on the log-hazard scale):

$$\pm\beta_j^\alpha = \frac{z_{1-\alpha/2} + z_\gamma}{\sigma_{x_j}\sqrt{n\psi(1-\rho_j^2)}}$$

*4.8 Sample Size, Power, and Detectable Effects

- In linear model, we calculate the sample size that would provide power of γ to reject $\beta_j = 0$ in a two-sided test with type-I error rate α, under the alternative hypothesis $\beta_j = \beta_j^\alpha$, assuming for now that $\beta_j^\alpha > 0$.
    - An **expression for power**, relying on the large-sample equivalence of the t and standard normal Z-distribution:

$$\begin{aligned}
\gamma &= P[|\hat{\beta}_j|/SE(\hat{\beta}_j) > z_{1-\alpha/2}] \\
&\approx P[\hat{\beta}_j/SE(\hat{\beta}_j) > z_{1-\alpha/2}] \\
&= P[(\hat{\beta}_j - \beta_j^\alpha)/SE(\hat{\beta}_j) > z_{1-\alpha/2} - \beta_j^\alpha/SE(\hat{\beta}_j)] \\
&= 1 - \phi[z_{1-\alpha/2} - \beta_j^\alpha/SE(\hat{\beta}_j)] \\
&= \phi[\beta_j^\alpha/SE(\hat{\beta}_j) - z_{1-\alpha/2}]
\end{aligned}$$

    - Apply the **inverse transformation** ϕ^{-1} , with $Var(\hat{\beta}_j) = \frac{\sigma^2_{y|x}}{(n-1)\sigma^2_{x_j}(1-r_j^2)}$ (replace (n-1) with n). Solve for n gives:

$$n = \frac{(z_{1-\alpha/2} + z_\gamma)^2 \sigma^2_{y|x}}{(\beta_j^\alpha \sigma_{x_j})^2 (1-\rho_j^2)}$$

    - The **variance inflation factor** :$1/(1-\rho_j^2)$ accounts for the potential loss of precision due to the inclusion of other predictors in the model
    - If we specify β_j^α, with $Var(\hat{\beta}_j) = \frac{\sigma^2_{y|x}}{(n-1)\sigma^2_{x_j}(1-r_j^2)}$ (replace (n-1) with n), we can calculate power:

$$\begin{aligned}\gamma &= 1 - \phi[z_{1-\alpha/2} - \beta_j^\alpha / SE(\hat{\beta}_j)] \\ &= 1 - \phi[z_{1-\alpha/2} - |\beta_j^\alpha| \sigma_{x_j} \sqrt{n(1-\rho_j^2)} / \sigma_{y|x}]\end{aligned}$$

- The **minimum detectable effect**: the smallest value of β_j^α for which a sample of size n would provide power of γ to reject the null hypothesis $\beta_j = 0$ in a two-sided test with type-I error of α.

$$\pm\beta_j^\alpha = \frac{(z_{1-\alpha/2} + z_\gamma)\sigma_{y|x}}{\sigma_{x_j}\sqrt{n(1-\rho_j^2)}}$$

Reference

"Regression Methods in Biostatistics - Linear, Logistic, Survival, and Repeated Measures Models"

2nd edition by Eric Vittinghoff, David V. Glidden, Stephen C. Shiboski, Charles E. McCulloch.