Statistical Practices and Research for Interdisciplinary Sciences (**SPRIS**)

Lecture 6

Yuanjia Wang, Ph.D.

Department of Biostatistics, Mailman School of Public Health
Columbia University

& Division of Biostatistics, New York State Psychiatric Institute

THE DEPARTMENT OF **BIOSTATISTICS**   Columbia University MAILMAN SCHOOL OF PUBLIC HEALTH

Missing data occur in many studies

- ► Nonresponse in survey sampling.

- ► Structural missing in survey instruments.

- ► Dropout or noncompliance in clinical trials: Subjects are scheduled to return to the clinic weekly to provide outcome assessments. However, some subjects fail to show up for all clinic visits after a certain point (drop out of the study). Still others may miss clinic visits sporadically (assessment dropout) or quit taking their assigned treatment (treatment dropout).

- ► Missingness by design (e.g., nested case-control): to save cost, obtain expensive covariates (e.g., genetic polymorphisms) only on a subsample of subject (e.g., genotype all cases and a subsample of controls).

Objective: making inference about some aspect (parameter) of the distribution of the "full data" (i.e., the data that would have been observed if no data were missing).

Problem: When some of the data are missing, depending on how and why they are missing, our ability to make an accurate inference may be compromised.

Missing data can be categorized as:

- ▶ Missing completely at random (MCAR): data are missing for reasons that are unrelated to any characteristics or responses for the subject, including the value of the missing data, were it to be known (e.g., a subject missed a clinic visit due to accident)

- ▶ Missing at random (MAR): Data are not missing at random, but the probability of missing depends on values of observed variables (observed outcomes and covariates).

- ▶ Missing not at random (MNAR): missingness depends on the unobserved true values of the variable (e.g., subjects are more likely to be missing if their true values of the variable are systematically higher or lower) sensitivity analysis, choose a method based on MAR first; Propose a missing data model and then do MNAR analysis

MCAR can be tested and easiest to handle. MNAR most difficult and requires untestable assumptions. Focus on the analysis of MAR.

Introduction to Missing Data

Before any analyses: characterize patterns of missingness using exploratory data analysis. Has implications on the analyses (e.g., remove variables with substantial missing).

Visualization of missing data pattern[1]
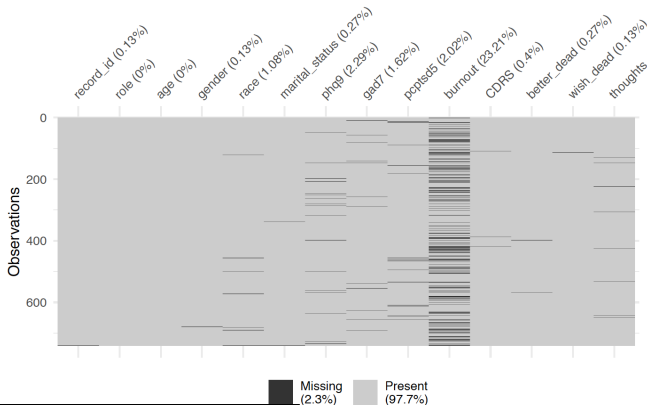
¹vis_mis() in visdat R package.

Strategies for Handling Missing Data

Statistical software packages use casewise deletion (listwise deletion) in handling missing predictors; i.e., any subject having any covariate or outcome missing will be excluded from a regression analysis.

Listwise deletion produces biased results unless MCAR is satisfied.

Even if MCAR is satisfied, it can be inefficient because a subject missing any of the predictors will be excluded from the analyses.

In the data example in Multiple Imputation Random Lasso (MIRL), listwise deletion will discard 81% of subjects.

Strategies for Handling Missing Data

Three main strategies:

- ▶ Likelihood based methods: express the full data likelihood in terms of observed data and unobserved data likelihood. Under the MAR assumption, the unobserved data likelihood does not provide information on parameters of interest. Use observed data likelihood for estimation and inference (e.g., LME).

- ▶ Imputation methods: valid under MAR and assumptions of imputation model when correctly accounting for variability due to imputation.

- ▶ Inverse probability weighting (IPW) of complete cases: valid under MAR and correct specification of model for missingness

$$\hat{\mu} = \frac{1}{n} \sum_i \frac{R_i Y_i}{P(R_i = 1 | X_i)},$$

where Y_i is the outcome, R_i is the indicator for observing an outcome, X_i are predictors of likelihood of missing.

Missing Data in Longitudinal Studies

Generalized estimating equations (GEE) is based on a working independence model: ordinary univariate regressions are fitted on a combined dataset as if all observations are uncorrelated. Then an after-the-fit correction for intra-cluster correlation is done using the robust sandwich covariance estimator.

GEE only need the mean model is correct

GEE assumes missing response values are MCAR. When it doesn't hold, under MAR, need to use IPW weighted GEE.

Linear mixed effects model and likelihood-based inference is valid under MAR: use each subject's observed trend as a model for "imputation". LME can provide robust sandwich covariance estimator as well.

Some practical guidelines:

► Missing covariates: multiple imputation

► Missing outcomes : inverse probability weighting (IPW), likelihood based approaches (LME)

If a covariate is unrelated to all other X's, the mean or median may be substituted for missing values without much loss of efficiency.

When the variable of interest is related to the other X's, more efficient to use an individual predictive model for each X based on the other variables, in turn. For example, use software MICE: Multiple Imputation by Chained Equations; SAS PROC MI

Tree-based models (e.g., random forest) can be used for imputation and they do not require linearity or additivity assumptions.

Multiple Imputation

Imputing missing values and then doing an ordinary analysis as if the imputed values were observed is usually better than excluding subjects with incomplete data. However, standard errors and other statistics need to take imputation into account.

Multiple imputation uses random draws from the conditional distribution of the target variable given the other variables.

Imputation is repeated M times, $M \geq 3$. Each repetition results in a "completed" dataset that is analyzed using the standard method. Parameter estimates are averaged. Variance-covariance formulae:

$$V = \frac{1}{M} \sum_{i=1}^{M} V_i + \frac{M+1}{M} B,$$

V_i the complete data estimate of var-cov matrix, B the between-imputation sample var-cov matrix.

Summary and Practical Guidelines

▶ Investigate reasons for missing (consider plausibility of MCAR, MAR and MNAR assumptions)

▶ Missing proportion $\leq 5\%$: It doesn't matter very much how we impute missings or whether we adjust variance of regression coefficient estimates for having imputed data

▶ Missing proportion between 5% and 15%: Mean or median imputation if X is not related to other predictors; Multiple imputation using regression; Consider classification tree for categorical variables; Single imputation might perform well

▶ Missing proportion $> 15\%$: more important to adjust for variability of imputation by multiple imputation

▶ Multiple predictors with frequently missing: effects of imputations are more pronounced.

▶ Consider sensitivity analysis

Strategies for Sensitivity Analyses

If different analyses all suggest same result, can be more comfortable with conclusions

- ▶ Worse-case scenario imputation in RCT: simplest sensitivity analysis
    - ▶ Assume all on investigational drug were treatment failures, all on control group were successes
    - ▶ If drug still appears significantly better than control, even under this extreme assumption, very robust effect
    - ▶ Note: conservative
- ▶ Completers analyses as another sensitivity
- ▶ MNAR: pattern-mixture model (include patterns of missingness as covariates in the model) or selection model. See Tutorial on Missing Data (Hogan et al. 2004).
- ▶ For categorical variables, code missing values as a separate category

Example: HEAL Analysis Plan

Non-normal outcomes (categorical, time-to-event)

Generalized Linear Models (GLM)

In the linear model, $Y_i = \mu_i + \varepsilon_i$, $\mu_i = E(Y_i|\boldsymbol{X}_i) = \boldsymbol{X}_i^T\boldsymbol{\beta}$ is assumed to be a linear function of the covariates, and $\varepsilon_i \sim N(0, \sigma^2)$.

As an extension to non-normal outcomes, generalized linear models assume

- Link function: $\eta_i = g(\mu_i) = \boldsymbol{X}_i^T\boldsymbol{\beta}$. Link functions are central ideas of generalized linear models. It is used to link the linear predictor to the conditional expectation of the response in a wider class of models.

- Random component usually follows the exponential family with density function

$$f(y|\theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right].$$

Normal, Poisson, Bernoulli, binomial, Gamma distributions are important members of this class. Poisson and binomial, $\phi = 1$; Normal and Gamma ϕ is a free parameter

Examples of Members in the Exponential Family

1. Normal:

$$
\begin{aligned}
f(y|\theta, \phi) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \\
&= \exp\left\{\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right\}
\end{aligned}
$$

so $\theta = \mu$, $\phi = \sigma^2$, $a(\phi) = \phi$, $b(\theta) = \theta^2/2$, and
$c(y, \phi) = -(y^2/\phi + \log(2\pi\phi))/2$.

2. Poisson:

$$
f(y|\theta, \phi) = e^{-\mu}\mu^y/y! = \exp(y\log\mu - \mu - \log y!)
$$

so $\theta = \log(\mu)$, $\phi \equiv 1$, $a(\phi) = 1$, $b(\theta) = \exp(\theta)$, and
$c(y, \phi) = -\log y!$.

Properties

Properties of exponential families:

1. Mean: $E(Y) = \mu = b'(\theta)$
2. Variance: $\text{var}(Y) = b''(\theta)a(\phi)$

θ: canonical parameter, ϕ: dispersion parameter, $b''(\theta)$ variance function to describe how the variance relates to the mean.

Gaussian: $b''(\theta) = 1$, variance independent of the mean.

We can introduce weights $a(\phi) = \phi/w$, w is known and varies across subjects.

Canonical links: $g(\cdot)$ is such that $\theta = g(\mu)$, the canonical parameter of the exponential family. Canonical link function has good interpretation, mathematically convenient.

Why is Exponential Family Important?

- ▶ At the core of generalized linear models (GLM).
- ▶ The only family that has conjugate priors to facilitate Bayesian inference (computing posterior).
- ▶ At the core of variational inference (with a conjugate prior, the best approximating posterior distribution is in the same family).
- ▶ Under conditions, the only family that has finite-sized sufficient statistics.
- ▶ Can be used as building blocks to compose more complicated distributions (e.g., mixture Gaussian, beta-binomial).

Fitting GLM by MLE

▶ The MLE maximizes the log-likelihood and is the solution to the score equation (estimating equation) $S(\beta)$, i.e.,

$$S(\widehat{\beta}) = \mathbf{0},$$

$$S(\beta) = \sum_i \frac{X_i h'(\eta_i)}{\sigma_i^2}(Y_i - \mu_i) = X\mathcal{D}\Sigma^{-1}(Y - \mu),$$

where $h = g^{-1}$, $\sigma_i^2 = var(Y_i) = \phi b''(\eta_i)$, $\Sigma = \text{diag}(\sigma_1^2, \cdots, \sigma_n^2)$, and $\mathcal{D} = \text{diag}(h'(\eta_1), \cdots, h'(\eta_n))$.

▶ Fisher information matrix

$$F(\beta) = X^T W X, W = \text{diag}(w_1, \cdots, w_n), w_i = \frac{(h'(\eta_i))^2}{\sigma_i^2}$$

▶ Numeric computation by iteratively reweighted least squares

$$\beta^{(t+1)} = (X^T W^{(t)} X)^{-1} X^T W^{(t)} \tilde{Y}^{(t)},$$

$$\tilde{Y}^{(t)} = X^T \beta^{(t)} + \mathcal{D}^{(t)}(Y - \mu(\beta^{(t)})$$

Measuring Model Goodness of Fit

The fitted model leads to $\hat{\mu}_i = g^{-1}(\hat{\eta}_i) = g^{-1}(\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}})$. The discrepancy between Y_i and $\hat{\mu}_i$ measures the fit of the model.

Fit the model under the full model and fitted model. Define discrepancy in terms of likelihood ratio, or the scaled deviance

$$D(y, \hat{\mu})/\phi = 2[l(y, \phi) - l(\hat{\mu}, \phi)],$$

where $l(y, \phi)$ is the log-likelihood for the full (saturated) model. Deviance:

- Normal: deviance is $\sum_i (y_i - \hat{\mu}_i)^2$
- Poisson: $2 \sum_i [y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)]$
- Bionmial: $2 \sum_i [y_i \log(y_i/\hat{\mu}_i) + (n - y_i) \log((n - y_i)/(n - \hat{\mu}_i))]$

Tests of comparing different models are based on χ^2 distribution (df=number of parameters when the model is correct).

Large Deviance

For binomial or Poisson GLM (overdispersion parameter $\phi = 1$), if model specification is correct, we expect that the residual deviance will be approximately distributed χ^2 with the appropriate degrees of freedom. If observe a deviance that is much larger than would be expected, need to determine which aspect of the model specification is incorrect.

Potential reasons for large deviances:

- Fail to include the right predictors or have not transformed them in the correct way (diagnostics can be performed based on "residuals" versus predictors or partial residuals versus predictors)

- Presence of a small number of outliers

- Sparse data (few subjects in each covariate class, distribution of deviance is not chi-squared)

- Overdispersion

Overdispersion

For Binomial case, model specifies that within groups defined by covariates, the variance is $var(Y_k) = np(1 - p)$. In other words, the variance is determined by mean (i.e., p). If there is unexplained heterogeniety, the variance will be greater than specified by the Binomial model. This is referred as overdispersion, usually due to heterogeneity or dependence among subjects.

To adjust for overdispersion to compute the variance, introduce an additional dispersion parameter and estimate from data:

$$\hat{\phi} = \frac{\chi^2}{n - p}, \quad \chi^2 : \text{Pearson chi-square statistic.}$$

This dispersion parameter method is only appropriate when the covariate classes are roughly equal in size.

model p(R = 1| X)
model log(Y|X)

If not, more sophisticated methods such as the beta-binomial distribution where we assume that p follows a beta distribution. For Poisson regression, consider zero inflated Poisson, two parts model.

Similar to linear models, use residual plots $\hat{\eta}$ versus r_{SD}, where r_{SD} is developed for the GLM. Examine:

- ▶ Choice of link
- ▶ Choice of scale of covariate (need transformation)
- ▶ Omission of quadratic trend in X.
- ▶ Most general model: generalized additive models, e.g., $\eta(X) = f_1(X_1) + f_2(X_2) + \cdots + f_k(X_k)$ where f_k are nonparametric functions.

Plot r_{SD} versus covariates also useful as in linear models.

Ordinal Logistic Regression

Ordinal response variable is common in medical and epidemiologic studies (e.g., severity of a symptom related to a disease: none, mild, moderate, severe).

Assumption of all commonly used ordinal regression models: the response variable behaves in an ordinal fashion with respect to each predictor; assume a predictor X is linearly related to the log-odds of levels of Y.

Check ordinality assumption: plot mean of X against level of Y; they should be in consistent order.

Example: cumulative logistic regression under proportional odds (PO) assumption for $Y = 0, 1, 2, \cdots, k$

$$P(Y \geq j | \boldsymbol{X}) = \frac{1}{1 + \exp[\alpha_j + \boldsymbol{X}^T \boldsymbol{\beta}]}, j = 1, \cdots, k$$

prevalence can be different, odds ratios can not

Assumption: the regression coefficient is independent of j, the cutoff level of Y.

Ordinal Logistic Regression

Interpretation of regression coefficient in PO model

$$P(Y \geq j | X) = \frac{1}{1 + \exp[\alpha_j + \boldsymbol{X}^T \boldsymbol{\beta}]}.$$

- ▶ Interpretation using odds ratios just like for usual logistic regression

- ▶ Difference is that a single odds ratio is used to estimate all events $Y \geq j, j = 1, \cdots, C$.

- ▶ If linearity holds in terms of the kth predictor X_k, β_k is interpreted as the log odds ratio comparing groups $X_k + 1$ versus X_k of the event $Y \geq j$, whatever the cutoff j.

- ▶ α_j is the log-odds of the event $Y \geq j$ in the group with $\boldsymbol{X} = \boldsymbol{0}$.

Continuation Ratio Model

Unlike the PO model, which is based on cumulative probabilities, the continuation ratio (CR) model is based on conditional probabilities:

$$P(Y = j | Y \geq j, \boldsymbol{X}) = \frac{1}{1 + \exp[-(\theta_j + \boldsymbol{X}^T \boldsymbol{\gamma})]}$$

The CR model: fits ordinal responses when subjects have to "pass through" one category to get to the next.

$$\text{logit}[P(Y = 0 | Y \geq 0, \boldsymbol{X})] = \text{logit}[P(Y = 0 | \boldsymbol{X})] = \theta_0 + \boldsymbol{X}^T \boldsymbol{\gamma}$$

$$\text{logit}[P(Y = 1 | Y \geq 1, \boldsymbol{X})] = \theta_1 + \boldsymbol{X}^T \boldsymbol{\gamma}$$

$$\cdots$$

$$\text{logit}[P(Y = C - 1 | Y \geq C - 1, \boldsymbol{X})] = \theta_{C-1} + \boldsymbol{X}^T \boldsymbol{\gamma}$$

The CR model is a discrete version of the Cox proportional hazards model. The discrete hazard function is defined as $P(Y = j | Y \geq j)$.

Continuation Ratio Model

Assumption of the CR model: the vector of regression coefficients, $\boldsymbol{\gamma}$, is the same regardless of which conditional probability is being computed.

Interpretation: β_k is the log odds ratio of $Y = j$ given that $Y \geq j$ comparing group with $X_k + 1$ versus group X_k, whatever level j is.

To obtain unconditional probabilities, multiple conditional probabilities (similar to how to obtain survival function for Cox model)

$$
\begin{aligned}
P(Y > 1 | \boldsymbol{X}) &= P(Y > 1 | Y \geq 1, \boldsymbol{X}) \times P(Y \geq 1 | \boldsymbol{X}) \\
&= [1 - P(Y = 1 | Y \geq 1, \boldsymbol{X})] \times [1 - P(Y = 0 | \boldsymbol{X})],
\end{aligned}
$$

where

$$
P(Y = 1 | Y \geq 1, \boldsymbol{X}) = \frac{\exp(\theta_1 + \boldsymbol{X}^T \boldsymbol{\gamma})}{1 + \exp(\theta_1 + \boldsymbol{X}^T \boldsymbol{\gamma})}, P(Y = 0 | \boldsymbol{X}) = \frac{\exp(\theta_0 + \boldsymbol{X}^T \boldsymbol{\gamma})}{1 + \exp(\theta_0 + \boldsymbol{X}^T \boldsymbol{\gamma})}
$$

can be estimated based on model parameters.

Conclusions

Consider PO and CR models for parsimonious estimation of ordinal outcomes. If both fail to adequately fit data, consider multinomial model.

For ordinal regression modeling, the modeling steps are (1) choice of predictor variables (consider interaction), (2) selecting or modeling predictor transformations (nonlinear trend), and (3) allowing for unequal effects across Y-cutoffs (i.e., non-PO or non-CR).

A model can have significant lack of fit with respect to some of the predictors and still yield quite accurate predictions.

Time-to-event Data

Time-to-event outcomes are often collected in medical studies (e.g., time-to-disease onset, time-to-death, time to hospital readmission, time to recovery from disease).

Why need new set of models/analyses methods for time-to-event data?

▶ Time to event can have an unusual distribution. Event time is restricted to be positive so it has a skewed distribution and will never be normally distributed.

▶ Predicting the entire survival curve.

▶ The probability of surviving past a certain time is often more relevant than the expected survival time

▶ Expected survival time may be difficult to estimate if censoring is high.

Introduction to the Analysis of Time-to-event Data

- Each patient with covariates X has its own survival curve
- Patients in group A are universally worse than B at any time point
- Group C worse than A, B at some time points, better other time
- Can compare mean survival time (area under the curve) and median survival time (time $S(t|X)$ crosses 0.5).
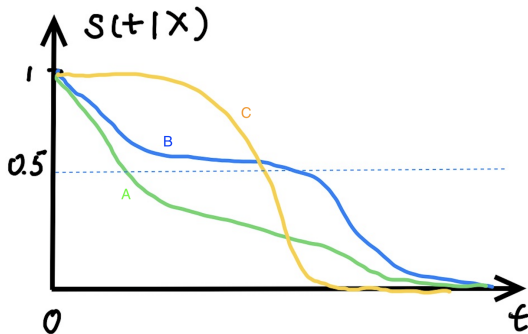
Censoring and Truncation

- ▶ Right censoring: An event occurs beyond a certain time ($T > C$, where T is event time, C is censoring time)

- ▶ Left censoring: An event is known to have occurred before a certain time ($T < C$)

- ▶ Interval censoring: the event is known to occur in an interval of time ($T \in [C_L, C_R]$); occurs when a medical condition is assessed during periodic exams.

- ▶ Left-truncation: An unknown subset of subjects failed before a certain time and the subjects didn't get into the study.

Assumptions on censoring mechanisms:

- ▶ Noninformative censoring: the censoring is caused by factors that are independent of the event of interest; can be relaxed as censoring is conditionally independent of events given covariates

- ▶ Informative censoring will cause biased estimates unless properly considered

- Survival function: $P(T > t)$

- Hazard function: instantaneous event (death, failure) rate. The hazard at time t is the probability that the event will occur in a small interval around t, given that the event has not occurred before time t, $\lambda(t) = \lim_{dt \to 0} P(T \in (t, t + dt) | T > t) = f(t)/S(t)$. Cumulative hazard is $\Lambda(t) = -\log S(t)$.

- Observed response is $Y = \min(T, C)$

T_i	C_i	Y_i	e_i
75	81	75	1
7	76	7	1
68+	68	68	0
52+	52	52	0
20	56	20	1

Termination of Study
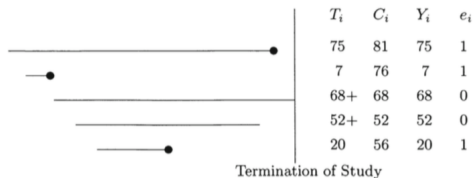
Fig. Censored Data; circles denote events

Survival Function and Hazard Function

Survival function → **Cumulative hazard function** → **Hazard function**

$$S(t|x) \xrightarrow[\text{Negative log}]{} -\log S(t|x) \xrightarrow[\text{Derivative w.r.t time}]{} -\frac{d}{dt}\log S(t|x) = \frac{f(t|x)}{S(t|x)}$$

$H(t|x)$

$h(t|x)$

Survival $S(t|x)$

$H(t|x)$ Cumulative hazard

$h(t|x)$ Hazard

1

0

t
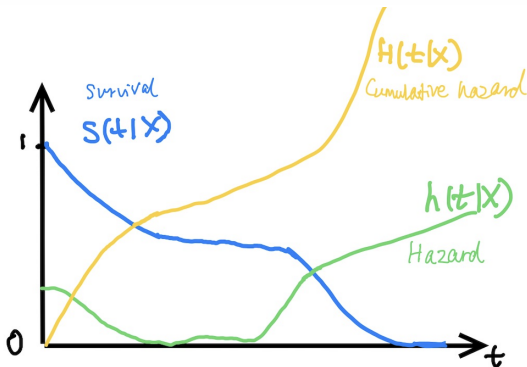
- Parametric models: Weibull, exponential, log-normal; estimation through MLE

- Nonparametric estimation of survival function: Kaplan-Meier

$$\Pr(\text{survive beyond time } t) \approx \prod_{j=1,\dots,L \text{ such that } t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

	t_1	t_2	t_3	t_4	$\cdots$	t_L
# people who fail	d_1	d_2	d_3	d_4	$\cdots$	d_L
# people at risk	n_1	n_2	n_3	n_4	$\cdots$	n_L
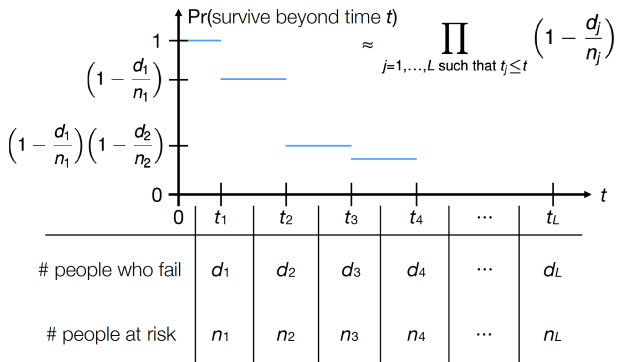
- Cumulative hazard function: Nelson-Aalen estimator
$H(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}$

KM does not account for covariates.

Regression models with covariates X:

- ▶ Proportional hazards (PH) model:

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t) \exp(\boldsymbol{X}^T \boldsymbol{\beta})$$

- ▶ $\lambda_0(t)$ is the baseline hazard function. Can be specified under parametric models. If baseline hazard function is left unspecified, the model is referred as Cox proportional hazards model.

PH model: $\log \lambda(t|\boldsymbol{X}) = \log \lambda_0(t) + \boldsymbol{X}^T \boldsymbol{\beta}$

Model assumptions:

- ▶ The relationship between the predictors and log hazard should be linear.

- ▶ Predictors affect the distribution of the response by multiplying the hazard or cumulative hazard by $\exp(\boldsymbol{X}^T \boldsymbol{\beta})$.

- ▶ The effect of the predictors is assumed to be the same at all values of t since $\log \lambda_0(t)$ can be separated from $\boldsymbol{X}^T \boldsymbol{\beta}$, i.e., the PH assumption implies no time by predictor interaction.

- ▶ β_k is the log hazard ratio comparing group $X_k + 1$ versus X_k

PH Model Assumptions and Interpretation

Effect of hazard ratio (HR) on distribution functions for small t,

$$F_2(t) \approx HR_{2|1}F_1(t),$$

$F_k(t) = P(T_k \leq t) = \lambda_k(t)S_k(t)$ is the cumulative distribution function, $HR_{2|1}$ is the hazard ratio, that is, $\lambda_2(t) = HR_{2|1}\lambda_1(t)$.
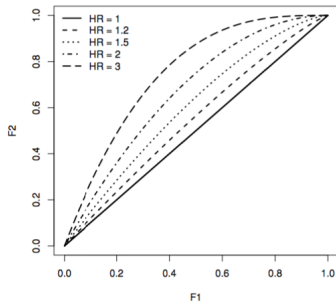
- ▶ HR is the slope of F_2 vs F_1 close to $t = 0$
- ▶ HR can be interpreted as a relative risk if the probability of failure is small

PH Model Assumptions and Interpretation

Fig. Visualizing effect of hazard ratio (HR=1.5; dotted versus solid) on the survival functions
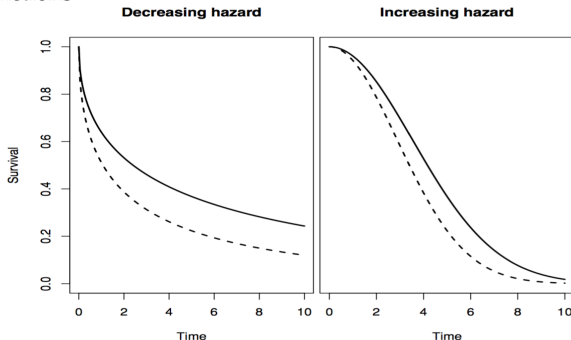
- ▶ The 5-year survival probabilities for the control group (solid line) are the same for both panels. Clear difference in effect on the life expectancy.
- ▶ Left: the median survival is reduced by about 50%
- ▶ Right: the reduction in median survival is about 20%

Under the PH assumption, all survival curves for different X are power functions of the baseline survival curve:

$$S(t|X) = [S_0(t)]^{\exp(\beta^T X)}$$
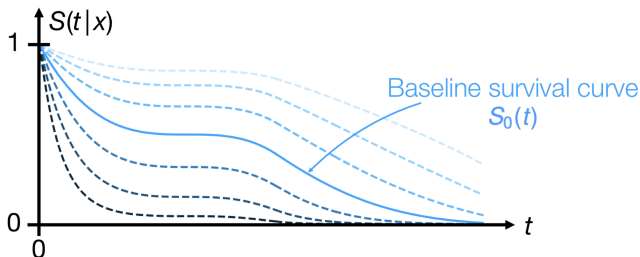
Which directions correspond to positive/negative values of $\beta^T X$?

Estimation for Cox Model

Estimation of hazard ratios $\boldsymbol{\beta}$ in Cox model is through partial likelihood. Estimation of baseline hazard is ignored.

Start from the full data log-likelihood for the observed data $(t_1, d_1, \boldsymbol{x}_1), \cdots, (t_n, d_n, \boldsymbol{x}_n)$

$$
\begin{aligned}
l(\lambda_0, \boldsymbol{\beta}) &= \sum_{i=1}^{n} \log(f(t_i)^{d_i} S(t_i)^{1-d_i}) = \sum_{i=1}^{n} \log(\lambda(t_i)^{d_i} S(t_i)) \\
&= \sum_{i=1}^{n} \left(-\Lambda_0(t_i) \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}) + d_i[\log(\lambda_0(t_i)) + \boldsymbol{x}_i^T \boldsymbol{\beta}] \right)
\end{aligned}
$$

where $t_i = \min(\tilde{t}_i, c_i)$ (minimum of event time and censoring time), d_i event indicator, $\boldsymbol{x}_i$ covariates. This leads to a discrete version of the hazard and cumulative hazard

$$
\Lambda_0(t) = \sum_{t_i \leq t} \lambda_0(t_i).
$$

Plugging cumulative hazard into the log-likelihood to obtain

$$l(\lambda_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left(-\lambda_0(t_i) \sum_{j \in R(t_i)} \exp(\boldsymbol{x}_j^T \boldsymbol{\beta}) + d_i[\log(\lambda_0(t_i)) + \boldsymbol{x}_i^T \boldsymbol{\beta}] \right),$$

where $R(t_i)$ is the at-risk set at time t_i, i.e., it includes all subjects who haven't failed by time t_i.

For fixed value of β, the maximal for baseline hazard is

$$\hat{\lambda}_0(t_i | \boldsymbol{\beta}) = \frac{d_i}{\sum_{j \in R(t_i)} \exp(\boldsymbol{x}_j^T \boldsymbol{\beta})}.$$

The baseline hazard function only has a jump at the observed event times.

Estimation for Cox Model

Plugging in the maximized baseline hazard into log-likelihood, the resulting maximized (or profile) log-likelihood is

$$l(\hat{\lambda}_0(.|\boldsymbol{\beta}), \boldsymbol{\beta}) = \sum_{i=1}^{n} d_i \left(-1 + \log(\hat{\lambda}_0(t_i|\boldsymbol{\beta}) + \boldsymbol{x}_i^T\boldsymbol{\beta}) \right) = -\sum_i d_i + PL(\boldsymbol{\beta}),$$

$$PL(\boldsymbol{\beta}) = \sum_{i=1}^{n} d_i \log \left(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\sum_{j \in R(t_i)} \exp(\boldsymbol{x}_j^T\boldsymbol{\beta})} \right)$$

$PL(\boldsymbol{\beta})$ is Cox's partial likelihood[2]

▶ Interpretation: $\dfrac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\sum_{j \in R(t_i)} \exp(\boldsymbol{x}_j^T\boldsymbol{\beta})}$ is the conditional probability that individual i is the one that failed at event time t_i given the risk set $R(t_i)$ of all individuals who were still event free just prior to time t_i.

▶ Maximize the partial likelihood will provide estimation of $\boldsymbol{\beta}$.

[2]Cox, D. R. 1972. "Regression Models and Life Tables". *Journal of the Royal Statistical Society*, 34: 187-220. (with discussion), Ser. B.

How about baseline hazard (need to compute survival function)?

▶ Plug $\widehat{\boldsymbol{\beta}}$ into the maximizer of baseline hazard function:

$$\hat{\lambda}_0(t_i) = \frac{d_i}{\sum_{j \in R(t_i)} \exp(\boldsymbol{x}_j^T \widehat{\boldsymbol{\beta}})}.$$

Why is Cox's partial likelihood approach a good approach?

▶ Enjoys all the good properties as a regular MLE, i.e., has an asymptotic normal distribution with mean $\boldsymbol{\beta}$ and covariance matrix given by the inverse observed Fisher information matrix[3].

▶ Theoretically optimal: it is semiparametrically efficient[4]!

---

[3]Tsiatis, A. A. (1981). A large sample study of Cox's regression model. The Annals of Statistics, 93-108.

[4]Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. The Annals of Statistics, 1100-1120.

Cox model has played critical role in survival analysis and is still used widely.
Assumptions:

- The covariates information can be summarized through a linear predictor $Z = \boldsymbol{X}^T \boldsymbol{\beta}$

- The hazard function satisfies the PH assumption

Other alternative models (proportional odds etc) are equivalent when $t \approx 0$, $z \approx 0$, the violation to PH assumption can be shown statistically if the predictor has a large effect on survival, or if there is long follow-up or there are many events.

Some violation of the PH assumption can be relaxed within Cox model framework.

Consider likely mechanism of violation of PH assumption:

- ▶ Heterogeneity between individuals: frailties (random effects)
- ▶ Dynamic behavior of covariates (time-dependent covariates)
- ▶ Measurement error in covariates
- ▶ Informative dropout; competing risks

Heterogeneity can be expressed in terms of a non-uniform survival function ($S_i(t)$, each subject has his/her own survival function), cumulative hazard or hazard function itself. Consider frailty model:

$$\lambda_i(t) = \gamma_i \lambda(t),$$

γ_i is a latent variable with a certain distribution (gamma frailty).

Extending the Model by Stratification

Cox PH model can be adjusted for factors that are not modeled or do not satisfy the PH assumption. Clinical study site may take on many levels and the sample size may not be large enough to model this nominal variable with many dummy variables.

Stratified Cox PH model allows the underlying hazard function to vary across levels of the stratification factors. A stratified Cox analysis ranks the failure times separately within strata.

$$\lambda(t|\boldsymbol{X}, C = j) = \lambda_{j0}(t) \exp(\boldsymbol{X}^T\boldsymbol{\beta}),$$

$\lambda_{j0}(t)$ is the underlying baseline hazard function for the jth stratum.

- ▶ does not assume any connection between the shapes of these functions for different strata.
- ▶ a common vector of regression coefficients is fitted across strata
- ▶ Example: a Cox model with age as a (modeled) predictor and sex as a stratification variable estimates the common effect of age by pooling information about age over the two sexes. The effect of age is adjusted by sex differences, but no assumption is made about how sex affects survival.

Violation of PH Assumptions

"Aging covariates": covariates measured at baseline is a "snap-shot" of the health state of a subject. The true covariate values can be time-dependent, but only observed at $t = 0$.

Cox PH model can be extended to include time-dependent covariates. Considerer a time-invariant covariate (e.g., gender) X_1 and a time-dependent covariate $X_2(t)$ (e.g., body mass index):

$$\lambda(t|X_1, X_2(t)) = \lambda_0(t) \exp(\beta_1 X_1 + \beta_2 X_2(t))$$

Consider $X_2(t)$ expressed as a regression model in terms of $X_2(0)$ $X_2(t) = \gamma(t)X_2(0) + \varepsilon(t)$. The hazard model using baseline covariates:

$$\lambda(t|X_1, X_2(0), \varepsilon(t)) = \lambda_0(t) \exp(\beta_1 X_1 + \beta_2 \gamma(t) X_2(0) + \beta_2 \varepsilon(t))$$

The effect of $X_2(0)$ can vary over time, depending on the covariance structure of the $X_2(.)$ process. The effect of X_1 will be influenced by the presence of random variable $\varepsilon(t)$. The effect will be negligible for small t because the variance of $\varepsilon(t)$ will be very small for small t.

Violation of PH Assumptions

Measurement error in covariates has similar effects as "aging covariates". Consider X_2 as a measurement of true covariate Z observed with error:

$$X_2 = \gamma Z + \varepsilon$$

The hazard function is

$$\lambda(t|X_1, X_2) = \lambda_0(t) \exp(\beta_1 X_1 + \beta_2 \gamma Z + \beta_2 \varepsilon)$$

The regression coefficient of Z is $\beta_2 \gamma$ instead of β_2 (referred as attenuation in measurement error model literature). The effect of observed covariates X_1 and X_2 cannot be described by a Cox model, because of the presence of the frailty term $\exp(\beta_2 \varepsilon)$, which has a log-normal distribution.

Graphs can be used to check PH assumption. Statistical tests of PH assumption are developed based on residuals and martingales.

When a factor violates the PH assumption and a test of association is not needed, the factor can be adjusted through stratification. For continuous predictors, stratify into quantile groups.

Consider time-dependent effect or time-dependent covariates

Semiparametric models that assume an effect other than PH, for example, the proportional odds model.

Partial likelihood:

$$PL(\boldsymbol{\beta}) = \sum_{i=1}^{n} d_i \log \left(\frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{\sum_{j \in R(t_i)} \exp(\boldsymbol{x}_j^T \boldsymbol{\beta})} \right)$$

Penalized partial likelihood minimizes $-PL$ with a penalty function L (lasso, SCAD, MCP etc)

$$
\begin{aligned}
loss &= -PL(\beta) + \lambda L(\beta) \\
&= \sum_i d_i \left[-\boldsymbol{x}_i^T \beta + \log(\sum_{j \in R(t_i)} \exp(\boldsymbol{x}_j^T \boldsymbol{\beta}) \right] + \lambda L(\beta)
\end{aligned}
$$

Relax linear function $\boldsymbol{x}_j^T \boldsymbol{\beta}$ as nonparametric function $\phi(\boldsymbol{x}_i)$. Fit $\phi(\boldsymbol{x})$ using machine learning approaches (e.g., neural nets). Or relax to be time-dependent $\phi(\boldsymbol{x}, t)$.

Alternative Models

Other ways of modeling predictors: increase in expected failure time, additive effect, effect described as a mortality ratio (relative risk), odds ratio. However, hazard ratio is often a natural way to describe an effect on survival time.

Accelerated failure time (AFT) model:

$$\log(T) = \boldsymbol{X}^T\boldsymbol{\beta} + \sigma\varepsilon,$$

where ε is an error term with cumulative distribution $F(s) = P(\varepsilon \leq s)$.

Proportional odds model

$$S(t|\boldsymbol{X}) = \frac{1}{1 + \exp(\boldsymbol{X}^T\boldsymbol{\beta} + A(t))}$$

Additive hazards model

$$\lambda(t|\boldsymbol{X}) = \boldsymbol{X}^T\boldsymbol{\beta}(t) + \lambda_0(t)$$

All are much more complicated to fit than Cox model.

Case Study 1: Remdesivir Trial for the
Treatment of COVID-19

An Interview with Dr. Dean Follman

Remdesivir for the Treatment of Covid-19 — Final Report

- ▶ Patients hospitalized with lower respiratory tract involvement and COVID-19 positive
- ▶ 1062 randomized, 60 sites (int'l), double-blind
- ▶ Intervention: remdesivir (200mg, 100mg QD 9 days) versus placebo
- ▶ Primary outcome: time to recovery (discharge from the hospital) by day 29.
- ▶ Secondary outcomes: clinical status at day 15 measured on an ordinal scale of 8 categories; mortality

Competing risks are events that prevent or affect the chance of the outcome event. Examples of competing risks

- ▶ Among COVID patients, death is a competing risk for recovery.
- ▶ Among dialysis patients, kidney transplant is a competing risk for death.

Why does competing risks need special care? Censoring is not independent of time to outcome events.

Primary censoring forms:
- death
- end of trial period (day 29)
- adverse events or withdraw consent (rare)

Death is a competing risk for the primary event of recovery

- Once death occurs, recovery cannot occur (recovery rate=0)
- Recovery is modeled based on subdistribution hazard approach (Fine-Gray, 1999)
- In the analysis, death is considered as censoring at day 29 (not at the time of death)

Methods for Competing Risks

	t_1	t_2	t_3	t_4	$\cdots$	t_L
# people who die	d_1	d_2	d_3	d_4	$\cdots$	d_L
# people at risk	n_1	n_2	n_3	n_4	$\cdots$	n_L
# people have competing risk	c_1	c_2	c_3	c_4	...	c_L

Remove competing risks from people at risk?

No: **subdistribution hazard (SDH)**
$n_2 = n_1 - d_1 - k_1$; k_1 (# censored in t_1)
Yes: **cause-specific hazard (CSH)**
$n_2 = n_1 - d_1 - k_1 - c_1$

	t_1	t_2	t_3	t_4	$\cdots$	t_L
SDH	1/4	0/3	1/3	0/2		
CSH	1/4	0/2	1/2	0/1		

Subdistribution hazard approach (SDH; Fine and Gray 1999):

- ▶ Compute cumulative incidence function (CIF) for event j

$$\hat{F}_j(t) = \sum_{t_{(i)} \leq t} \hat{S}(t_{(i)}) d_{ij} / n_i.$$

- ▶ Sum of CIFs estimates the overall survival (over all types of events)

$$\sum_j \hat{F}_j(t) = 1 - \hat{S}_{KM}(t)$$

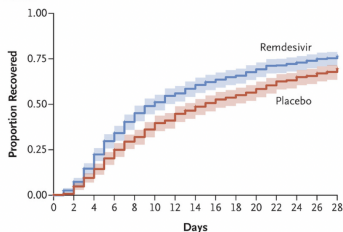- ▶ Mostly for prognostic analysis

Cause specific hazard (CSH) approach

- ▶ $CSH_j(t)$: rate of outcome j occurring given that no event has yet occurred

- ▶ Useful for association analysis with multivariate events (e.g. HRs from Cox), estimates of HR are directly based on at-risk group, rather than mixture of at-risk and had competing-outcome groups

- ▶ Mostly for etiologic analyses

Primary analysis of time-to-recovery: Stratified Cox model
(stratify by disease severity), subdistribution hazard for
handling competing risk by death

	Overall	
	Remdesivir (N=541)	Placebo (N=521)
Recovery		
No. of recoveries	399	352
Median time to recovery (95% CI) — days	10 (9–11)	15 (13–18)
Rate ratio (95% CI)†	1.29 (1.12–1.49 [P<0.001])	

Table S14. Odds Ratio for Better (Lower) Clinical Status Score at Day 15 by Treatment Using a Proportional Odds Model, Remdesivir Relative to Placebo – ITT Population

Analysis/Subgroup	Treatment Group	Odds Ratio		P-value
		Estimate	95% CI	
Analysis of Key Secondary Endpoint[a], Full ITT Population	Remdesivir (N=541)	1.6	1.3, 1.9	<0.001
	Placebo (N=521)			
Geographic region (US Site)	Remdesivir (N=427)	1.6	1.3, 2.0	<0.001
	Placebo (N=410)			
Geographic region (Non-US Site)	Remdesivir (N=114)	1.3	0.8, 2.1	0.233
	Placebo (N=111)			
Geographic region (North America)	Remdesivir (N=431)	1.6	1.3, 2.0	<0.001
	Placebo (N=416)			
Geographic region (Asia)	Remdesivir (N=26)	1.3	0.5, 3.5	0.592
	Placebo (N=26)			
Geographic region (Europe)	Remdesivir (N=84)	1.4	0.8, 2.4	0.246
	Placebo (N=79)			
Duration of symptoms prior to enrollment (First Quartile (≤ 6 Days))	Remdesivir (N=158)	2.7	1.8, 4.2	<0.001
	Placebo (N=124)			
Duration of symptoms prior to enrollment (Second Quartile (7 to ≤ 9 Days))	Remdesivir (N=148)	1.1	0.8, 1.7	0.573
	Placebo (N=152)			

Case Study 2: Multilayer Exponential Family Factor models for integrative analysis and learning disease progression[5]

[5]Qinxia Wang, Yuanjia Wang. Multilayer Exponential Family Factor models for integrative analysis and learning disease progression, Biostatistics, 2022.
https://doi.org/10.1093/biostatistics/kxac042