Midterm Review Session

Biost 515/518

Jon Fintzi and LaNae Schaal

General regression model - setup

Every regression model we have studied this quarter has the form

$$g(\theta(Y)) = \beta_0 + \beta_1 X + \beta_2 W_1 + \cdots + \beta_{p+1} W_p$$

General regression model - setup

Every regression model we have studied this quarter has the form

$$g(\theta(Y)) = \beta_0 + \beta_1 X + \beta_2 W_1 + \cdots + \beta_{p+1} W_p$$

- Y is the response variable.
- X is the predictor of interest.
- $W_1, \ldots, W_p$ are variables used to further stratify the population, to adjust for confounding, or to provide additional precision.

General regression model - setup

Every regression model we have studied this quarter has the form

$$g(\theta(Y)) = \beta_0 + \beta_1 X + \beta_2 W_1 + \cdots + \beta_{p+1} W_p$$

- $\theta(Y)$ is a summary measure of the distribution of Y
  - Choice of $\theta(Y)$ depends on the type of Y - binary, unordered categorical, ordered categorical, or continuous.
  - Common choices for $\theta(Y)$ are the
    - Mean
    - Geometric mean (for positive Y)
    - Median of the distribution of Y
    - Probability or odds that $Y > c$ for some interesting value of c
    - Hazard function

General regression model - setup

Every regression model we have studied this quarter has the form

$$g(\theta(Y)) = \beta_0 + \beta_1 X + \beta_2 W_1 + \cdots + \beta_{p+1} W_p$$

General regression model - setup

Every regression model we have studied this quarter has the form

$$g(\theta(Y)) = \beta_0 + \beta_1 X + \beta_2 W_1 + \cdots + \beta_{p+1} W_p$$

The *link function*, $g(\cdot)$, specifies which function of $\theta(Y)$ is linear in X and $W_1, \ldots, W_p$.

General regression model - setup

Every regression model we have studied this quarter has the form

$$g(\theta(Y)) = \beta_0 + \beta_1 X + \beta_2 W_1 + \cdots + \beta_{p+1} W_p$$

The *link function*, $g(\cdot)$, specifies which function of $\theta(Y)$ is linear in X and $W_1, \ldots, W_p$.

- Additive models are described using an *identity link*:

$$g(\theta(Y)) = \theta(Y)$$

- Multiplicative models are described using a *log link*:

$$g(\theta(Y)) = log(\theta(Y))$$

General regression model - β_0

Suppose we fit a simple regression to study the association between Y and X:

$$g(\theta(Y)) = \beta_0 + \beta_1 X$$

- □ If we are interested in the value of $\theta(Y)$ when $X = x$ (i.e. the *fitted value*), we simply plug in the value x for X in our model. So,

$$g(\theta(Y|X = x)) = \beta_0 + \beta_1 x$$

General regression model - β_0

Suppose we fit a simple regression to study the association between Y and X:

$$g(\theta(Y)) = \beta_0 + \beta_1 X$$

- □ If we are interested in the value of $\theta(Y)$ when $X = x$ (i.e. the *fitted value*), we simply plug in the value x for X in our model. So,

$$g(\theta(Y|X = x)) = \beta_0 + \beta_1 x$$

- □ In particular, the fitted value when $X = 0$ is

$$g(\theta(Y|X = 0)) = \beta_0$$

General regression model - β_0

Suppose we fit a simple regression to study the association between Y and X:

$$g(\theta(Y)) = \beta_0 + \beta_1 X$$

- When $g(\cdot)$ is the identity link,

$$\theta(Y|X=0) = \beta_0$$

- When $g(\cdot)$ is the log link,

$$log(\theta(Y|X=0)) = \beta_0$$

- However, we are interested in $\theta(Y|X=0)$. Therefore, we exponentiate both sides to obtain $\theta(Y|X=0) = e^{\beta_0}$

General regression model - β_1

Now, suppose we want to compare $\theta(Y|X)$ in groups differing in their level of X.

General regression model - β_1

Now, suppose we want to compare $\theta(Y|X)$ in groups differing in their level of X. Then,

$$\begin{aligned} g(\theta(Y|X=x+1)) - g(\theta(Y|X=x)) &= \beta_0 + \beta_1(x+1) - \beta_0 - \beta_1 x \\ &= \beta_1 \end{aligned}$$

General regression model - β_1

Now, suppose we want to compare $\theta(Y|X)$ in groups differing in their level of X. Then,

$$\begin{aligned} g(\theta(Y|X = x+1)) - g(\theta(Y|X = x)) &= \beta_0 + \beta_1(x+1) - \beta_0 - \beta_1 x \\ &= \beta_1 \end{aligned}$$

Therefore, β_1 is the difference in $g(\theta(Y))$ between two groups differing in their level of X by one unit.

General regression model - β_1

Now, suppose we want to compare $\theta(Y|X)$ in groups differing in their level of X. Then,

$$\begin{aligned} g(\theta(Y|X=x+1)) - g(\theta(Y|X=x)) &= \beta_0 + \beta_1(x+1) - \beta_0 - \beta_1 x \\ &= \beta_1 \end{aligned}$$

Therefore, β_1 is the difference in $g(\theta(Y))$ between two groups differing in their level of X by one unit.

- □ Note that we are interested in $\theta(Y|X=x+1) - \theta(Y|X=x+1)$.
- □ Also, we have not yet said anything about which $g(\cdot)$ we are using, or about what type of variables are Y and X.

General regression model - identity link

If $g(\cdot)$ is the identity link, we are done since

$$\begin{aligned} g(\theta(Y|X=x+1)) - g(\theta(Y|X=x)) &= \theta(Y|X=x+1) - \theta(Y|X=x) \\ &= \beta_1 \end{aligned}$$

General regression model - identity link

If $g(\cdot)$ is the identity link, we are done since

$$\begin{aligned} g(\theta(Y|X=x+1)) - g(\theta(Y|X=x)) &= \theta(Y|X=x+1) - \theta(Y|X=x) \\ &= \beta_1 \end{aligned}$$

So, β_1 is the difference in $\theta(Y|X)$ for two groups differing in their level of X by one unit.

General regression model - log link

If $g(\cdot)$ is the log link, then

$$g(\theta(Y|X=x+1)) - g(\theta(Y|X=x)) \quad = \quad \log(\theta(Y|X=x+1)) - \log(\theta(Y|X=x))$$

General regression model - log link

If $g(\cdot)$ is the log link, then

$$\begin{aligned} g(\theta(Y|X=x+1)) - g(\theta(Y|X=x)) &= \log(\theta(Y|X=x+1)) - \log(\theta(Y|X=x)) \\ &= \beta_0 + \beta_1(x+1) - \beta_0 - \beta_1 x \end{aligned}$$

General regression model - log link

If $g(\cdot)$ is the log link, then

$$\begin{aligned} g(\theta(Y|X=x+1)) - g(\theta(Y|X=x)) &= \log(\theta(Y|X=x+1)) - \log(\theta(Y|X=x)) \\ &= \beta_0 + \beta_1(x+1) - \beta_0 - \beta_1 x \end{aligned}$$

So,

$$\log\left(\frac{\theta(Y|X=x+1)}{\theta(Y|X=x)}\right) = \beta_1$$

General regression model - log link

If $g(\cdot)$ is the log link, then

$$\begin{aligned} g(\theta(Y|X=x+1)) - g(\theta(Y|X=x)) &= \log(\theta(Y|X=x+1)) - \log(\theta(Y|X=x)) \\ &= \beta_0 + \beta_1(x+1) - \beta_0 - \beta_1 x \end{aligned}$$

So,

$$\log\left(\frac{\theta(Y|X=x+1)}{\theta(Y|X=x)}\right) = \beta_1$$

However, we are interested in $\frac{\theta(Y|X=x)}{\theta(Y|X=x)}$. So, we exponentiate to obtain

$$\frac{\theta(Y|X=x+1)}{\theta(Y|X=x)} = e^{\beta_1}$$

Therefore, e^{β_1} is the ratio of the summary measure in two groups differing in their value of X by one unit.

General regression model - comments

Some comments:

- If we had included additional covariates in our model, our interpretations of β_1 would have changed only in that we further stipulate that the values of the additional covariates be held constant.
- In the above interpretations, we used that $g(\theta(Y|X))$ is linear in X. If this linearity assumption does not hold, we interpret the slope parameter as an average difference or average ratio across all possible subpopulations differing in their levels of X by 1 unit. However, we no longer claim that that the ratio or difference is constant for each unit difference in X.

General regression model - log transformed predictors

We are often interested in contrasts of our summary measure for groups defined by a multiplicative change in the level of the covariate.

General regression model - log transformed predictors

We are often interested in contrasts of our summary measure for groups defined by a multiplicative change in the level of the covariate.

□ If we have the model $g(\theta(Y|X)) = \beta_0 + \beta_1 \log_k(X)$, then we have seen that β_1 is the difference in $g(\theta(Y))$ associated with a 1 unit difference in $\log_k(X)$.

□ A 1 unit increase in $log_k(X)$ corresponds to a k-fold increase in X since

$$\begin{aligned} 1 &= \log_k(x') - \log_k(x) \\ &= \log_k\left(\frac{x'}{x}\right) \\ \Longrightarrow k &= \frac{x'}{x} \\ \Longrightarrow kx &= x' \end{aligned}$$

Linear regression - setup

The linear regression specializes the general regression model by:

- □ $\theta(Y|X,W) = \mathrm{E}(Y|X,W)$
- □ $g(\theta) = \theta$, i.e. g is the identity link

Linear regression - setup

The linear regression specializes the general regression model by:

- $\theta(Y|X,W) = \mathrm{E}(Y|X,W)$
- $g(\theta) = \theta$, i.e. g is the identity link
- $\beta_0 = \mathrm{E}(Y|X=0, W_1 = \cdots = W_p = 0)$

Linear regression - setup

The linear regression specializes the general regression model by:

- $\theta(Y|X,W) = \mathrm{E}(Y|X,W)$
- $g(\theta) = \theta$, i.e. g is the identity link
- $\beta_0 = \mathrm{E}(Y|X=0, W_1 = \cdots = W_p = 0)$
- β_1 is the difference in the mean of Y in two populations differing in X by 1 unit, holding constant $W_1, \ldots, W_p$.
  - Note that when Y is binary,
    $$Y = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{cases}$$
  - $\mathrm{E}(Y) = \Pr(Y=1)$, i.e. the expectation of Y is the probability, or risk, of Y taking on a value of 1. In this case, β_1 is the risk difference for two groups differing in X by 1 unit, holding constant $W_1, \ldots, W_p$.

Linear regression - properties

- The ordinary lease squares regression estimates are the most precise estimates among all unbiased linear estimates under the following assumptions:
  - Independent observations
  - Homoscedasticity (constant variance of Y|X)
  - Linearity
- When homoscedasticity does not hold, we may use weighted least squares to arrive at optimal estimates.
- In practice, we use the Huber-White robust standard errors. *Note:* Robust standard errors do not assume heteroscedasticity. They merely allow for the possibility that the error variances are not constant.
- If $Y|X$ is normally distributed, the OLS estimates are maximum likelihood estimates and are efficient estimators. If $Y|X$ is not normal, the OLSEs are asymptotically normal.

Linear regression - facts

- When we perform simple linear regression on a binary predictor:
  - Classical SLR gives us exactly the t-test that *assumes* equal variances
  - SLR with robuse SE approximates the t-test that *allows* for unequal variances

Linear regression - facts

- When we perform simple linear regression on a binary predictor:
  - Classical SLR gives us exactly the t-test that *assumes* equal variances
  - SLR with robuse SE approximates the t-test that *allows* for unequal variances
- In classical simple SLR, the test for statistical significance of the slope is exactly equivalent to the test for statistically significant correlation

Linear regression - facts

In classical SLR, suppose we have wrongly assumed homoscedasticity. Then our inference will be:

- Anti-conservative (p-values too small and CIs too narrow) if $Var(Y|X)$ is higher for outlying groups of X
- Conservative if $Var(Y|X)$ is lower for outlying groups of X.
- These results are similar to what we know about the behavior of the t-test when we wrongly assume equal variances between groups.

Pearson correlation and linear associations

Recall that the formula for the Pearson correlation coefficient could be rewritten as

$$r \approx \beta \sqrt{\frac{Var(X)}{\beta^2 Var(X) + Var(Y|X = x)}}$$

Where

- β is the LS slope for the regression line
- Var(X) is the variance of X in the sample
- Var(Y|X=x) is the within group variance of Y for a fixed level of X

Pearson correlation and linear associations

Then correlation tends to increase as:

- □ The absolute value of the slope increases
- □ The variance of X in the sample increases
- □ The variance of $Y|X$ decreases

Pearson correlation and linear associations

Then correlation tends to increase as:

- The absolute value of the slope increases
- The variance of X in the sample increases
- The variance of Y|X decreases

Note the following though:

- Sample size does not influence whether or not the correlation is higher or lower.
- Correlation is a measure of the strength of the *linear* association between Y and X. A correlation of 0 does not mean that Y and X are not associated - e.g. if $Y = (X - \bar{X})^2$, the correlation between $Y$and X will be 0, but the variables are clearly related.

Linear regression - log transformed response

Linear regression with a log transformed response variable specializes the general regression model by taking:

- $\theta(Y) = GM(Y|X, W)$, i.e. the geometric mean of the distribution of Y.
- $g(\theta) = \log(\theta)$, i.e. the log link.

Linear regression - log transformed response

Linear regression with a log transformed response variable specializes the general regression model by taking:

- □ $\theta(Y) = GM(Y|X, W)$, i.e. the geometric mean of the distribution of Y.
- □ $g(\theta) = \log(\theta)$, i.e. the log link.
- □ $e^{\beta_0} = GM(Y|X = 0, W_1 = \cdots = W_p = 0)$

Linear regression - log transformed response

Linear regression with a log transformed response variable specializes the general regression model by taking:

- $\theta(Y) = GM(Y|X, W)$, i.e. the geometric mean of the distribution of Y.
- $g(\theta) = \log(\theta)$, i.e. the log link.
- $e^{\beta_0} = GM(Y|X = 0, W_1 = \cdots = W_p = 0)$
- $e^{\beta_1} = \frac{GM(Y|X=x+1, W_1=w_1,\ldots,W_p=w_p)}{GM(Y|X=x, W_1=w_1,\ldots,W_p=w_p)}$
  - i.e. e^{β_1} is the ratio of geometric means for two groups differing in their value of X by one unit, but agreeing in their levels of $W_1, \ldots, W_p$.
- *Important: linear regression with a log transformed response only makes sense when* $Y > 0$

Side bar - geometric means and logs

The geometric mean is estimated by

$$\widehat{GM(Y)} \quad = \quad \left(\prod_{i=1}^{n} Y_i \right)^{\frac{1}{n}}$$

Side bar - geometric means and logs

The geometric mean is estimated by

$$\begin{aligned}\widehat{GM(Y)} &= \left(\prod_{i=1}^{n} Y_i\right)^{\frac{1}{n}} \\ &= \exp\left(\log\left(\left(\prod_{i=1}^{n} Y_i\right)^{\frac{1}{n}}\right)\right)\end{aligned}$$

Side bar - geometric means and logs

The geometric mean is estimated by

$$\begin{aligned}
\widehat{GM(Y)} &= \left(\prod_{i=1}^{n} Y_i\right)^{\frac{1}{n}} \\
&= \exp\left(\log\left(\left(\prod_{i=1}^{n} Y_i\right)^{\frac{1}{n}}\right)\right) \\
&= \exp\left(\frac{1}{n}\sum_{i=1}^{n} \log(Y_i)\right)
\end{aligned}$$

Side bar - geometric means and logs

The geometric mean is estimated by

$$
\begin{aligned}
\widehat{GM(Y)} &= \left(\prod_{i=1}^{n} Y_i\right)^{\frac{1}{n}} \\
&= \exp\left(\log\left(\left(\prod_{i=1}^{n} Y_i\right)^{\frac{1}{n}}\right)\right) \\
&= \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log(Y_i)\right)
\end{aligned}
$$

So the log of the geometric mean, $\mathrm{E}(\log(Y))$ is estimated by $\frac{1}{n}\sum_{i=1}^{n}\log(Y_i)$ Exponentiating β_0 and β_1 gets us back to the geometric mean scale.

Logistic regression - setup

Suppose that Y is a binary variable, so

$$Y = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1 - p \end{cases}$$

Logistic regression specializes the general regression model by taking:

Logistic regression - setup

Suppose that Y is a binary variable, so

$$Y = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{cases}$$

Logistic regression specializes the general regression model by taking:

- □ $\theta(Y) = \frac{\Pr(Y=1)}{\Pr(Y=0)}$, i.e. the odds that $Y = 1$.
- □ $g(\theta) = \log(\theta)$, i.e. $g(\cdot)$ is the log link.

Logistic regression - setup

Suppose that Y is a binary variable, so

$$Y = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1 - p \end{cases}$$

Logistic regression specializes the general regression model by taking:

□ $\theta(Y) = \frac{\Pr(Y=1)}{\Pr(Y=0)}$, i.e. the odds that $Y = 1$.

□ $g(\theta) = \log(\theta)$, i.e. $g(\cdot)$ is the log link.

Logistic regression - setup

Suppose that Y is a binary variable, so

$$Y = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{cases}$$

Logistic regression specializes the general regression model by taking:

□ $\theta(Y) = \frac{\Pr(Y=1)}{\Pr(Y=0)}$, i.e. the odds that $Y = 1$.

□ $g(\theta) = \log(\theta)$, i.e. $g(\cdot)$ is the log link.

Exercise: derive that e^{β_0} is the odds of $Y = 1$ when $X = 0, W_1 = \cdots = W_p = 0$, and that e^{β_1} is the ratio of the odds that $Y = 1$ between two groups differing in 1 unit in X. *Hint:* Follow the steps in the derivation for the general regression model with log link.

Poisson regression - setup

Poisson regression specializes the general regression model by taking:

- $\theta(Y) = E(Y)$
- $g(\theta) = log(\theta)$

Poisson regression - setup

Poisson regression specializes the general regression model by taking:

- $\theta(Y) = E(Y)$
- $g(\theta) = log(\theta)$

Classically, Poisson regression is used for count data. We have also seen Poisson regression used when Y is a binary variable, and when we are interested in multiplicative contrasts in means.

- When Y records Poisson count data, the mean of Y is the mean number of events. So, the response is an event rate.
- When Y is a binary variable, the mean of Y is the probability that $Y = 1$. So, the response is a risk.

Poisson regression - setup

Poisson regression specializes the general regression model by taking:

- □ $\theta(Y) = E(Y)$
- □ $g(\theta) = log(\theta)$

Classically, Poisson regression is used for count data. We have also seen Poisson regression used when Y is a binary variable, and when we are interested in multiplicative contrasts in means.

- □ When Y records Poisson count data, the mean of Y is the mean number of events. So, the response is an event rate.
- □ When Y is a binary variable, the mean of Y is the probability that $Y = 1$. So, the response is a risk.

Important: Poisson regression assumes that the mean is equal to the variance. Robust standard errors permit us to circumvent this assumption and should be used.

Poisson regression - parameter interpretation

□ As with our other regressions involving a log link, e^{β_0} is the mean of Y (event rate, or risk) when
$X = 0,\ W_1 = \cdots = W_p = 0$.

Poisson regression - parameter interpretation

- As with our other regressions involving a log link, e^{β_0} is the mean of Y (event rate, or risk) when $X = 0,\ W_1 = \cdots = W_p = 0$.
- Similarly, e^{β_1} is the ratio of the mean of Y between groups differing in their level of X by 1 unit, but otherwise agreeing in their values of $W_1, \ldots, W_p$ (*Derive this!*)

Poisson regression - parameter interpretation

- As with our other regressions involving a log link, e^{β_0} is the mean of Y (event rate, or risk) when $X = 0, \; W_1 = \cdots = W_p = 0$.
- Similarly, e^{β_1} is the ratio of the mean of Y between groups differing in their level of X by 1 unit, but otherwise agreeing in their values of $W_1, \ldots, W_p$ (*Derive this!*)
  - When Y records count data, e^{β_1} denotes a relative rate.
  - When Y is binary, e^{β_1} is the relative risk.

Poisson regression - normalized counts

□ Frequently, we need to normalize our counts, for example over space-time, when we apply Poisson regression. This is accomplished via an offset.

Suppose $Y_i \sim Poisson(\frac{\lambda_i}{t_i})$. Then,

$$\begin{aligned}
\log \mathrm{E}(Y_i | T_i, X_i) &= \log\left(\frac{\lambda_i}{t_i}\right) \\
&= \log(\lambda_i) - \log(t_i) \\
&= \beta_0 + \beta_1 X_i \\
\implies \log(\lambda_i) &= \log(t_i) + \beta_0 + \beta_1 X_i \\
\therefore \lambda_i &= t_i e^{\beta_0 + \beta_1 X_i}
\end{aligned}$$

Odds ratios and relative risks in case-control studies

□ Even though we are often interested in estimating the probability of disease among exposed and unexposed individuals, we cannot do so in a case-control study. Subjects in case-control studies are sampled based on their disease status, not their exposure status.

Odds ratios and relative risks in case-control studies

- □ Even though we are often interested in estimating the probability of disease among exposed and unexposed individuals, we cannot do so in a case-control study. Subjects in case-control studies are sampled based on their disease status, not their exposure status.
- □ However, we can still estimate the odds ratio of disease given exposure status.

Odds ratios and relative risks in case-control studies

- □ Even though we are often interested in estimating the probability of disease among exposed and unexposed individuals, we cannot do so in a case-control study. Subjects in case-control studies are sampled based on their disease status, not their exposure status.
- □ However, we can still estimate the odds ratio of disease given exposure status.
- □ The odds ratio is the same, whether we condition on disease status or exposure status.

Odds ratios and relative risks in case-control studies

The OR given disease is equal to the OR conditioning on exposure:

$$\begin{aligned}
OR(D|E) = \frac{odds(D|E)}{odds(D|Ec)} &= \frac{\Pr(D|E)}{\Pr(D^c|E)}\Bigg/\frac{\Pr(D|E^c)}{\Pr(D^c|E^c)} \\
&= \frac{\Pr(D|E)\Pr(E)}{\Pr(D^c|E)\Pr(E)}\Bigg/\frac{\Pr(D|E^c)\Pr(E^c)}{\Pr(D^c|E^c)\Pr(E^c)} \\
&= \frac{\Pr(D,E)}{\Pr(D^c,E)}\Bigg/\frac{\Pr(D,E^c)}{\Pr(D^c,E^c)} \\
&= \frac{\Pr(E|D)\Pr(D)}{\Pr(E|D^c)\Pr(D^c)}\Bigg/\frac{\Pr(E^c|D)\Pr(D)}{\Pr(E^c|D^c)\Pr(D^c)} \\
&= \frac{\Pr(E|D)}{\Pr(E|D^c)}\Bigg/\frac{\Pr(E^c|D)}{\Pr(E^c|D^c)} \\
&= \frac{\Pr(E|D)}{\Pr(E^c|D)}\Bigg/\frac{\Pr(E|D^c)}{\Pr(E^c|D^c)} \\
&= \frac{odds(E|D)}{odds(E|D^c)} = OR(E|D)
\end{aligned}$$

Proportional hazards regression - setup

Proportional hazards regression is classically used with time-to-event data. The key PH assumption is that the hazards (i.e. the instantaneous rate of failure) is proportional across subpopulations.

Proportional hazards regression - setup

Proportional hazards regression is classically used with time-to-event data. The key PH assumption is that the hazards (i.e. the instantaneous rate of failure) is proportional across subpopulations. Here, we specialize the general regression model by taking:

- $\theta(Y)$ is the hazard function for Y.
- $g(\theta) = \log(\theta)$

Proportional hazards regression - setup

Proportional hazards regression is classically used with time-to-event data. The key PH assumption is that the hazards (i.e. the instantaneous rate of failure) is proportional across subpopulations. Here, we specialize the general regression model by taking:

- □ $\theta(Y)$ is the hazard function for Y.
- □ $g(\theta) = \log(\theta)$

As opposed to an accelerated failure time model, which studies how the rate of death is associated with the predictors (see PH notes, slides 15, 17), PH models compare the which people death chooses relative to their prevalence in the population.

Proportional hazards regression - setup

- The PH model only has to consider the covariates of those subjects still in the risk set at any particular time.
- The model can be expressed in terms of an odds ratio of an event at each time an event occurs. The model averages these odds ratios across all observed event times.
- Information is borrowed between groups over time, as the instantaneous relative risk of an event at any point in time is constant under the PH assumption (Note: the hazard ratio is constant, but the hazards are not).

Proportional hazards regression - setup

Letting $\lambda(t|X_i)$ denote the hazard at time t conditional on X_i, the simple PH regression model specifies that

$$\log(\lambda(t|X_i)) \quad = \quad log(\lambda_{i0}(t)) + \beta_1 X_i)$$

The log baseline hazard function is

$$\log(\lambda(t)|X_i = 0) \quad = \quad \log(\lambda_0(t))$$

while,

$$\begin{aligned} \log(\lambda(t|X_i = x+1)) \quad &= \quad \log(\lambda_0(t)) + \beta_1(x+1) \\ \log(\lambda(t|X_i = x)) \quad &= \quad \log(\lambda_0(t)) + \beta_1(x) \end{aligned}$$

Proportional hazards regression - setup

Exponentiating, we obtain that our model is,

$$\lambda(t|X_i) = \lambda_0(t)e^{\beta_1 X_i}$$

- The baseline hazard function is $\lambda_0(t)$. This effectively functions as the "intercept" for the model.
- The hazard ratio of two groups, differing in their level of X by one unit (and agreeing in the values of any other covariates, had we chosen to include them) is e^{β_1}.

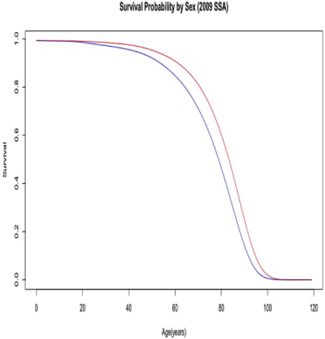
Relating hazards to survival

The hazard function determines the survival function.

Hazard	$\lambda(t\|X_i) = \lambda_0(t)e^{\beta_x X_i}$
Cumulative Hazard	$\Lambda(t\|X_i) = \int_0^t \lambda_0(u)e^{\beta_1 X_i}du$
Survival Function	$S(t\|X_i) = e^{-\Lambda(t\|X_i)} = (S_0(t))^{e^{\beta_1 X_i}}$

Comparing survival curves

We can compare common summary measures of survival distributions between two groups by inspecting their survival curves.



□ Difference in survival at t_0: vertical separation at t_0

□ Difference in quantiles: horizontal separation at p

□ Difference in means: area between curves

□ Hazard: Slope divided by height of curve (difficult to see)