Political Methodology III: Model Based Inference

Justin Grimmer

Professor
Department of Political Science
Stanford University

April 30th, 2019

Model Based Inference

1) Likelihood inference
2) Logit/Probit
3) Ordered Probit
4) Choice Models:
5) Count Models
    - Negative Binomial Regression
        - DGP
        - Quantities of Interest
        - Interpetation
- Clustered Standard Errors
6) Survival Models

Event Count Outcomes

- Outcome: number of times an event occurs

$$Y_i \quad \in \quad \{0, 1, 2, 3, \ldots, \}$$

- Examples:
  1) Number of militarized disputes a country is involved in
  2) Number of times a phrase is used
  3) Number of messages into a Congressional office
  4) Number of votes cast for a particular candidate
- Goal:
  - Model the rate at which events occur
  - Understand how an intervention (e.g. country becoming a democracy) affects rate
  - Predict number of future events

Overdispersion in Poisson Model

- The Poisson model assumes $\mathsf{E}(Y_i \mid X_i) = \mathsf{Var}(Y_i \mid X_i)$
- But for many count data, $\mathsf{E}(Y_i \mid X_i) < \mathsf{V}(Y_i \mid X_i)$
- Potential sources of overdispersion:
  1. unobserved heterogeneity
  2. clustering
  3. contagion or diffusion
  4. (classical) measurement error
- Underdispersion could occur, but rare
- One solution to this is to modify the Poisson model by assuming:

$$\mathsf{E}(Y_i \mid X_i) = \lambda_i = \exp(X_i'\beta) \quad \text{and} \quad \mathsf{Var}(Y_i \mid X_i) = V_i = \sigma^2\lambda_i$$

- This is called the overdispersed Poisson regression model
- When $\sigma^2 > 1$, this corresponds to the negative binomial regression model

Overdispersion in Poisson Model

- The Poisson model assumes $\mathsf{E}(Y_i \mid X_i) = \mathsf{Var}(Y_i \mid X_i)$
- But for many count data, $\mathsf{E}(Y_i \mid X_i) < \mathsf{V}(Y_i \mid X_i)$
- Potential sources of overdispersion:
  1. unobserved heterogeneity
  2. clustering
  3. contagion or diffusion
  4. (classical) measurement error
- Underdispersion could occur, but rare
- One solution to this is to modify the Poisson model by assuming:

$$\mathsf{E}(Y_i \mid X_i) = \lambda_i = \exp(X_i^{'}\beta) \quad \text{and} \quad \mathsf{Var}(Y_i \mid X_i) = V_i = \sigma^2 \lambda_i$$

- This is called the overdispersed Poisson regression model
- When $\sigma^2 > 1$, this corresponds to the negative binomial regression model

Overdispersion in Poisson Model

- The Poisson model assumes $\mathsf{E}(Y_i \mid X_i) = \mathsf{Var}(Y_i \mid X_i)$
- But for many count data, $\mathsf{E}(Y_i \mid X_i) < \mathsf{V}(Y_i \mid X_i)$
- Potential sources of overdispersion:
  1. unobserved heterogeneity
  2. clustering
  3. contagion or diffusion
  4. (classical) measurement error
- Underdispersion could occur, but rare

- One solution to this is to modify the Poisson model by assuming:

$$\mathsf{E}(Y_i \mid X_i) = \lambda_i = \exp(X_i^{'}\beta) \quad \text{and} \quad \mathsf{Var}(Y_i \mid X_i) = V_i = \sigma^2\lambda_i$$

- This is called the overdispersed Poisson regression model
- When $\sigma^2 > 1$, this corresponds to the negative binomial regression model

Overdispersion in Poisson Model

- The Poisson model assumes $\mathsf{E}(Y_i \mid X_i) = \mathsf{Var}(Y_i \mid X_i)$
- But for many count data, $\mathsf{E}(Y_i \mid X_i) < \mathsf{V}(Y_i \mid X_i)$
- Potential sources of overdispersion:
  1. unobserved heterogeneity
  2. clustering
  3. contagion or diffusion
  4. (classical) measurement error
- Underdispersion could occur, but rare
- One solution to this is to modify the Poisson model by assuming:

$$\mathsf{E}(Y_i \mid X_i) = \lambda_i = \exp(X_i^{'}\beta) \quad \text{and} \quad \mathsf{Var}(Y_i \mid X_i) = V_i = \sigma^2 \lambda_i$$

- This is called the overdispersed Poisson regression model
- When $\sigma^2 > 1$, this corresponds to the negative binomial regression model

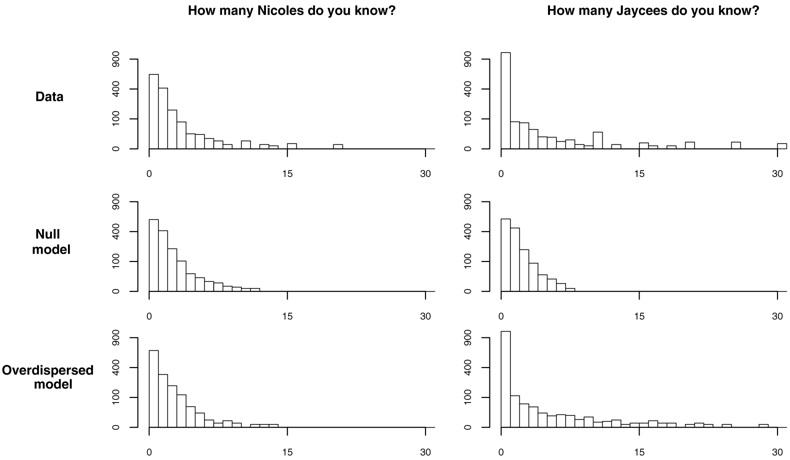
Overdispersion in Poisson Model

- The Poisson model assumes $\mathsf{E}(Y_i \mid X_i) = \mathsf{Var}(Y_i \mid X_i)$
- But for many count data, $\mathsf{E}(Y_i \mid X_i) < \mathsf{V}(Y_i \mid X_i)$
- Potential sources of overdispersion:
  1. unobserved heterogeneity
  2. clustering
  3. contagion or diffusion
  4. (classical) measurement error
- Underdispersion could occur, but rare
- One solution to this is to modify the Poisson model by assuming:

$$\mathsf{E}(Y_i \mid X_i) = \lambda_i = \exp(X_i^{'}\beta) \quad \text{and} \quad \mathsf{Var}(Y_i \mid X_i) = V_i = \sigma^2 \lambda_i$$

- This is called the overdispersed Poisson regression model
- When $\sigma^2 > 1$, this corresponds to the negative binomial regression model

Example: Social Network Survey Data



(Zheng, et al., 2006 *JASA*)

Negative Binomial Distribution

Suppose $Y_i \in \{0, 1, 2, \ldots, \}$. If Y_i has pmf

$$p(y_i) \;=\; \frac{\Gamma\left(\frac{\lambda}{\sigma^2-1} + y_i\right)}{y_i!\Gamma\left(\frac{\lambda}{\sigma^2-1}\right)}\left(\frac{\sigma^2-1}{\sigma^2}\right)^{y_i}\left(\sigma^2\right)^{\frac{-\lambda}{\sigma^2-1}}$$

with $\lambda > 0$ and $\sigma^2 > 1$
Then we will say

$$\begin{aligned} Y_i &\sim \mathsf{NegBin}(\lambda, \sigma^2) \\ E[Y_i] &= \lambda \\ \mathsf{Var}(Y_i) &= \lambda\sigma^2 \end{aligned}$$

Negative Binomial Regression

Suppose:

$$\begin{aligned}
Y_i &\sim \text{Negative Binomial}(\lambda_i, \sigma^2) \\
\lambda_i &= \exp(\boldsymbol{X}_i^{'}\boldsymbol{\beta})
\end{aligned}$$

This implies a likelihood of:

$$\begin{aligned}
L(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}) &= f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}) \\
&= \prod_{i=1}^{N} f(Y_i|\boldsymbol{X}_i, \boldsymbol{\beta}) \\
&= \prod_{i=1}^{N} \frac{\Gamma\left(\frac{\lambda_i}{\sigma^2-1} + y_i\right)}{y_i!\Gamma\left(\frac{\lambda_i}{\sigma^2-1}\right)} \left(\frac{\sigma^2-1}{\sigma^2}\right)^{y_i} \left(\sigma^2\right)^{\frac{-\lambda_i}{\sigma^2-1}}
\end{aligned}$$

Optimize numerically. Usual theorems about asymptotic distributions apply.

Negative Binomial Regression

Negative Binomial Regression:

1) Careful! Variance is sometimes:

$$\text{Var}(Y_i|\boldsymbol{X}_i) \quad = \quad \lambda_i(1 + \sigma^2\lambda_i)$$

2) Run in `R` using

```
library(MASS)
out<- glm.nb(Y~X)
```

Clustered Standard Errors

Misspecification and Quasi-MLE

- When the model is exactly correct, MLE is the best estimator
- But your model is usually wrong!
- What if you assumed $f(Y \mid \theta)$ but the true DGP is $g(Y \mid \theta)$?

Point estimate:

- Generally, $\hat{\theta}$ maximizing $L_f = f$ is inconsistent: $\lim_{n\to\infty} \hat{\theta} \to^p \theta^* \neq \theta$
- Instead, θ^* minimizes the "divergence" between f and g, defined as:

$$\mathrm{E}\left[\log g(Y \mid \theta) - \log f(Y \mid \theta)\right]$$

  $\longrightarrow$ "best possible" assuming g, but no guarantee θ^* is *substantively* close to θ

- This $\hat{\theta}$ is called the quasi-maximum likelihood estimator (QMLE)
- However, for some models and types of misspecification, $\hat{\theta}_{QMLE}$ is still consistent for true θ

Misspecification and Quasi-MLE

- When the model is exactly correct, MLE is the best estimator
- But your model is usually wrong!
- What if you assumed $f(Y \mid \theta)$ but the true DGP is $g(Y \mid \theta)$?

Point estimate:

- Generally, $\hat{\theta}$ maximizing $L_f = f$ is inconsistent: $\lim_{n\to\infty} \hat{\theta} \to^p \theta^* \neq \theta$
- Instead, θ^* minimizes the "divergence" between f and g, defined as:

$$\mathrm{E}\left[\log g(Y \mid \theta) - \log f(Y \mid \theta)\right]$$

$\longrightarrow$ "best possible" assuming g, but no guarantee θ^* is *substantively* close to θ

- This $\hat{\theta}$ is called the quasi-maximum likelihood estimator (QMLE)
- However, for some models and types of misspecification, $\hat{\theta}_{QMLE}$ is still consistent for true θ

Misspecification and Quasi-MLE

- When the model is exactly correct, MLE is the best estimator
- But your model is usually wrong!
- What if you assumed $f(Y \mid \theta)$ but the true DGP is $g(Y \mid \theta)$?

Point estimate:

- Generally, $\hat{\theta}$ maximizing $L_f = f$ is inconsistent: $\lim_{n\to\infty} \hat{\theta} \to^p \theta^* \neq \theta$
- Instead, θ^* minimizes the "divergence" between f and g, defined as:

$$\mathrm{E}\left[\log g(Y \mid \theta) - \log f(Y \mid \theta)\right]$$

  $\longrightarrow$ "best possible" assuming g, but no guarantee θ^* is *substantively* close to θ

- This $\hat{\theta}$ is called the quasi-maximum likelihood estimator (QMLE)
- However, for some models and types of misspecification, $\hat{\theta}_{QMLE}$ is still consistent for true θ

Misspecification and Quasi-MLE

- When the model is exactly correct, MLE is the best estimator
- But your model is usually wrong!
- What if you assumed $f(Y \mid \theta)$ but the true DGP is $g(Y \mid \theta)$?

Point estimate:

- Generally, $\hat{\theta}$ maximizing $L_f = f$ is inconsistent: $\lim_{n \to \infty} \hat{\theta} \to^p \theta^* \neq \theta$
- Instead, θ^* minimizes the "divergence" between f and g, defined as:

$$\mathsf{E}\left[\log g(Y \mid \theta) - \log f(Y \mid \theta)\right]$$

  $\longrightarrow$ "best possible" assuming g, but no guarantee θ^* is *substantively* close to θ
- This $\hat{\theta}$ is called the quasi-maximum likelihood estimator (QMLE)
- However, for some models and types of misspecification, $\hat{\theta}_{QMLE}$ is still consistent for true θ

Misspecification and Quasi-MLE

- When the model is exactly correct, MLE is the best estimator
- But your model is usually wrong!
- What if you assumed $f(Y \mid \theta)$ but the true DGP is $g(Y \mid \theta)$?

Point estimate:

- Generally, $\hat{\theta}$ maximizing $L_f = f$ is inconsistent: $\lim_{n\rightarrow\infty} \hat{\theta} \rightarrow^p \theta^* \neq \theta$
- Instead, θ^* minimizes the "divergence" between f and g, defined as:

$$\mathsf{E}\left[\log g(Y \mid \theta) - \log f(Y \mid \theta)\right]$$

$\longrightarrow$ "best possible" assuming g, but no guarantee θ^* is *substantively* close to θ
- This $\hat{\theta}$ is called the quasi-maximum likelihood estimator (QMLE)
- However, for some models and types of misspecification, $\hat{\theta}_{QMLE}$ is still consistent for true θ

Misspecification and Quasi-MLE

- When the model is exactly correct, MLE is the best estimator
- But your model is usually wrong!
- What if you assumed $f(Y \mid \theta)$ but the true DGP is $g(Y \mid \theta)$?

Point estimate:

- Generally, $\hat{\theta}$ maximizing $L_f = f$ is inconsistent: $\lim_{n\to\infty} \hat{\theta} \to^p \theta^* \neq \theta$
- Instead, θ^* minimizes the "divergence" between f and g, defined as:

$$\mathsf{E}\left[\log g(Y \mid \theta) - \log f(Y \mid \theta)\right]$$

$\longrightarrow$ "best possible" assuming g, but no guarantee θ^* is *substantively* close to θ

- This $\hat{\theta}$ is called the quasi-maximum likelihood estimator (QMLE)
- However, for some models and types of misspecification, $\hat{\theta}_{QMLE}$ is still consistent for true θ

Misspecification and Quasi-MLE

- When the model is exactly correct, MLE is the best estimator
- But your model is usually wrong!
- What if you assumed $f(Y \mid \theta)$ but the true DGP is $g(Y \mid \theta)$?

Point estimate:

- Generally, $\hat{\theta}$ maximizing $L_f = f$ is inconsistent: $\lim_{n\to\infty} \hat{\theta} \to^p \theta^* \neq \theta$
- Instead, θ^* minimizes the "divergence" between f and g, defined as:

$$\mathsf{E}\left[\log g(Y \mid \theta) - \log f(Y \mid \theta)\right]$$

  $\longrightarrow$ "best possible" assuming g, but no guarantee θ^* is *substantively* close to θ

- This $\hat{\theta}$ is called the quasi-maximum likelihood estimator (QMLE)
- However, for some models and types of misspecification, $\hat{\theta}_{QMLE}$ is still consistent for true θ

Misspecification and Quasi-MLE

- When the model is exactly correct, MLE is the best estimator
- But your model is usually wrong!
- What if you assumed $f(Y \mid \theta)$ but the true DGP is $g(Y \mid \theta)$?

Point estimate:

- Generally, $\hat{\theta}$ maximizing $L_f = f$ is inconsistent: $\lim_{n \to \infty} \hat{\theta} \to^p \theta^* \neq \theta$
- Instead, θ^* minimizes the "divergence" between f and g, defined as:

$$\mathsf{E}\left[\log g(Y \mid \theta) - \log f(Y \mid \theta)\right]$$

  $\longrightarrow$ "best possible" assuming g, but no guarantee θ^* is *substantively* close to θ
- This $\hat{\theta}$ is called the quasi-maximum likelihood estimator (QMLE)
- However, for some models and types of misspecification, $\hat{\theta}_{QMLE}$ is still consistent for true θ

Robust Standard Errors

Variance estimate:

- For $\hat{\theta}_{QMLE}$ we can show

$$\hat{\theta} \overset{\text{approx.}}{\sim} \mathsf{N}\left(\theta^*,\ A^{-1}BA^{-1}\right)$$

where $\quad A = -\mathsf{E}[H(\theta^*)] \quad$ and $\quad B = \mathsf{E}[s(\theta^*)s(\theta^*)^\top]$

- If $f = g$, $\theta^* = \theta$ (consistency) and $A = B$ (information equality) $\longrightarrow$ We get $\hat{\theta}_{QMLE} \overset{\text{approx.}}{\sim} \mathsf{N}(\theta, A^{-1})$
- Thus, as expected, $\hat{\theta}_{QMLE} = \hat{\theta}_{MLE}$

- If $f \neq g$, usually $A \neq B$
- So we need to estimate A and B separately and use a sandwich estimator for variance:

$$\mathsf{V}(\widehat{\hat{\theta}_{QMLE}}) \;=\; \hat{A}^{-1}\hat{B}\hat{A}^{-1}$$

where $\quad \hat{A} = -\sum_{i=1}^n H_i(\hat{\theta}) \quad$ and $\quad \hat{B} = \sum_{i=1}^n s_i(\hat{\theta})s_i(\hat{\theta})^\top$

- This gives the Huber-White robust standard errors

Robust Standard Errors

Variance estimate:

- For $\hat{\theta}_{QMLE}$ we can show

$$\hat{\theta} \overset{\text{approx.}}{\sim} \mathsf{N}\left(\theta^*,\ A^{-1}BA^{-1}\right)$$

where $\quad A = -\mathsf{E}[H(\theta^*)] \quad$ and $\quad B = \mathsf{E}[s(\theta^*)s(\theta^*)^\top]$

- If $f = g$, $\theta^* = \theta$ (consistency) and $A = B$ (information equality) $\longrightarrow$ We get $\hat{\theta}_{QMLE} \overset{\text{approx.}}{\sim} \mathsf{N}(\theta, A^{-1})$
- Thus, as expected, $\hat{\theta}_{QMLE} = \hat{\theta}_{MLE}$
- If $f \neq g$, usually $A \neq B$
- So we need to estimate A and B separately and use a sandwich estimator for variance:

$$\widehat{\mathsf{V}(\hat{\theta}_{QMLE})} \ = \ \hat{A}^{-1}\hat{B}\hat{A}^{-1}$$

where $\quad \hat{A} = -\sum_{i=1}^{n} H_i(\hat{\theta}) \quad$ and $\quad \hat{B} = \sum_{i=1}^{n} s_i(\hat{\theta})s_i(\hat{\theta})^\top$

- This gives the Huber-White robust standard errors

Robust Standard Errors

Variance estimate:

- For $\hat{\theta}_{QMLE}$ we can show

$$\hat{\theta} \stackrel{\text{approx.}}{\sim} \mathsf{N}\left(\theta^*,\ A^{-1}BA^{-1}\right)$$

where $\quad A = -\mathsf{E}[H(\theta^*)] \quad$ and $\quad B = \mathsf{E}[s(\theta^*)s(\theta^*)^\top]$

- If $f = g$, $\theta^* = \theta$ (consistency) and $A = B$ (information equality) $\longrightarrow$ We get $\hat{\theta}_{QMLE} \stackrel{\text{approx.}}{\sim} \mathsf{N}(\theta, A^{-1})$
- Thus, as expected, $\hat{\theta}_{QMLE} = \hat{\theta}_{MLE}$

- If $f \neq g$, usually $A \neq B$
- So we need to estimate A and B separately and use a sandwich estimator for variance:

$$\widehat{\mathsf{V}(\hat{\theta}_{QMLE})} = \hat{A}^{-1}\hat{B}\hat{A}^{-1}$$

where $\quad \hat{A} = -\sum_{i=1}^{n} H_i(\hat{\theta}) \quad$ and $\quad \hat{B} = \sum_{i=1}^{n} s_i(\hat{\theta})s_i(\hat{\theta})^\top$

- This gives the Huber-White robust standard errors

Robust Standard Errors

Variance estimate:

- For $\hat{\theta}_{QMLE}$ we can show

$$\hat{\theta} \overset{\text{approx.}}{\sim} \mathsf{N}\left(\theta^*,\ A^{-1}BA^{-1}\right)$$

where $\quad A = -\mathsf{E}[H(\theta^*)] \quad$ and $\quad B = \mathsf{E}[s(\theta^*)s(\theta^*)^\top]$

- If $f = g$, $\theta^* = \theta$ (consistency) and $A = B$ (information equality) $\longrightarrow$ We get $\hat{\theta}_{QMLE} \overset{\text{approx.}}{\sim} \mathsf{N}(\theta, A^{-1})$
- Thus, as expected, $\hat{\theta}_{QMLE} = \hat{\theta}_{MLE}$

- If $f \neq g$, usually $A \neq B$
- So we need to estimate A and B separately and use a sandwich estimator for variance:

$$\widehat{\mathsf{V}(\hat{\theta}_{QMLE})} = \hat{A}^{-1}\hat{B}\hat{A}^{-1}$$

where $\quad \hat{A} = -\sum_{i=1}^{n} H_i(\hat{\theta}) \quad$ and $\quad \hat{B} = \sum_{i=1}^{n} s_i(\hat{\theta})s_i(\hat{\theta})^\top$

- This gives the Huber-White robust standard errors

Robust Standard Errors

Variance estimate:

- For $\hat{\theta}_{QMLE}$ we can show

$$\hat{\theta} \overset{\text{approx.}}{\sim} \mathsf{N}\left(\theta^*,\ A^{-1}BA^{-1}\right)$$

where $\quad A = -\mathsf{E}[H(\theta^*)] \quad$ and $\quad B = \mathsf{E}[s(\theta^*)s(\theta^*)^\top]$

- If $f = g$, $\theta^* = \theta$ (consistency) and $A = B$ (information equality) $\longrightarrow$ We get $\hat{\theta}_{QMLE} \overset{\text{approx.}}{\sim} \mathsf{N}(\theta, A^{-1})$
- Thus, as expected, $\hat{\theta}_{QMLE} = \hat{\theta}_{MLE}$

- If $f \neq g$, usually $A \neq B$
- So we need to estimate A and B separately and use a sandwich estimator for variance:

$$\mathsf{V}(\widehat{\hat{\theta}_{QMLE}}) \ = \ \hat{A}^{-1}\hat{B}\hat{A}^{-1}$$

where $\quad \hat{A} = -\sum_{i=1}^n H_i(\hat{\theta}) \quad$ and $\quad \hat{B} = \sum_{i=1}^n s_i(\hat{\theta})s_i(\hat{\theta})^\top$

- This gives the Huber-White robust standard errors

Robust Standard Errors

Variance estimate:

- For $\hat{\theta}_{QMLE}$ we can show

$$\hat{\theta} \overset{\text{approx.}}{\sim} \mathsf{N}\left(\theta^*,\ A^{-1}BA^{-1}\right)$$

where $\quad A = -\mathsf{E}[H(\theta^*)] \quad$ and $\quad B = \mathsf{E}[s(\theta^*)s(\theta^*)^\top]$

- If $f = g$, $\theta^* = \theta$ (consistency) and $A = B$ (information equality) $\longrightarrow$ We get $\hat{\theta}_{QMLE} \overset{\text{approx.}}{\sim} \mathsf{N}(\theta, A^{-1})$
- Thus, as expected, $\hat{\theta}_{QMLE} = \hat{\theta}_{MLE}$

- If $f \neq g$, usually $A \neq B$
- So we need to estimate A and B separately and use a sandwich estimator for variance:

$$\mathsf{V}\widehat{(\hat{\theta}_{QMLE})} \ = \ \hat{A}^{-1}\hat{B}\hat{A}^{-1}$$

where $\quad \hat{A} = -\sum_{i=1}^n H_i(\hat{\theta}) \quad$ and $\quad \hat{B} = \sum_{i=1}^n s_i(\hat{\theta})s_i(\hat{\theta})^\top$

- This gives the Huber-White robust standard errors

Standard Errors for Cluster Sampled Data

- Suppose that data are collected via cluster sampling, i.e., first sampling M clusters and then N_m within each cluster m
- There may be dependence between units within each cluster
- The correct likelihood would take into account this dependence:

$$l(\theta \mid Y) \ = \ \sum_{m=1}^{M} f_m(Y_{1m}, ..., Y_{N_m m}|\theta)$$

- This joint likelihood will be intractable for most models (e.g. logit)
- Instead, we could look at a quasi-log-likelihood:

$$l^*(\theta \mid Y) \ = \ \sum_{m=1}^{M} \sum_{i=1}^{N_m} f_i(Y_i \mid \theta)$$

And "fix" the standard errors by allowing within-cluster correlation:

$$\hat{B}^* = \sum_{m=1}^{M} \left\{ \left(\sum_{i=1}^{N_m} s_i(\hat{\theta}) \right) \left(\sum_{i=1}^{N_m} s_i(\hat{\theta}) \right)' \right\}$$

- $\hat{A}^{-1} \hat{B}^* \hat{A}^{-1}$ gives the cluster robust standard errors
- Asymptotics is w.r.t. $M \Rightarrow$ may be badly behaved when M small

Standard Errors for Cluster Sampled Data

- Suppose that data are collected via cluster sampling, i.e., first sampling M clusters and then N_m within each cluster m
- There may be dependence between units within each cluster
- The correct likelihood would take into account this dependence:

$$l(\theta \mid Y) \; = \; \sum_{m=1}^{M} f_m(Y_{1m}, ..., Y_{N_m m}|\theta)$$

- This joint likelihood will be intractable for most models (e.g. logit)
- Instead, we could look at a quasi-log-likelihood:

$$l^*(\theta \mid Y) \; = \; \sum_{m=1}^{M} \sum_{i=1}^{N_m} f_i(Y_i \mid \theta)$$

And "fix" the standard errors by allowing within-cluster correlation:

$$\hat{B}^* = \sum_{m=1}^{M} \left\{ \left(\sum_{i=1}^{N_m} s_i(\hat{\theta}) \right) \left(\sum_{i=1}^{N_m} s_i(\hat{\theta}) \right)' \right\}$$

- $\hat{A}^{-1}\hat{B}^*\hat{A}^{-1}$ gives the cluster robust standard errors
- Asymptotics is w.r.t. $M \Rightarrow$ may be badly behaved when M small

Standard Errors for Cluster Sampled Data

- Suppose that data are collected via cluster sampling, i.e., first sampling M clusters and then N_m within each cluster m
- There may be dependence between units within each cluster
- The correct likelihood would take into account this dependence:

$$l(\theta \mid Y) \;=\; \textstyle\sum_{m=1}^{M} f_m(Y_{1m}, ..., Y_{N_m m}|\theta)$$

- This joint likelihood will be intractable for most models (e.g. logit)
- Instead, we could look at a quasi-log-likelihood:

$$l^*(\theta \mid Y) \;=\; \textstyle\sum_{m=1}^{M} \sum_{i=1}^{N_m} f_i(Y_i \mid \theta)$$

And "fix" the standard errors by allowing within-cluster correlation:

$$\hat{B}^* = \sum_{m=1}^{M} \left\{ \left(\sum_{i=1}^{N_m} s_i(\hat{\theta}) \right) \left(\sum_{i=1}^{N_m} s_i(\hat{\theta}) \right)' \right\}$$

- $\hat{A}^{-1}\hat{B}^*\hat{A}^{-1}$ gives the cluster robust standard errors
- Asymptotics is w.r.t. $M \Rightarrow$ may be badly behaved when M small

Standard Errors for Cluster Sampled Data

- Suppose that data are collected via cluster sampling, i.e., first sampling M clusters and then N_m within each cluster m
- There may be dependence between units within each cluster
- The correct likelihood would take into account this dependence:

$$l(\theta \mid Y) \ = \ \sum_{m=1}^{M} f_m(Y_{1m}, ..., Y_{N_m m}|\theta)$$

- This joint likelihood will be intractable for most models (e.g. logit)
- Instead, we could look at a quasi-log-likelihood:

$$l^*(\theta \mid Y) \ = \ \sum_{m=1}^{M} \sum_{i=1}^{N_m} f_i(Y_i \mid \theta)$$

  And "fix" the standard errors by allowing within-cluster correlation:

$$\hat{B}^* = \sum_{m=1}^{M} \left\{ \left(\sum_{i=1}^{N_m} s_i(\hat{\theta}) \right) \left(\sum_{i=1}^{N_m} s_i(\hat{\theta}) \right)' \right\}$$

- $\hat{A}^{-1}\hat{B}^*\hat{A}^{-1}$ gives the cluster robust standard errors
- Asymptotics is w.r.t. $M \Rightarrow$ may be badly behaved when M small

Standard Errors for Cluster Sampled Data

- Suppose that data are collected via cluster sampling, i.e., first sampling M clusters and then N_m within each cluster m
- There may be dependence between units within each cluster
- The correct likelihood would take into account this dependence:

$$l(\theta \mid Y) \ = \ \sum_{m=1}^{M} f_m(Y_{1m}, ..., Y_{N_m m}|\theta)$$

- This joint likelihood will be intractable for most models (e.g. logit)
- Instead, we could look at a quasi-log-likelihood:

$$l^*(\theta \mid Y) \ = \ \sum_{m=1}^{M} \sum_{i=1}^{N_m} f_i(Y_i \mid \theta)$$

And "fix" the standard errors by allowing within-cluster correlation:

$$\hat{B}^* = \sum_{m=1}^{M} \left\{ \left(\sum_{i=1}^{N_m} s_i(\hat{\theta}) \right) \left(\sum_{i=1}^{N_m} s_i(\hat{\theta}) \right)' \right\}$$

- $\hat{A}^{-1}\hat{B}^*\hat{A}^{-1}$ gives the cluster robust standard errors
- Asymptotics is w.r.t. $M \Rightarrow$ may be badly behaved when M small

Standard Errors for Cluster Sampled Data

- Suppose that data are collected via cluster sampling, i.e., first sampling M clusters and then N_m within each cluster m
- There may be dependence between units within each cluster
- The correct likelihood would take into account this dependence:

$$l(\theta \mid Y) \ = \ \sum_{m=1}^{M} f_m(Y_{1m}, ..., Y_{N_m m}|\theta)$$

- This joint likelihood will be intractable for most models (e.g. logit)
- Instead, we could look at a quasi-log-likelihood:

$$l^*(\theta \mid Y) \ = \ \sum_{m=1}^{M} \sum_{i=1}^{N_m} f_i(Y_i \mid \theta)$$

  And "fix" the standard errors by allowing within-cluster correlation:

$$\hat{B}^* = \sum_{m=1}^{M} \left\{ \left(\sum_{i=1}^{N_m} s_i(\hat{\theta}) \right) \left(\sum_{i=1}^{N_m} s_i(\hat{\theta}) \right)' \right\}$$

- $\hat{A}^{-1}\hat{B}^*\hat{A}^{-1}$ gives the cluster robust standard errors
- Asymptotics is w.r.t. $M \Rightarrow$ may be badly behaved when M small

Standard Errors for Cluster Sampled Data

- Suppose that data are collected via cluster sampling, i.e., first sampling M clusters and then N_m within each cluster m
- There may be dependence between units within each cluster
- The correct likelihood would take into account this dependence:

$$l(\theta \mid Y) \ = \ \sum_{m=1}^{M} f_m(Y_{1m}, ..., Y_{N_m m}|\theta)$$

- This joint likelihood will be intractable for most models (e.g. logit)
- Instead, we could look at a quasi-log-likelihood:

$$l^*(\theta \mid Y) \ = \ \sum_{m=1}^{M} \sum_{i=1}^{N_m} f_i(Y_i \mid \theta)$$

  And "fix" the standard errors by allowing within-cluster correlation:

$$\hat{B}^* = \sum_{m=1}^{M} \left\{ \left(\sum_{i=1}^{N_m} s_i(\hat{\theta}) \right) \left(\sum_{i=1}^{N_m} s_i(\hat{\theta}) \right)' \right\}$$

- $\hat{A}^{-1} \hat{B}^* \hat{A}^{-1}$ gives the cluster robust standard errors
- Asymptotics is w.r.t. M $\Rightarrow$ may be badly behaved when M small

Survival analysis

What is Survival Analysis?

- Analyze the length of time spent in a given state
- $Y_i \in [0, \infty)$: Duration, "time to an event"
- Suppose Y_i has density $f(y)$.
- Example: Cabinet duration
  - Are cabinets more likely to dissolve early or late?
  - What factors predict the length of time until dissolution?
  - King, Alt, Burns & Laver (1990 AJPS) .... Exponential model
  - Warwick & Easton (1992 AJPS) .... Weibull model
  - Warwick (1992 AJPS) .... Cox PH model
  - Diermeier & Stevenson (1999 AJPS) .... Competing risks model
- One of the most sophisticated subfields of statistical modeling, developed in multiple disciplines
- We will only be able to scratch the surface

What is Survival Analysis?

- Analyze the length of time spent in a given state
- $Y_i \in [0, \infty)$: Duration, "time to an event"
- Suppose Y_i has density $f(y)$.
- Example: Cabinet duration
  - Are cabinets more likely to dissolve early or late?
  - What factors predict the length of time until dissolution?
  - King, Alt, Burns & Laver (1990 AJPS) .... Exponential model
  - Warwick & Easton (1992 AJPS) .... Weibull model
  - Warwick (1992 AJPS) .... Cox PH model
  - Diermeier & Stevenson (1999 AJPS) .... Competing risks model
- One of the most sophisticated subfields of statistical modeling, developed in multiple disciplines
- We will only be able to scratch the surface

What is Survival Analysis?

- Analyze the length of time spent in a given state
- $Y_i \in [0, \infty)$: Duration, "time to an event"
- Suppose Y_i has density $f(y)$.
- Example: Cabinet duration
  - Are cabinets more likely to dissolve early or late?
  - What factors predict the length of time until dissolution?
  - King, Alt, Burns & Laver (1990 AJPS) .... Exponential model
  - Warwick & Easton (1992 AJPS) .... Weibull model
  - Warwick (1992 AJPS) .... Cox PH model
  - Diermeier & Stevenson (1999 AJPS) .... Competing risks model
- One of the most sophisticated subfields of statistical modeling, developed in multiple disciplines
- We will only be able to scratch the surface

What is Survival Analysis?

- Analyze the length of time spent in a given state
- $Y_i \in [0, \infty)$: Duration, "time to an event"
- Suppose Y_i has density $f(y)$.
- Example: Cabinet duration
  - Are cabinets more likely to dissolve early or late?
  - What factors predict the length of time until dissolution?
  - King, Alt, Burns & Laver (1990 AJPS) .... Exponential model
  - Warwick & Easton (1992 AJPS) .... Weibull model
  - Warwick (1992 AJPS) .... Cox PH model
  - Diermeier & Stevenson (1999 AJPS) .... Competing risks model
- One of the most sophisticated subfields of statistical modeling, developed in multiple disciplines
- We will only be able to scratch the surface

Survival Function

- Survival function: Probability of surviving at least up to time y

$$S(y) \equiv \Pr(Y_i > y) = \int_y^{\infty} f(y)dy = 1 - \int_0^y f(y)dy = 1 - F(y)$$

- How likely am I to live at least y years?
- Properties:
  - $S(0) = 1$ and $S(\infty) = 0$; monotonically decreasing
  - Area under $S(y)$ is the average survival time:

$$\begin{aligned} \mathrm{E}(Y_i) &= \int_0^{\infty} y f(y)dy \\ &= y\left(F(y)|_0^{\infty}\right) - \int_0^{\infty} F(y)dy \\ &= \int_0^{\infty} (1 - F(y))dy \\ &= \int_0^{\infty} S(y)dy \end{aligned}$$

Survival Function

- Survival function: Probability of surviving at least up to time y

$$S(y) \equiv \Pr(Y_i > y) = \int_y^{\infty} f(y)dy = 1 - \int_0^y f(y)dy = 1 - F(y)$$

- How likely am I to live at least y years?
- Properties:
  - $S(0) = 1$ and $S(\infty) = 0$; monotonically decreasing
  - Area under $S(y)$ is the average survival time:

$$\begin{aligned} \mathsf{E}(Y_i) &= \int_0^{\infty} y f(y) dy \\ &= y \left(F(y)|_0^{\infty}\right) - \int_0^{\infty} F(y)dy \\ &= \int_0^{\infty} (1 - F(y))dy \\ &= \int_0^{\infty} S(y)dy \end{aligned}$$

Survival Function

- Survival function: Probability of surviving at least up to time y

$$S(y) \equiv \Pr(Y_i > y) = \int_y^{\infty} f(y)dy = 1 - \int_0^y f(y)dy = 1 - F(y)$$

- How likely am I to live at least y years?
- Properties:
  - $S(0) = 1$ and $S(\infty) = 0$; monotonically decreasing
  - Area under $S(y)$ is the average survival time:

$$\begin{aligned}
\mathsf{E}(Y_i) &= \int_0^{\infty} y f(y)dy \\
&= y\left(F(y)|_0^{\infty}\right) - \int_0^{\infty} F(y)dy \\
&= \int_0^{\infty} (1 - F(y))dy \\
&= \int_0^{\infty} S(y)dy
\end{aligned}$$

Survival Function

- Survival function: Probability of surviving at least up to time y

$$S(y) \equiv \Pr(Y_i > y) = \int_y^{\infty} f(y)dy = 1 - \int_0^y f(y)dy = 1 - F(y)$$

- How likely am I to live at least y years?
- Properties:
  - $S(0) = 1$ and $S(\infty) = 0$; monotonically decreasing
  - Area under $S(y)$ is the average survival time:

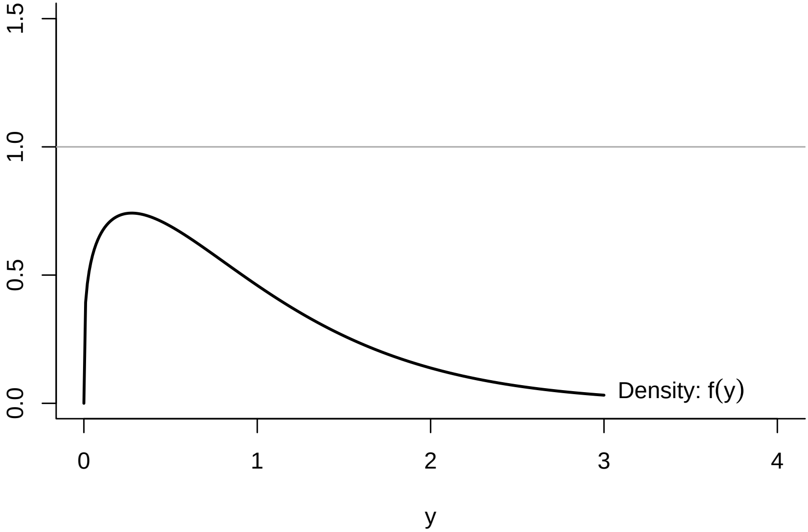
$$\begin{aligned}
\mathsf{E}(Y_i) &= \int_0^{\infty} y f(y)dy \\
&= y\left(F(y)|_0^{\infty}\right) - \int_0^{\infty} F(y)dy \\
&= \int_0^{\infty} (1 - F(y))dy \\
&= \int_0^{\infty} S(y)dy
\end{aligned}$$

Survival Function

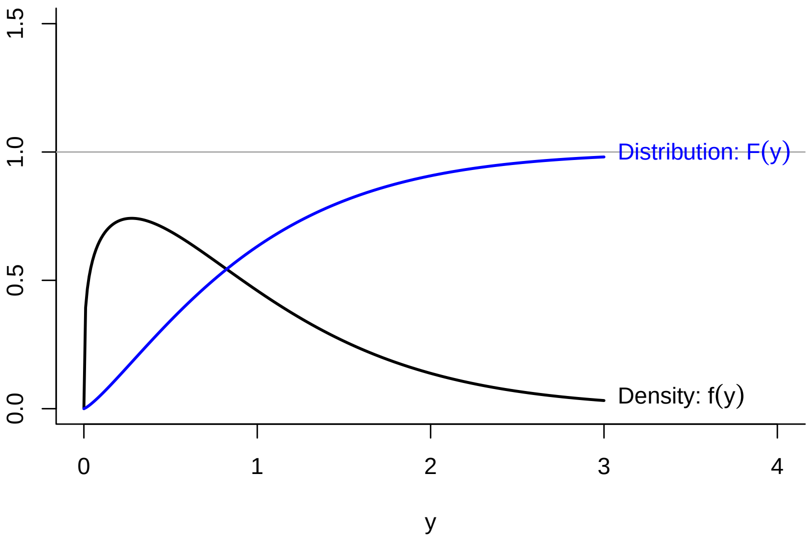
- One-to-one relationships with density and probability:

$$f(y) = -\frac{d}{dy}S(y) \quad \text{and} \quad S(y) = \int_y^{\infty} f(t)dt$$
$$\Pr(y \leq Y_i < y + h) = S(y) - S(y+h)$$

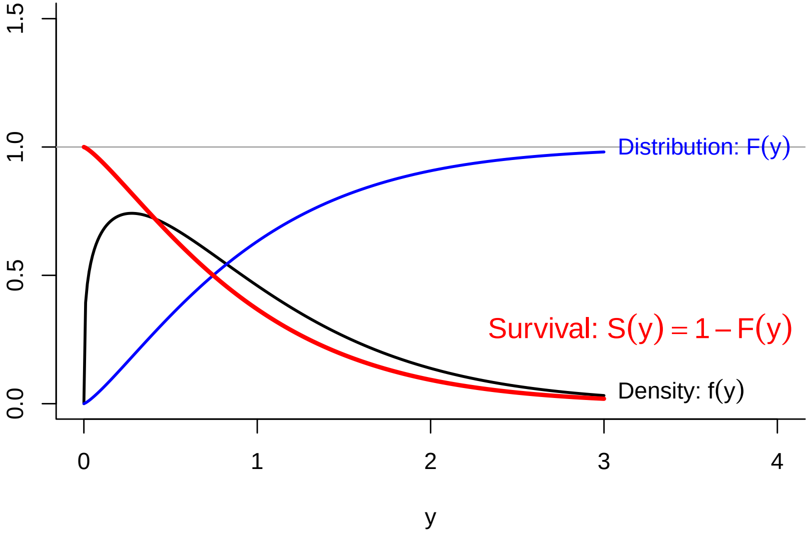
Survival Function



Survival Function



Survival Function



Hazard Function

- Hazard function: Instantaneous rate of leaving a state at time t conditional on survival up to that time

$$\lambda(y) \equiv \lim_{h \downarrow 0} \frac{\Pr(y \leq Y_i < y + h \mid Y_i \geq y)}{h} = \frac{f(y)}{S(y)}$$

- "Force of mortality" — what is the 'risk' that I die at time y given that I have lived up until y?
- Difficult to directly interpret, but useful for model checking, etc.
- One-to-one relationship with survival function:

$$\lambda(y) = -\frac{d}{dy} \log S(y) \quad \text{and} \quad S(y) = \exp\left(-\int_0^y \lambda(t) dt\right)$$

Hazard Function

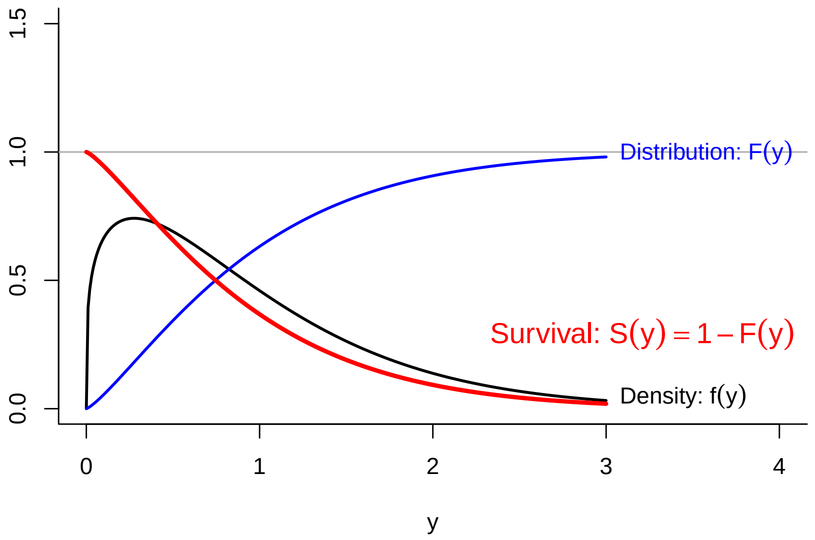
- Hazard function: Instantaneous rate of leaving a state at time t conditional on survival up to that time

$$\lambda(y) \equiv \lim_{h \downarrow 0} \frac{\Pr(y \leq Y_i < y + h \mid Y_i \geq y)}{h} = \frac{f(y)}{S(y)}$$

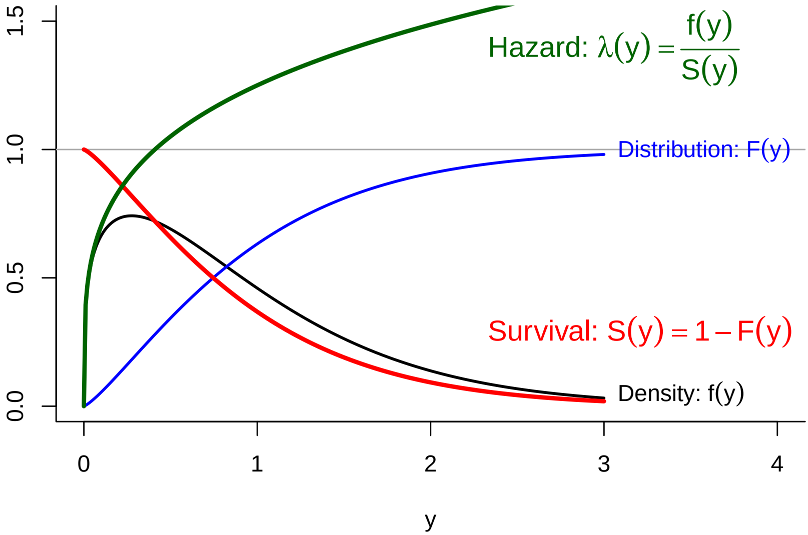
- "Force of mortality" — what is the 'risk' that I die at time y given that I have lived up until y?
- Difficult to directly interpret, but useful for model checking, etc.
- One-to-one relationship with survival function:

$$\lambda(y) = -\frac{d}{dy}\log S(y) \quad \text{and} \quad S(y) = \exp\left(-\int_0^y \lambda(t)dt\right)$$

Hazard Function



Hazard Function



Quantities of Interest

- Shape of the survival curve
- Expected (remaining) time to event (= life expectancy at age y):

$$\mu(y) \equiv \mathsf{E}(Y_i - y \mid Y_i > y) = \frac{1}{S(y)} \int_y^{\infty} S(t)dt$$

- Given that I'm alive at y, how much longer should I expect to live?
- Predicted differences in the above
- Causal effects on survival outcomes:
  - One-shot treatment administered at the beginning of study period — needs conditional ignorability given observed pre-trial covariates
  - Time-varying treatment, possibly given in response to covariates — needs "sequential ignorability"

Quantities of Interest

- Shape of the survival curve
- Expected (remaining) time to event (= life expectancy at age y):

$$\mu(y) \;\equiv\; \mathsf{E}(Y_i - y \mid Y_i > y) \;=\; \frac{1}{S(y)} \int_y^{\infty} S(t)dt$$

- Given that I'm alive at y, how much longer should I expect to live?
- Predicted differences in the above
- Causal effects on survival outcomes:
  - One-shot treatment administered at the beginning of study period — needs conditional ignorability given observed pre-trial covariates
  - Time-varying treatment, possibly given in response to covariates — needs "sequential ignorability"

Quantities of Interest

- Shape of the survival curve
- Expected (remaining) time to event (= life expectancy at age y):

$$\mu(y) \equiv \mathsf{E}(Y_i - y \mid Y_i > y) = \frac{1}{S(y)} \int_y^\infty S(t)dt$$

- Given that I'm alive at y, how much longer should I expect to live?
- Predicted differences in the above
- Causal effects on survival outcomes:
  - One-shot treatment administered at the beginning of study period — needs conditional ignorability given observed pre-trial covariates
  - Time-varying treatment, possibly given in response to covariates — needs "sequential ignorability"

Quantities of Interest

- Shape of the survival curve
- Expected (remaining) time to event (= life expectancy at age y):

$$\mu(y) \equiv \mathsf{E}(Y_i - y \mid Y_i > y) = \frac{1}{S(y)} \int_y^{\infty} S(t)dt$$

- Given that I'm alive at y, how much longer should I expect to live?
- Predicted differences in the above
- Causal effects on survival outcomes:
  - One-shot treatment administered at the beginning of study period — needs conditional ignorability given observed pre-trial covariates
  - Time-varying treatment, possibly given in response to covariates — needs "sequential ignorability"

Censoring

- Observation is right-censored when only the lower bound of duration is known: $Y_i \in (c, \infty)$
- The independent censoring assumption: Censored observations do not systematically differ from complete observations in terms of hazard rates
- A sufficient condition: $Y_i \perp\!\!\!\perp C_i \mid X_i$ where $C_i =$ time to censoring
- Either Y_i or C_i is actually observed
- Examples:
  - Random attrition of study sample
  - Study begins and ends at exogenously fixed calendar dates
  - Study ends after fixed duration (type I censoring)
  - Study ends after a fixed number of failures (type II censoring)
- Other types of censoring (left, interval) are less common

Censoring

- Observation is right-censored when only the lower bound of duration is known: $Y_i \in (c, \infty)$
- The independent censoring assumption: Censored observations do not systematically differ from complete observations in terms of hazard rates
- A sufficient condition: $Y_i \perp\!\!\!\perp C_i \mid X_i$ where $C_i =$ time to censoring
- Either Y_i or C_i is actually observed
- Examples:
  - Random attrition of study sample
  - Study begins and ends at exogenously fixed calendar dates
  - Study ends after fixed duration (type I censoring)
  - Study ends after a fixed number of failures (type II censoring)
- Other types of censoring (left, interval) are less common

Censoring

- Observation is right-censored when only the lower bound of duration is known: $Y_i \in (c, \infty)$
- The independent censoring assumption: Censored observations do not systematically differ from complete observations in terms of hazard rates
- A sufficient condition: $Y_i \perp\!\!\!\perp C_i \mid X_i$ where $C_i =$ time to censoring
- Either Y_i or C_i is actually observed
- Examples:
  - Random attrition of study sample
  - Study begins and ends at exogenously fixed calendar dates
  - Study ends after fixed duration (type I censoring)
  - Study ends after a fixed number of failures (type II censoring)
- Other types of censoring (left, interval) are less common

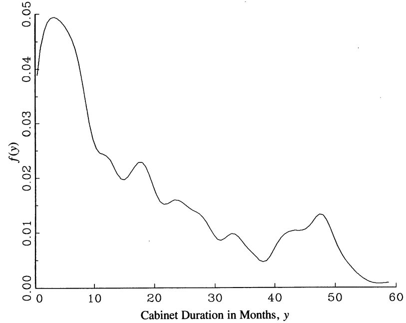
Censoring

- Observation is right-censored when only the lower bound of duration is known: $Y_i \in (c, \infty)$
- The independent censoring assumption: Censored observations do not systematically differ from complete observations in terms of hazard rates
- A sufficient condition: $Y_i \perp\!\!\!\perp C_i \mid X_i$ where $C_i =$ time to censoring
- Either Y_i or C_i is actually observed
- Examples:
  - Random attrition of study sample
  - Study begins and ends at exogenously fixed calendar dates
  - Study ends after fixed duration (type I censoring)
  - Study ends after a fixed number of failures (type II censoring)
- Other types of censoring (left, interval) are less common

Cabinet Duration Example: Censoring

King, Alt, Burns, and Laver (1990 AJPS):

- Y_i: Duration of parliamentary cabinets, $n = 314$

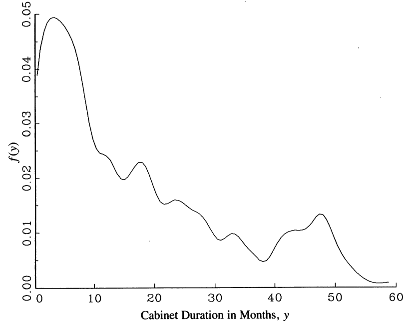


- Notice the "bump"?
- Some cabinets end their lives "naturally"
- Others end because of constitutional interelection periods (CIEP)
- King et al. treat CIEPs as censored observations
- But is this "censoring" independent?

Cabinet Duration Example: Censoring

King, Alt, Burns, and Laver (1990 AJPS):

- Y_i: Duration of parliamentary cabinets, $n = 314$

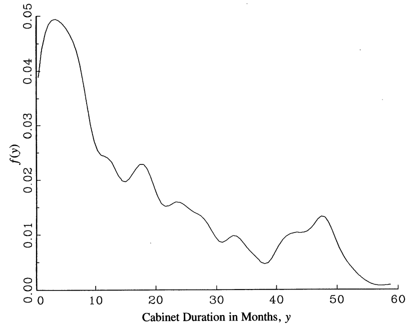


- Notice the "bump"?
- Some cabinets end their lives "naturally"
- Others end because of constitutional interelection periods (CIEP)
- King et al. treat CIEPs as censored observations
- But is this "censoring" independent?

Cabinet Duration Example: Censoring

King, Alt, Burns, and Laver (1990 AJPS):

- Y_i: Duration of parliamentary cabinets, $n = 314$



- Notice the "bump"?

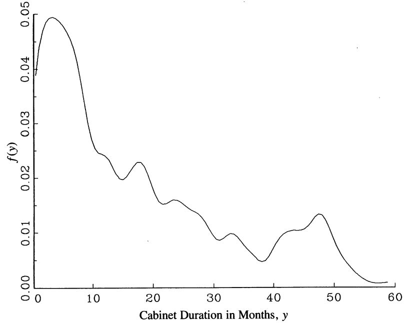
- Some cabinets end their lives "naturally"
- Others end because of constitutional interelection periods (CIEP)

- King et al. treat CIEPs as censored observations
- But is this "censoring" independent?

Cabinet Duration Example: Censoring

King, Alt, Burns, and Laver (1990 AJPS):

- Y_i: Duration of parliamentary cabinets, $n = 314$



- Notice the "bump"?
- Some cabinets end their lives "naturally"
- Others end because of constitutional interelection periods (CIEP)
- King et al. treat CIEPs as censored observations
- But is this "censoring" independent?

Discrete Time Approximation

- Time is continuous but we observe discrete time: $t_1 < t_2 < \cdots$
- Density function: $f(t_j) = \Pr(Y_i = t_j)$
- Survival function: $S(t_j) = \Pr(Y_i > t_j) = \sum_{\{k:\ t_k > t_j\}} f(t_k)$
- Hazard function: $\lambda(t_j) = \Pr(Y_i = t_j \mid Y_i \geq t_j) = f(t_j)/S(t_{j-1})$
- Key relationships:
  - $S(t_j) = \prod_{k=1}^{j}(1 - \lambda(t_k))$
  - $f(t_j) = S(t_{j-1}) - S(t_j) = \lambda(t_j) \prod_{k=1}^{j-1}(1 - \lambda(t_k))$
- Expected remaining time to event:

$$\begin{aligned} \mu(t_j) &= \mathsf{E}(Y_i - t_j \mid Y_i > t_j) \\ &= \frac{1}{S(t_j)} \sum_{k=j+1}^{\infty} (t_k - t_j) f(t_k) \end{aligned}$$

Discrete Time Approximation

- Time is continuous but we observe discrete time: $t_1 < t_2 < \cdots$
- Density function: $f(t_j) = \Pr(Y_i = t_j)$
- Survival function: $S(t_j) = \Pr(Y_i > t_j) = \sum_{\{k:\ t_k > t_j\}} f(t_k)$
- Hazard function: $\lambda(t_j) = \Pr(Y_i = t_j \mid Y_i \geq t_j) = f(t_j)/S(t_{j-1})$
- Key relationships:
  - $S(t_j) = \prod_{k=1}^{j}(1 - \lambda(t_k))$
  - $f(t_j) = S(t_{j-1}) - S(t_j) = \lambda(t_j)\prod_{k=1}^{j-1}(1 - \lambda(t_k))$
- Expected remaining time to event:

$$
\begin{aligned}
\mu(t_j) &= \mathsf{E}(Y_i - t_j \mid Y_i > t_j) \\
&= \frac{1}{S(t_j)} \sum_{k=j+1}^{\infty} (t_k - t_j) f(t_k)
\end{aligned}
$$

Discrete Time Approximation

- Time is continuous but we observe discrete time: $t_1 < t_2 < \cdots$
- Density function: $f(t_j) = \Pr(Y_i = t_j)$
- Survival function: $S(t_j) = \Pr(Y_i > t_j) = \sum_{\{k:\ t_k > t_j\}} f(t_k)$
- Hazard function: $\lambda(t_j) = \Pr(Y_i = t_j \mid Y_i \geq t_j) = f(t_j) / S(t_{j-1})$
- Key relationships:
  - $S(t_j) = \prod_{k=1}^{j}(1 - \lambda(t_k))$
  - $f(t_j) = S(t_{j-1}) - S(t_j) = \lambda(t_j)\prod_{k=1}^{j-1}(1 - \lambda(t_k))$
- Expected remaining time to event:

$$\begin{aligned}
\mu(t_j) &= \mathsf{E}(Y_i - t_j \mid Y_i > t_j) \\
&= \frac{1}{S(t_j)} \sum_{k=j+1}^{\infty} (t_k - t_j) f(t_k)
\end{aligned}$$

Discrete Time Approximation

- Time is continuous but we observe discrete time: $t_1 < t_2 < \cdots$
- Density function: $f(t_j) = \Pr(Y_i = t_j)$
- Survival function: $S(t_j) = \Pr(Y_i > t_j) = \sum_{\{k:\ t_k > t_j\}} f(t_k)$
- Hazard function: $\lambda(t_j) = \Pr(Y_i = t_j \mid Y_i \geq t_j) = f(t_j) \big/ S(t_{j-1})$
- Key relationships:
  - $S(t_j) = \prod_{k=1}^{j}(1 - \lambda(t_k))$
  - $f(t_j) = S(t_{j-1}) - S(t_j) = \lambda(t_j) \prod_{k=1}^{j-1}(1 - \lambda(t_k))$
- Expected remaining time to event:

$$\begin{aligned} \mu(t_j) &= \mathsf{E}(Y_i - t_j \mid Y_i > t_j) \\ &= \frac{1}{S(t_j)} \sum_{k=j+1}^{\infty} (t_k - t_j) f(t_k) \end{aligned}$$

Discrete Time Approximation

- Time is continuous but we observe discrete time: $t_1 < t_2 < \cdots$
- Density function: $f(t_j) = \Pr(Y_i = t_j)$
- Survival function: $S(t_j) = \Pr(Y_i > t_j) = \sum_{\{k:\ t_k > t_j\}} f(t_k)$
- Hazard function: $\lambda(t_j) = \Pr(Y_i = t_j \mid Y_i \geq t_j) = f(t_j) \big/ S(t_{j-1})$
- Key relationships:
  - $S(t_j) = \prod_{k=1}^{j}(1 - \lambda(t_k))$
  - $f(t_j) = S(t_{j-1}) - S(t_j) = \lambda(t_j) \prod_{k=1}^{j-1}(1 - \lambda(t_k))$
- Expected remaining time to event:

$$\begin{aligned} \mu(t_j) &= \mathsf{E}(Y_i - t_j \mid Y_i > t_j) \\ &= \frac{1}{S(t_j)} \sum_{k=j+1}^{\infty} (t_k - t_j) f(t_k) \end{aligned}$$

Discrete Time Approximation

- Time is continuous but we observe discrete time: $t_1 < t_2 < \cdots$
- Density function: $f(t_j) = \Pr(Y_i = t_j)$
- Survival function: $S(t_j) = \Pr(Y_i > t_j) = \sum_{\{k:\ t_k > t_j\}} f(t_k)$
- Hazard function: $\lambda(t_j) = \Pr(Y_i = t_j \mid Y_i \geq t_j) = f(t_j) \big/ S(t_{j-1})$
- Key relationships:
  - $S(t_j) = \prod_{k=1}^{j} (1 - \lambda(t_k))$
  - $f(t_j) = S(t_{j-1}) - S(t_j) = \lambda(t_j) \prod_{k=1}^{j-1} (1 - \lambda(t_k))$
- Expected remaining time to event:

$$
\begin{aligned}
\mu(t_j) &= \mathsf{E}(Y_i - t_j \mid Y_i > t_j) \\
&= \frac{1}{S(t_j)} \sum_{k=j+1}^{\infty} (t_k - t_j) f(t_k)
\end{aligned}
$$

Estimating the Survival Curve Without a Model

- Goal: Get the sense of what $S(t_j)$ looks like before introducing X_i
- Easy if no censoring; just count # of units failing at each t_j
- Censored observations make things a bit more complicated
- Setup:
  - Observed failure times: $t_1 < t_2 < \cdots < t_J$
  - $d_j =$ # of units that failed at time t_j
  - $m_j =$ # of units censored at time t_j
  - $r_j = \sum_{k=j}^{J}(d_k + m_k)$
    $=$ # of units at risk at time t_j, i.e., those that have neither failed nor been censored until right before t_j
- A natural estimate for the hazard function will then be:

$$\hat{\lambda}(t_j) \;=\; \widehat{\Pr}(Y_i = t_j \mid Y_i \geq t_j) \;=\; \frac{d_j}{r_j}$$

- In fact, this is the MLE of $\lambda(t_j)$

Estimating the Survival Curve Without a Model

- Goal: Get the sense of what $S(t_j)$ looks like before introducing X_i
- Easy if no censoring; just count # of units failing at each t_j
- Censored observations make things a bit more complicated
- Setup:
  - Observed failure times: $t_1 < t_2 < \cdots < t_J$
  - d_j = # of units that failed at time t_j
  - m_j = # of units censored at time t_j
  - $r_j = \sum_{k=j}^{J}(d_k + m_k)$
    = # of units at risk at time t_j, i.e., those that have neither failed nor been censored until right before t_j
- A natural estimate for the hazard function will then be:

$$\hat{\lambda}(t_j) \;=\; \widehat{\Pr}(Y_i = t_j \mid Y_i \geq t_j) \;=\; \frac{d_j}{r_j}$$

- In fact, this is the MLE of $\lambda(t_j)$

Estimating the Survival Curve Without a Model

- Goal: Get the sense of what $S(t_j)$ looks like before introducing X_i
- Easy if no censoring; just count # of units failing at each t_j
- Censored observations make things a bit more complicated
- Setup:
  - Observed failure times: $t_1 < t_2 < \cdots < t_J$
  - $d_j = \#$ of units that failed at time t_j
  - $m_j = \#$ of units censored at time t_j
  - $r_j = \sum_{k=j}^{J}(d_k + m_k)$
    $= \#$ of units at risk at time t_j, i.e., those that have neither failed nor been censored until right before t_j
- A natural estimate for the hazard function will then be:

$$\hat{\lambda}(t_j) \;=\; \widehat{\Pr}(Y_i = t_j \mid Y_i \geq t_j) \;=\; \frac{d_j}{r_j}$$

- In fact, this is the MLE of $\lambda(t_j)$

Estimating the Survival Curve Without a Model

- Goal: Get the sense of what $S(t_j)$ looks like before introducing X_i
- Easy if no censoring; just count # of units failing at each t_j
- Censored observations make things a bit more complicated
- Setup:
  - Observed failure times: $t_1 < t_2 < \cdots < t_J$
  - $d_j = \#$ of units that failed at time t_j
  - $m_j = \#$ of units censored at time t_j
  - $r_j = \sum_{k=j}^{J} (d_k + m_k)$
    $= \#$ of units at risk at time t_j, i.e., those that have neither failed nor been censored until right before t_j
- A natural estimate for the hazard function will then be:

$$\hat{\lambda}(t_j) \;=\; \widehat{\Pr}(Y_i = t_j \mid Y_i \geq t_j) \;=\; \frac{d_j}{r_j}$$

- In fact, this is the MLE of $\lambda(t_j)$

Kaplan-Meier Estimator

- This leads to the Kaplan-Meier estimator:

$$\widehat{S}(t_j) = \prod_{k=j}^{J}(1 - \hat{\lambda}(t_k)) = \prod_{k=j}^{J} \frac{r_k - d_k}{r_k}$$

- Using the MLE derivation for $\hat{\lambda}(t_j)$, we obtain the Hessian-based estimate of the asymptotic variance:

$$\widehat{\text{Var}(\widehat{S}(t_j))} = \widehat{S}^2(t_j) \sum_{k=j}^{J} \frac{d_k}{r_k(r_k - d_k)}$$

Kaplan-Meier Estimator

- This leads to the Kaplan-Meier estimator:

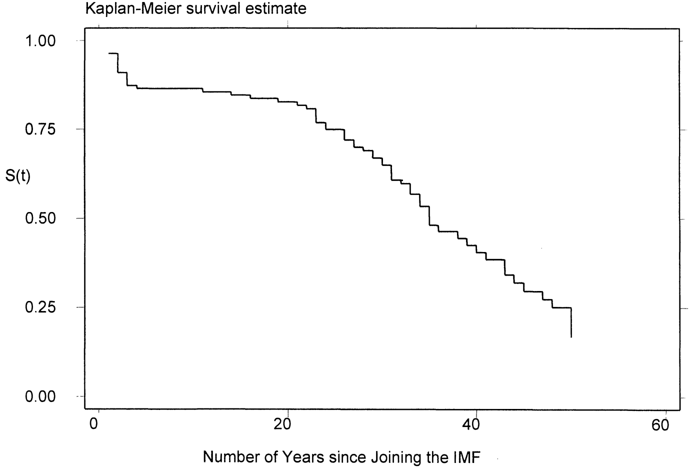
$$\widehat{S}(t_j) = \prod_{k=j}^{J}(1 - \hat{\lambda}(t_k)) = \prod_{k=j}^{J}\frac{r_k - d_k}{r_k}$$

- Using the MLE derivation for $\hat{\lambda}(t_j)$, we obtain the Hessian-based estimate of the asymptotic variance:

$$\widehat{\text{Var}(\widehat{S}(t_j))} = \widehat{S}^2(t_j)\sum_{k=j}^{J}\frac{d_k}{r_k(r_k - d_k)}$$

Example: Time Until Commitment to IMF Article VIII

FIGURE 2. The Kaplan-Meier Survival Function Duration of Article XIV Status over Time



Simmons (2000 APSR)

Exponential Regression Model

- Suppose that failures occur according to a Poisson process (i.e. continuously, independently, and with constant probability)
- Then the "time to an event" follows the exponential distribution

- Model: $Y_i \mid X_i \sim_{\text{ind}} \text{Exponential}(\mu_i)$ where $\mu_i = \exp(X_i'\beta)$
- Density: $f(y \mid \mu_i) = \frac{1}{\mu_i}\exp(-y/\mu_i)$
- Mean $\text{E}(Y_i \mid \mu_i) = \mu_i$ and Variance $\text{Var}(Y_i \mid \mu_i) = \mu_i^2$
- Survival function: $S(y) = \exp(-y/\mu_i)$
- Hazard function: $\lambda(y) = 1/\mu_i = \exp(-X_i'\beta)$ (constant in y)

- A common alternative parameterization: $\gamma_i \equiv 1/\mu_i$
- This changes nothing except that the sign of β gets reversed

- The constant hazard assumption: λ_i does not vary across time

Exponential Regression Model

- Suppose that failures occur according to a Poisson process (i.e. continuously, independently, and with constant probability)
- Then the "time to an event" follows the exponential distribution

- Model: $Y_i \mid X_i \sim_{\text{ind}} \text{Exponential}(\mu_i)$ where $\mu_i = \exp(X_i'\beta)$
- Density: $f(y \mid \mu_i) = \frac{1}{\mu_i}\exp(-y/\mu_i)$
- Mean $\text{E}(Y_i \mid \mu_i) = \mu_i$ and Variance $\text{Var}(Y_i \mid \mu_i) = \mu_i^2$
- Survival function: $S(y) = \exp(-y/\mu_i)$
- Hazard function: $\lambda(y) = 1/\mu_i = \exp(-X_i'\beta)$ (constant in y)

- A common alternative parameterization: $\gamma_i \equiv 1/\mu_i$
- This changes nothing except that the sign of β gets reversed

- The constant hazard assumption: λ_i does not vary across time

Exponential Regression Model

- Suppose that failures occur according to a Poisson process (i.e. continuously, independently, and with constant probability)
- Then the "time to an event" follows the exponential distribution

- Model: $Y_i \mid X_i \sim_{\text{ind}} \text{Exponential}(\mu_i)$ where $\mu_i = \exp(X_i^{'}\beta)$
- Density: $f(y \mid \mu_i) = \frac{1}{\mu_i}\exp(-y/\mu_i)$
- Mean $\text{E}(Y_i \mid \mu_i) = \mu_i$ and Variance $\text{Var}(Y_i \mid \mu_i) = \mu_i^2$
- Survival function: $S(y) = \exp(-y/\mu_i)$
- Hazard function: $\lambda(y) = 1/\mu_i = \exp(-X_i^{'}\beta)$ (constant in y)

- A common alternative parameterization: $\gamma_i \equiv 1/\mu_i$
- This changes nothing except that the sign of β gets reversed

- The constant hazard assumption: λ_i does not vary across time

Exponential Regression Model

- Suppose that failures occur according to a Poisson process (i.e. continuously, independently, and with constant probability)
- Then the "time to an event" follows the exponential distribution
- Model: $Y_i \mid X_i \sim_{\text{ind}} \text{Exponential}(\mu_i)$ where $\mu_i = \exp(X_i^{\prime}\beta)$
- Density: $f(y \mid \mu_i) = \frac{1}{\mu_i} \exp(-y/\mu_i)$
- Mean $\text{E}(Y_i \mid \mu_i) = \mu_i$ and Variance $\text{Var}(Y_i \mid \mu_i) = \mu_i^2$
- Survival function: $S(y) = \exp(-y/\mu_i)$
- Hazard function: $\lambda(y) = 1/\mu_i = \exp(-X_i^{\prime}\beta)$ (constant in y)
- A common alternative parameterization: $\gamma_i \equiv 1/\mu_i$
- This changes nothing except that the sign of β gets reversed
- The constant hazard assumption: λ_i does not vary across time

MLE for the Exponential Model with Censoring

- Censoring indicator: $D_i = 1$ if censored
- Y_i is the censoring time (rather than failure time) if $D_i = 1$
- Likelihood function:

$$
\begin{aligned}
L_n(\beta \mid Y, X, D) &= \prod_{i=1}^{n} \underbrace{\{f(Y_i \mid \mu_i)\}^{1-D_i}}_{uncensored} \cdot \underbrace{\{S(Y_i \mid \mu_i)\}^{D_i}}_{censored} \\
&= \prod_{i=1}^{n} \left\{ \frac{1}{\mu_i} \exp(-Y_i/\mu_i) \right\}^{1-D_i} \{\exp(-Y_i/\mu_i)\}^{D_i} \\
&= \prod_{i=1}^{n} \exp\left\{-(1-D_i) X_i^\top \beta\right\} \exp\left\{-\exp(-X_i'\beta) Y_i\right\}
\end{aligned}
$$

- Log-likelihood, score and Hessian can be calculated as usual

MLE for the Exponential Model with Censoring

- Censoring indicator: $D_i = 1$ if censored
- Y_i is the censoring time (rather than failure time) if $D_i = 1$

- Likelihood function:

$$
\begin{aligned}
L_n(\beta \mid Y, X, D) &= \prod_{i=1}^{n} \underbrace{\{f(Y_i \mid \mu_i)\}^{1-D_i}}_{uncensored} \cdot \underbrace{\{S(Y_i \mid \mu_i)\}^{D_i}}_{censored} \\
&= \prod_{i=1}^{n} \left\{ \frac{1}{\mu_i} \exp(-Y_i/\mu_i) \right\}^{1-D_i} \{\exp(-Y_i/\mu_i)\}^{D_i} \\
&= \prod_{i=1}^{n} \exp\left\{ -(1-D_i) X_i^{\top} \beta \right\} \exp\left\{ -\exp(-X_i^{\prime}\beta) Y_i \right\}
\end{aligned}
$$

- Log-likelihood, score and Hessian can be calculated as usual

MLE for the Exponential Model with Censoring

- Censoring indicator: $D_i = 1$ if censored
- Y_i is the censoring time (rather than failure time) if $D_i = 1$

- Likelihood function:

$$
\begin{aligned}
L_n(\beta \mid Y, X, D) &= \prod_{i=1}^{n} \underbrace{\{f(Y_i \mid \mu_i)\}^{1-D_i}}_{uncensored} \cdot \underbrace{\{S(Y_i \mid \mu_i)\}^{D_i}}_{censored} \\
&= \prod_{i=1}^{n} \left\{ \frac{1}{\mu_i} \exp(-Y_i/\mu_i) \right\}^{1-D_i} \{\exp(-Y_i/\mu_i)\}^{D_i} \\
&= \prod_{i=1}^{n} \exp\left\{ -(1-D_i) X_i^\top \beta \right\} \exp\left\{ -\exp(-X_i^\prime \beta) Y_i \right\}
\end{aligned}
$$

- Log-likelihood, score and Hessian can be calculated as usual

Weibull Regression Model

- The constant hazard assumption is often too restrictive
- The Weibull model relaxes the assumption by introducing a "shape" parameter
- Model: $Y_i \mid X_i \sim_{\text{ind}}$ Weibull(μ_i, α) where $\mu_i = \exp(X_i^\top \beta)$ and $\alpha > 0$
- Density: $f(y \mid \mu_i, \alpha) = \frac{\alpha}{\mu_i^\alpha} y^{\alpha-1} \exp\{-(y/\mu_i)^\alpha\}$
- Reduces to the exponential model when $\alpha = 1$
- Survival function: $S(y) = \exp\{-(y/\mu_i)^\alpha\}$
- Hazard function: $\lambda(y) = \frac{\alpha}{\mu_i^\alpha} y^{\alpha-1}$
- The monotonic hazard assumption: increasing (decreasing) if $\alpha > 1$ (if $\alpha < 1$)
- Other parametric regression models:
  - Gompertz: Monotonic hazard
  - Log-normal, Log-logistic: Inverse U-shaped hazard

Weibull Regression Model

- The constant hazard assumption is often too restrictive
- The Weibull model relaxes the assumption by introducing a "shape" parameter

- Model: $Y_i \mid X_i \sim_{\text{ind}} \text{Weibull}(\mu_i, \alpha)$ where $\mu_i = \exp(X_i^\top \beta)$ and $\alpha > 0$
- Density: $f(y \mid \mu_i, \alpha) = \frac{\alpha}{\mu_i^\alpha} y^{\alpha-1} \exp\{-(y/\mu_i)^\alpha\}$
- Reduces to the exponential model when $\alpha = 1$
- Survival function: $S(y) = \exp\{-(y/\mu_i)^\alpha\}$
- Hazard function: $\lambda(y) = \frac{\alpha}{\mu_i^\alpha} y^{\alpha-1}$

- The monotonic hazard assumption: increasing (decreasing) if $\alpha > 1$ (if $\alpha < 1$)

- Other parametric regression models:
  - Gompertz: Monotonic hazard
  - Log-normal, Log-logistic: Inverse U-shaped hazard

Weibull Regression Model

- The constant hazard assumption is often too restrictive
- The Weibull model relaxes the assumption by introducing a "shape" parameter

- Model: $Y_i \mid X_i \sim_{\text{ind}} \text{Weibull}(\mu_i, \alpha)$ where $\mu_i = \exp(X_i^\top \beta)$ and $\alpha > 0$
- Density: $f(y \mid \mu_i, \alpha) = \frac{\alpha}{\mu_i^\alpha} y^{\alpha-1} \exp\{-(y/\mu_i)^\alpha\}$
- Reduces to the exponential model when $\alpha = 1$
- Survival function: $S(y) = \exp\{-(y/\mu_i)^\alpha\}$
- Hazard function: $\lambda(y) \;=\; \frac{\alpha}{\mu_i^\alpha} y^{\alpha-1}$

- The monotonic hazard assumption: increasing (decreasing) if $\alpha > 1$ (if $\alpha < 1$)

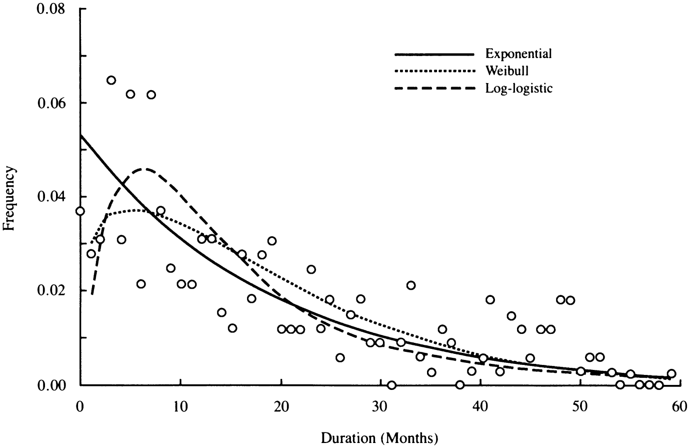
- Other parametric regression models:
  - Gompertz: Monotonic hazard
  - Log-normal, Log-logistic: Inverse U-shaped hazard

Cabinet Duration Example: Exponential or Weibull?

King et al. (Exponential) vs. Warwick and Easton (Weibull)

- Comparing density functions:

Figure 1. Duration Frequencies with Three Fitted Distributions

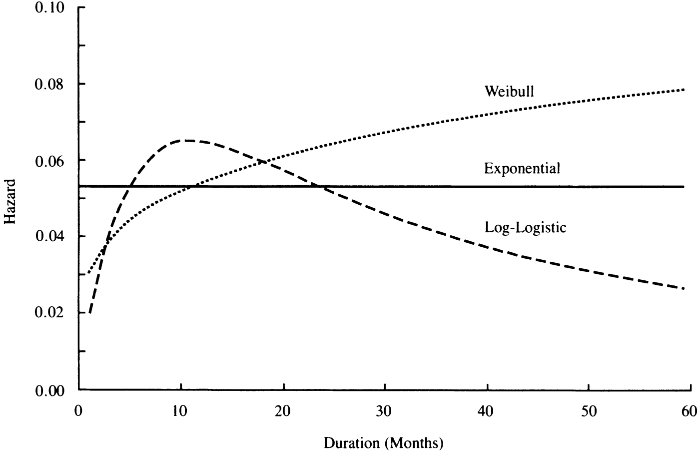


Cabinet Duration Example: Exponential or Weibull?

King et al. (Exponential) vs. Warwick and Easton (Weibull)

- Comparing hazard functions:

Figure 2. Estimated Hazard Functions for Three Distributions



Semi-Parametric Regression for Survival Data

- Less restriction on the hazard function
- Time-varying covariates to further model stochastic risks
- Note that both exponential and Weibull models are proportional hazard models:

$$\lambda(y \mid X_i) = \underbrace{\lambda_0(y)}_{\text{baseline hazard}} \exp(X_i^{'}\beta^*)$$

where $\lambda_0(y) = \begin{cases} 1 & \text{(exponential)} \\ \alpha y^{\alpha-1} & \text{(Weibull)} \end{cases}$ and $\beta^* = -\alpha\beta$

- The Cox Proportional Hazard Model generalizes this model:

$$\lambda(y \mid X_i(y)) = \lambda_0(y) \exp(X_i(y)^{'}\beta^*)$$

where

- $\lambda_0(y)$: Nonparametric baseline hazard common to all i across t
- $X_i(y)$: (Potentially) time-varying covariates

(Note: We omit $*$ from hereon)

Semi-Parametric Regression for Survival Data

- Less restriction on the hazard function
- Time-varying covariates to further model stochastic risks
- Note that both exponential and Weibull models are proportional hazard models:

$$\lambda(y \mid X_i) = \underbrace{\lambda_0(y)}_{\text{baseline hazard}} \exp(X_i' \beta^*)$$

where $\lambda_0(y) = \begin{cases} 1 & \text{(exponential)} \\ \alpha y^{\alpha - 1} & \text{(Weibull)} \end{cases}$ and $\beta^* = -\alpha\beta$

- The Cox Proportional Hazard Model generalizes this model:

$$\lambda(y \mid X_i(y)) = \lambda_0(y) \exp(X_i(y)' \beta^*)$$

where

  - $\lambda_0(y)$: Nonparametric baseline hazard common to all i across t
  - $X_i(y)$: (Potentially) time-varying covariates

(Note: We omit $*$ from hereon)

Semi-Parametric Regression for Survival Data

- Less restriction on the hazard function
- Time-varying covariates to further model stochastic risks
- Note that both exponential and Weibull models are proportional hazard models:

$$\lambda(y \mid X_i) \;=\; \underbrace{\lambda_0(y)}_{\text{baseline hazard}} \quad \exp(X_i^{'}\beta^*)$$

where $\lambda_0(y) \;=\; \begin{cases} 1 & \text{(exponential)} \\ \alpha y^{\alpha-1} & \text{(Weibull)} \end{cases}$ and $\beta^* = -\alpha\beta$

- The Cox Proportional Hazard Model generalizes this model:

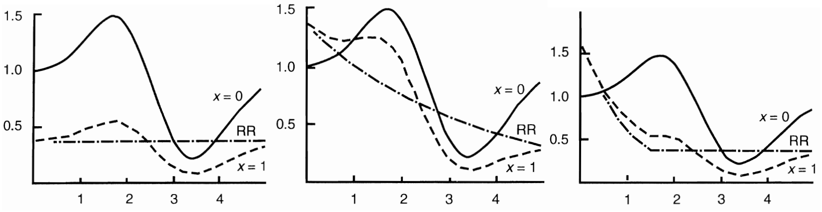
$$\lambda(y \mid X_i(y)) \;=\; \lambda_0(y) \exp(X_i(y)^{'}\beta^*)$$

where

  - $\lambda_0(y)$: Nonparametric baseline hazard common to all i across t
  - $X_i(y)$: (Potentially) time-varying covariates

(Note: We omit $*$ from hereon)

Example: Hazards Accommodated by the Cox Model



- The Cox PH model allows flexible shapes of hazard functions
- Suppose we have one binary predictor $x \in \{0, 1\}$ to model y:
  1. $\lambda(y \mid x) = \lambda_0(y) \exp(x\beta)$ —— no time-varying covariate
  2. $\lambda(y \mid x) = \lambda_0(y) \exp[x\beta_1 + xy\beta_2]$ —— interaction with time trend
  3. $\lambda(y \mid x) = \lambda_0(y) \exp[x\beta_1 + x(1.5 - y)\mathbf{1}\{y \leq 1.5\}\beta_2]$ —— allowing high initial risk

Note: In the figures, the relative risk (RR) stands for:

$$RR = \frac{\lambda(y \mid x = 1)}{\lambda(y \mid x = 0)} = \exp[g(y \mid x = 1) - g(y \mid x = 0)]$$

Estimation of Cox Proportional Hazard Models

- Joint MLE for $\lambda_0(y)$ and β is difficult (because $\lambda_0(y)$ is nonparametric)
- Instead, consider the partial likelihood function which only contains information about non-censored observations
- Because of the independent censoring assumption, this should give us a consistent (although not efficient) estimate for β
- For now, suppose that no two observations fail at the same time
- This implies we can unambiguously index observations by j
- Under this assumption, the partial likelihood function turns out to be:

$$L_P(\beta) = \prod_{j=1}^{J} \frac{\exp(X_i(t_j)^\top \beta)}{\sum_{k \in R(t_j)} \exp(X_k(t_j)^\top \beta)}$$

  where $R(t_j)$ = risk set at time t_j
- Note that $\lambda_0(y)$ drops out of the partial likelihood
- Take the log and maximize to obtain the estimate of β

Estimation of Cox Proportional Hazard Models

- Joint MLE for $\lambda_0(y)$ and β is difficult (because $\lambda_0(y)$ is nonparametric)
- Instead, consider the partial likelihood function which only contains information about non-censored observations
- Because of the independent censoring assumption, this should give us a consistent (although not efficient) estimate for β
- For now, suppose that no two observations fail at the same time
- This implies we can unambiguously index observations by j
- Under this assumption, the partial likelihood function turns out to be:

$$L_P(\beta) = \prod_{j=1}^{J} \frac{\exp(X_i(t_j)^\top \beta)}{\sum_{k \in R(t_j)} \exp(X_k(t_j)^\top \beta)}$$

where $R(t_j)$ = risk set at time t_j

- Note that $\lambda_0(y)$ drops out of the partial likelihood
- Take the log and maximize to obtain the estimate of β

Estimation of Cox Proportional Hazard Models

- Joint MLE for $\lambda_0(y)$ and β is difficult (because $\lambda_0(y)$ is nonparametric)
- Instead, consider the partial likelihood function which only contains information about non-censored observations
- Because of the independent censoring assumption, this should give us a consistent (although not efficient) estimate for β
- For now, suppose that no two observations fail at the same time
- This implies we can unambiguously index observations by j
- Under this assumption, the partial likelihood function turns out to be:

$$L_P(\beta) = \prod_{j=1}^{J} \frac{\exp(X_i(t_j)^\top \beta)}{\sum_{k \in R(t_j)} \exp(X_k(t_j)^\top \beta)}$$

  where $R(t_j)$ = risk set at time t_j
- Note that $\lambda_0(y)$ drops out of the partial likelihood
- Take the log and maximize to obtain the estimate of β

Estimation of Cox Proportional Hazard Models

- Joint MLE for $\lambda_0(y)$ and β is difficult (because $\lambda_0(y)$ is nonparametric)
- Instead, consider the partial likelihood function which only contains information about non-censored observations
- Because of the independent censoring assumption, this should give us a consistent (although not efficient) estimate for β
- For now, suppose that no two observations fail at the same time
- This implies we can unambiguously index observations by j
- Under this assumption, the partial likelihood function turns out to be:

$$L_P(\beta) = \prod_{j=1}^{J} \frac{\exp(X_i(t_j)^\top \beta)}{\sum_{k \in R(t_j)} \exp(X_k(t_j)^\top \beta)}$$

where $R(t_j)$ = risk set at time t_j

- Note that $\lambda_0(y)$ drops out of the partial likelihood
- Take the log and maximize to obtain the estimate of β

Estimation of Cox Proportional Hazard Models

- Joint MLE for $\lambda_0(y)$ and β is difficult (because $\lambda_0(y)$ is nonparametric)
- Instead, consider the partial likelihood function which only contains information about non-censored observations
- Because of the independent censoring assumption, this should give us a consistent (although not efficient) estimate for β
- For now, suppose that no two observations fail at the same time
- This implies we can unambiguously index observations by j
- Under this assumption, the partial likelihood function turns out to be:

$$L_P(\beta) = \prod_{j=1}^{J} \frac{\exp(X_i(t_j)^\top \beta)}{\sum_{k \in R(t_j)} \exp(X_k(t_j)^\top \beta)}$$

where $R(t_j)$ = risk set at time t_j

- Note that $\lambda_0(y)$ drops out of the partial likelihood
- Take the log and maximize to obtain the estimate of β

Estimation of Cox Proportional Hazard Models

- Joint MLE for $\lambda_0(y)$ and β is difficult (because $\lambda_0(y)$ is nonparametric)
- Instead, consider the partial likelihood function which only contains information about non-censored observations
- Because of the independent censoring assumption, this should give us a consistent (although not efficient) estimate for β
- For now, suppose that no two observations fail at the same time
- This implies we can unambiguously index observations by j
- Under this assumption, the partial likelihood function turns out to be:

$$L_P(\beta) = \prod_{j=1}^{J} \frac{\exp(X_i(t_j)^\top \beta)}{\sum_{k \in R(t_j)} \exp(X_k(t_j)^\top \beta)}$$

where $R(t_j)$ = risk set at time t_j

- Note that $\lambda_0(y)$ drops out of the partial likelihood
- Take the log and maximize to obtain the estimate of β

Hypothesis tests, Model checking, and Likelihood models!