Political Methodology III: Model Based Inference

Justin Grimmer

Associate Professor
Department of Political Science
Stanford University

May 2nd, 2019

Model Based Inference

1) Likelihood inference
2) Logit/Probit
3) Ordered Probit
4) Choice Models:
5) Count Models
6) Survival Models
7) Hypothesis Tests + Model Checking in Likelihood
    - Likelihood Ratios, Wald, and Score tests
    - Model Checking: analysis of residuals, hat values, etc.

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

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

- 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

$$S(y) \equiv \Pr(Y_i > y) = \int_y^\infty f(t)dt = 1 - \int_0^y f(t)dt = 1 - F(y)$$

- 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

$$S(y) \equiv \Pr(Y_i > y) = \int_y^\infty f(t)dt = 1 - \int_0^y f(t)dt = 1 - F(y)$$

- 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

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

- 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:

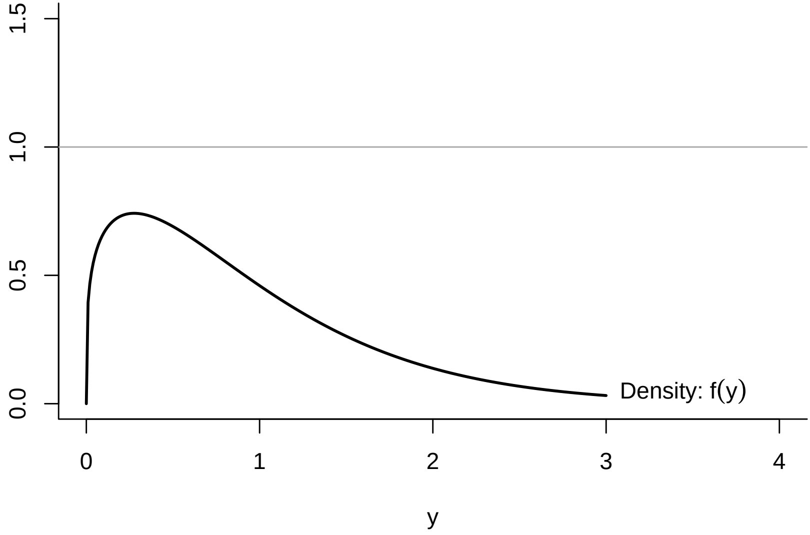
$$\begin{aligned}
\mathsf{E}(Y_i) &= \int_0^{\infty} y f(y)dy \\
&= y\left(F(y)|_0^{\infty}\right) - \int_0^{\infty} F(y)dy \\
&= \int_0^{\infty} (1 - F(y))dy \\
&= \int_0^{\infty} S(y)dy
\end{aligned}$$

Survival Function

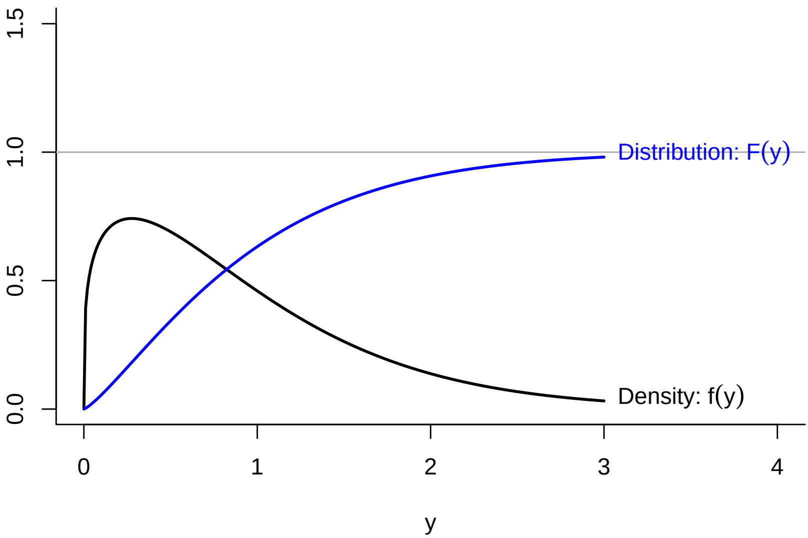
- One-to-one relationships with density and probability:

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

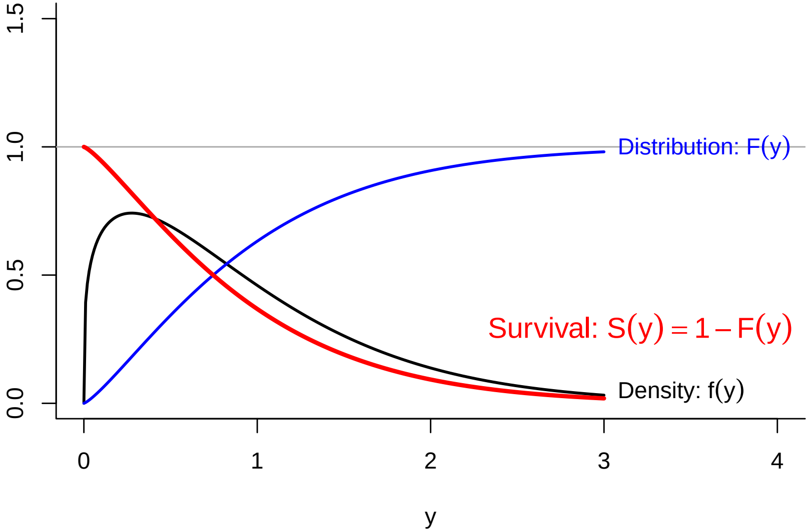
Survival Function



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

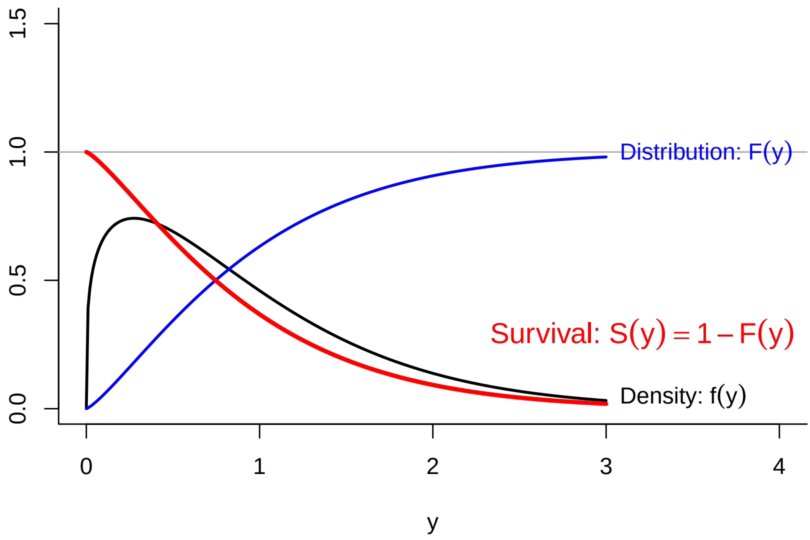
$$\lambda(y) \equiv \lim_{h \downarrow 0} \frac{\Pr(y \leq Y_i < y + h \mid Y_i \geq y)}{h} = \frac{f(y)}{S(y)}$$

- "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.

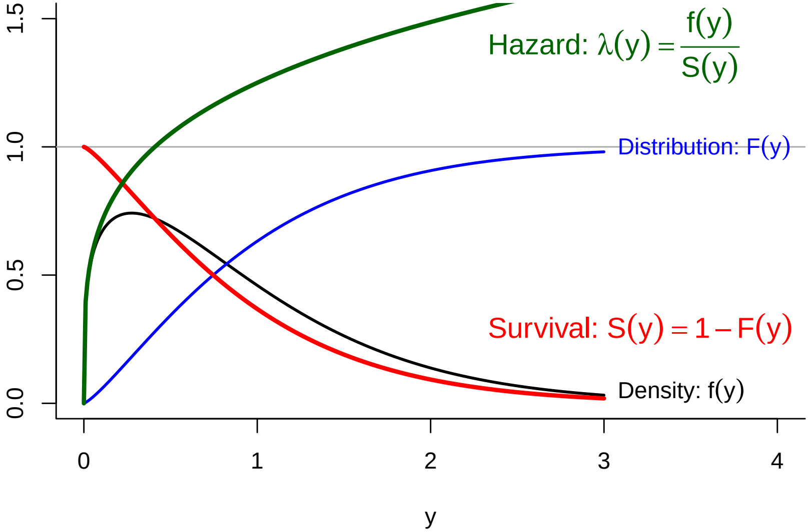
- One-to-one relationship with survival function:

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

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

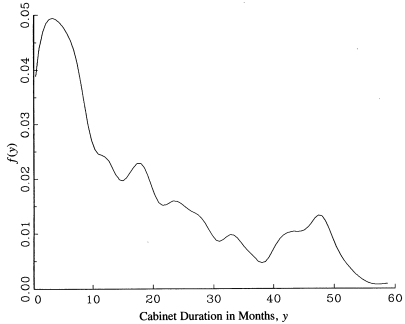
Censoring

- 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, $n = 314$

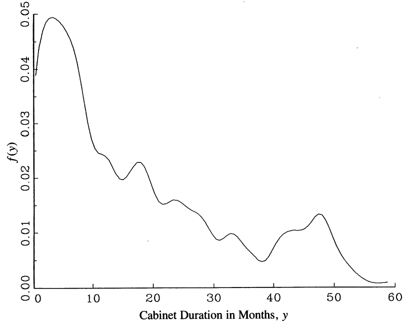


- Notice the "bump"?
- Some cabinets end their lives "naturally"
- Others end because of constitutional interelection periods (CIEP)
- King et al. treat CIEPs as censored observations
- But is this "censoring" independent?

Cabinet Duration Example: Censoring

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

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

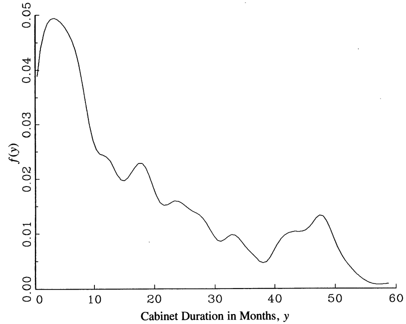


- Notice the "bump"?
- Some cabinets end their lives "naturally"
- Others end because of constitutional interelection periods (CIEP)
- King et al. treat CIEPs as censored observations
- But is this "censoring" independent?

Cabinet Duration Example: Censoring

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

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

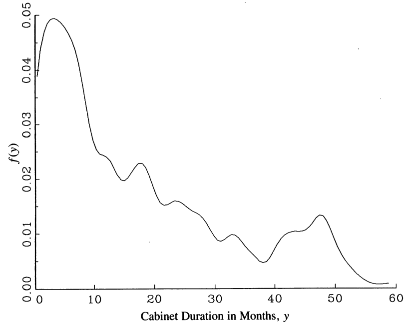


- Notice the "bump"?
- Some cabinets end their lives "naturally"
- Others end because of constitutional interelection periods (CIEP)
- King et al. treat CIEPs as censored observations
- But is this "censoring" independent?

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:

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

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:

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

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:

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

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:

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

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:

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

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:

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

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:

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

Kaplan-Meier Estimator

- This leads to the Kaplan-Meier estimator:

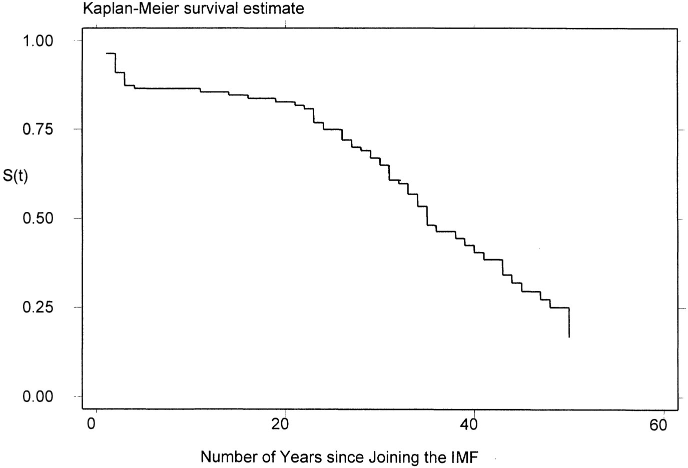
$$\widehat{S}(t_j) = \prod_{k=j}^{J}(1 - \hat{\lambda}(t_k)) = \prod_{k=j}^{J} \frac{r_k - d_k}{r_k}$$

- 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: $Y_i \mid X_i \sim_{\text{ind}}$ Exponential(μ_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: $Y_i \mid X_i \sim_{\text{ind}}$ Exponential(μ_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: $Y_i \mid X_i \sim_{\text{ind}}$ Exponential(μ_i) where $\mu_i = \exp(X_i'\beta)$
- Density: $f(y \mid \mu_i) = \frac{1}{\mu_i}\exp(-y/\mu_i)$
- Mean $\mathsf{E}(Y_i \mid \mu_i) = \mu_i$ and Variance $\mathsf{Var}(Y_i \mid \mu_i) = \mu_i^2$
- 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

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:

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

- 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:

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

- 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:

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

- 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: $Y_i \mid X_i \sim_{\text{ind}} \text{Weibull}(\mu_i, \alpha)$ where $\mu_i = \exp(X_i'\beta)$ 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 $\mu_i = \exp(X_i'\beta)$ 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 $\mu_i = \exp(X_i'\beta)$ 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 $\alpha < 1$)

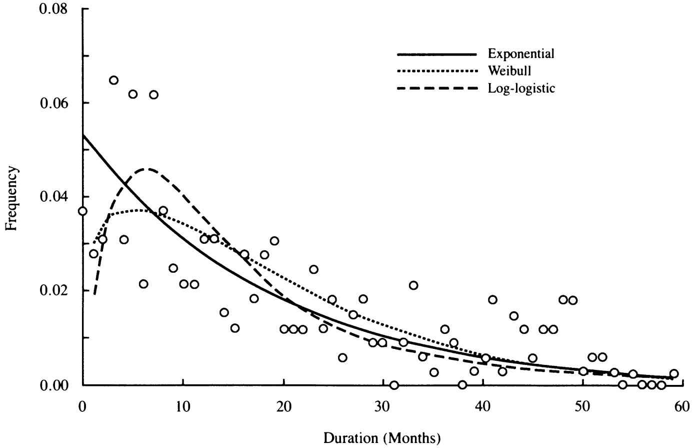
- Other parametric regression models:
  - Gompertz: Monotonic hazard
  - Log-normal, Log-logistic: Inverse U-shaped hazard

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

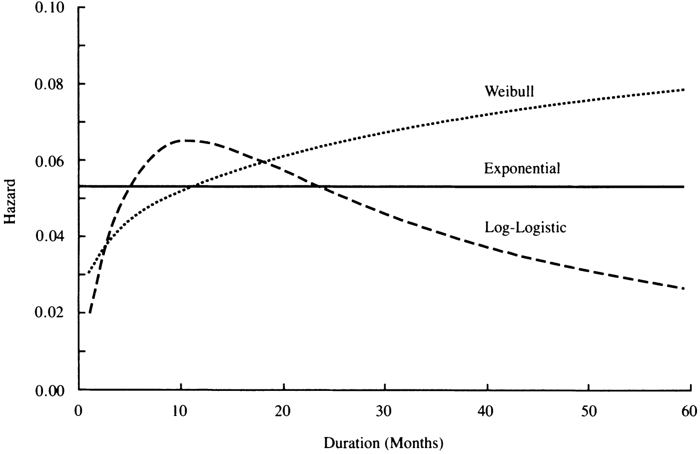


Cabinet Duration Example: Exponential or Weibull?

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:

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

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)

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)

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:

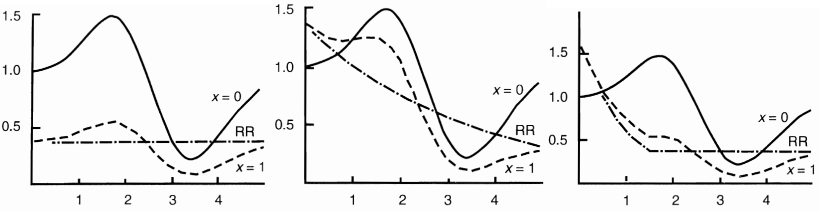
$$\lambda(y \mid X_i(y)) = \lambda_0(y) \exp(X_i(y)^{'}\beta^*)$$

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)

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 Testing

Simple Example: Antiobama Speech

We'll use the speech data from the problem set, as follows:

- $Y_i = 1$ if representative says `obamacare` or `big government` during the year, 0 otherwise
- $\boldsymbol{X}_i = (1, I(\text{Year} = 2010)_i, \text{Democrat}_i, \text{DW-Nom}_i)$

$$
\begin{aligned}
Y_i &\sim \text{Bernoulli}(\pi_i) \\
\pi_i &= \text{logit}^{-1}(\boldsymbol{X}_i' \boldsymbol{\beta}) = \frac{1}{1 + \exp(-\boldsymbol{X}_i' \boldsymbol{\beta})}
\end{aligned}
$$

Which covariates do we include? $\leadsto$ depends on goal.

- Predictive goal $\leadsto$ replicate task
- Model fitting $\leadsto$ do covariates increase likelihood? Can we drop them?

Hypothesis Testing — Likelihood Ratio Test

- Null (H_0): $h_1(\boldsymbol{\beta}) = \cdots = h_Q(\boldsymbol{\beta}) = 0$ (Q equality constraints)
- Alternative (H_1): No such constraints
- Let $\widehat{\boldsymbol{\beta}}_R = \widehat{\boldsymbol{\beta}}_{MLE|H_0}$ (restricted MLE) and $\widehat{\boldsymbol{\beta}}_{UR} = \widehat{\boldsymbol{\beta}}_{MLE}$ (original MLE)
- Likelihood ratio (LR) test: If H_0 is true, $L(\widehat{\boldsymbol{\beta}}_R)$ should be equal to $L(\widehat{\boldsymbol{\beta}}_{UR})$ except for sampling variability
- LR statistic:

$$LR(Y) \equiv -2 \log \frac{L(\widehat{\boldsymbol{\beta}}_R)}{L(\widehat{\boldsymbol{\beta}}_{UR})} = 2 \left[\ell(\widehat{\boldsymbol{\beta}}_{UR}) - \ell(\widehat{\boldsymbol{\beta}}_R) \right]$$

- We can show that $LR(Y) \xrightarrow{d} \chi^2_Q$
- Works for testing any nested models
  — model under H_0 has to be a special case of model under H_1

Hypothesis Testing — Likelihood Ratio Test

- Null (H_0): $h_1(\boldsymbol{\beta}) = \cdots = h_Q(\boldsymbol{\beta}) = 0$ (Q equality constraints)
- Alternative (H_1): No such constraints
- Let $\widehat{\boldsymbol{\beta}}_R = \widehat{\boldsymbol{\beta}}_{MLE|H_0}$ (restricted MLE) and $\widehat{\boldsymbol{\beta}}_{UR} = \widehat{\boldsymbol{\beta}}_{MLE}$ (original MLE)
- Likelihood ratio (LR) test: If H_0 is true, $L(\widehat{\boldsymbol{\beta}}_R)$ should be equal to $L(\widehat{\boldsymbol{\beta}}_{UR})$ except for sampling variability
- LR statistic:

$$LR(Y) \equiv -2 \log \frac{L(\widehat{\boldsymbol{\beta}}_R)}{L(\widehat{\boldsymbol{\beta}}_{UR})} = 2 \left[\ell(\widehat{\boldsymbol{\beta}}_{UR}) - \ell(\widehat{\boldsymbol{\beta}}_R) \right]$$

- We can show that $LR(Y) \xrightarrow{d} \chi^2_Q$
- Works for testing any nested models
  — model under H_0 has to be a special case of model under H_1

Hypothesis Testing — Likelihood Ratio Test

- Null (H_0): $h_1(\boldsymbol{\beta}) = \cdots = h_Q(\boldsymbol{\beta}) = 0$ (Q equality constraints)
- Alternative (H_1): No such constraints
- Let $\widehat{\boldsymbol{\beta}}_R = \widehat{\boldsymbol{\beta}}_{MLE|H_0}$ (restricted MLE) and $\widehat{\boldsymbol{\beta}}_{UR} = \widehat{\boldsymbol{\beta}}_{MLE}$ (original MLE)
- Likelihood ratio (LR) test: If H_0 is true, $L(\widehat{\boldsymbol{\beta}}_R)$ should be equal to $L(\widehat{\boldsymbol{\beta}}_{UR})$ except for sampling variability
- LR statistic:

$$LR(Y) \equiv -2\log\frac{L(\widehat{\boldsymbol{\beta}}_R)}{L(\widehat{\boldsymbol{\beta}}_{UR})} = 2\left[\ell(\widehat{\boldsymbol{\beta}}_{UR}) - \ell(\widehat{\boldsymbol{\beta}}_R)\right]$$

- We can show that $LR(Y) \xrightarrow{d} \chi^2_Q$
- Works for testing any nested models — model under H_0 has to be a special case of model under H_1

Hypothesis Testing — Likelihood Ratio Test

- Null (H_0): $h_1(\boldsymbol{\beta}) = \cdots = h_Q(\boldsymbol{\beta}) = 0$ (Q equality constraints)
- Alternative (H_1): No such constraints
- Let $\widehat{\boldsymbol{\beta}}_R = \widehat{\boldsymbol{\beta}}_{MLE|H_0}$ (restricted MLE) and $\widehat{\boldsymbol{\beta}}_{UR} = \widehat{\boldsymbol{\beta}}_{MLE}$ (original MLE)
- Likelihood ratio (LR) test: If H_0 is true, $L(\widehat{\boldsymbol{\beta}}_R)$ should be equal to $L(\widehat{\boldsymbol{\beta}}_{UR})$ except for sampling variability
- LR statistic:

$$LR(Y) \equiv -2\log\frac{L(\widehat{\boldsymbol{\beta}}_R)}{L(\widehat{\boldsymbol{\beta}}_{UR})} = 2\left[\ell(\widehat{\boldsymbol{\beta}}_{UR}) - \ell(\widehat{\boldsymbol{\beta}}_R)\right]$$

- We can show that $LR(Y) \xrightarrow{d} \chi^2_Q$
- Works for testing any nested models
  — model under H_0 has to be a special case of model under H_1

Hypothesis Testing — Likelihood Ratio Test

- Null (H_0): $h_1(\boldsymbol{\beta}) = \cdots = h_Q(\boldsymbol{\beta}) = 0$ (Q equality constraints)
- Alternative (H_1): No such constraints
- Let $\widehat{\boldsymbol{\beta}}_R = \widehat{\boldsymbol{\beta}}_{MLE|H_0}$ (restricted MLE) and $\widehat{\boldsymbol{\beta}}_{UR} = \widehat{\boldsymbol{\beta}}_{MLE}$ (original MLE)
- Likelihood ratio (LR) test: If H_0 is true, $L(\widehat{\boldsymbol{\beta}}_R)$ should be equal to $L(\widehat{\boldsymbol{\beta}}_{UR})$ except for sampling variability
- LR statistic:

$$LR(Y) \equiv -2 \log \frac{L(\widehat{\boldsymbol{\beta}}_R)}{L(\widehat{\boldsymbol{\beta}}_{UR})} = 2 \left[\ell(\widehat{\boldsymbol{\beta}}_{UR}) - \ell(\widehat{\boldsymbol{\beta}}_R) \right]$$

- We can show that $LR(Y) \xrightarrow{d} \chi^2_Q$
- Works for testing any nested models — model under H_0 has to be a special case of model under H_1

Hypothesis Testing — Likelihood Ratio Test

- Null (H_0): $h_1(\boldsymbol{\beta}) = \cdots = h_Q(\boldsymbol{\beta}) = 0$ (Q equality constraints)
- Alternative (H_1): No such constraints
- Let $\widehat{\boldsymbol{\beta}}_R = \widehat{\boldsymbol{\beta}}_{MLE|H_0}$ (restricted MLE) and $\widehat{\boldsymbol{\beta}}_{UR} = \widehat{\boldsymbol{\beta}}_{MLE}$ (original MLE)
- Likelihood ratio (LR) test: If H_0 is true, $L(\widehat{\boldsymbol{\beta}}_R)$ should be equal to $L(\widehat{\boldsymbol{\beta}}_{UR})$ except for sampling variability
- LR statistic:

$$LR(Y) \equiv -2\log\frac{L(\widehat{\boldsymbol{\beta}}_R)}{L(\widehat{\boldsymbol{\beta}}_{UR})} = 2\left[\ell(\widehat{\boldsymbol{\beta}}_{UR}) - \ell(\widehat{\boldsymbol{\beta}}_R)\right]$$

- We can show that $LR(Y) \xrightarrow{d} \chi^2_Q$
- Works for testing any nested models
  — model under H_0 has to be a special case of model under H_1

Hypothesis Testing — Likelihood Ratio Test

- Null (H_0): $h_1(\boldsymbol{\beta}) = \cdots = h_Q(\boldsymbol{\beta}) = 0$ (Q equality constraints)
- Alternative (H_1): No such constraints
- Let $\widehat{\boldsymbol{\beta}}_R = \widehat{\boldsymbol{\beta}}_{MLE|H_0}$ (restricted MLE) and $\widehat{\boldsymbol{\beta}}_{UR} = \widehat{\boldsymbol{\beta}}_{MLE}$ (original MLE)
- Likelihood ratio (LR) test: If H_0 is true, $L(\widehat{\boldsymbol{\beta}}_R)$ should be equal to $L(\widehat{\boldsymbol{\beta}}_{UR})$ except for sampling variability
- LR statistic:

$$LR(Y) \equiv -2\log\frac{L(\widehat{\boldsymbol{\beta}}_R)}{L(\widehat{\boldsymbol{\beta}}_{UR})} = 2\left[\ell(\widehat{\boldsymbol{\beta}}_{UR}) - \ell(\widehat{\boldsymbol{\beta}}_R)\right]$$

- We can show that $LR(Y) \xrightarrow{d} \chi^2_Q$
- Works for testing any nested models
  — model under H_0 has to be a special case of model under H_1

```
un_rest_reg<- glm(once~two_10 + dem + dw_nom,
   data = speech_dat, family = binomial(link = logit))

rest_reg<- glm(once~1, family= binomial(link = logit))

##calculating the likelihood ratio
log_lik<- function(pars, X, Y){
   y.tilde<- X%*%pars
   probs<- plogis(y.tilde)
   log_out<- Y%*%log(probs) + (1-Y)%*%log(1 - probs)
   return(log_out)
}
X<- cbind(1, two_10, dem, speech_dat$dw_nom)

un_rest<- log_lik(un_rest_reg$coef, X, once)

rest<- log_lik(rest_reg$coef, as.matrix(rep(1, nrow(X))), once)

> 2 * un_rest - 2*rest
   [,1]
[1,] 433.996
```

```
> 2 * un_rest - 2*rest
   [,1]
[1,] 433.996
##get the same statistic automatically from glm
diff<- un_rest_reg$null.deviance - un_rest_reg$deviance
1 - pchisq(diff, 3) ##very small!
[1] 0
```

Hypothesis Testing — Wald Test

- Wald test: If true, the null $h_1(\boldsymbol{\beta}) = \cdots = h_Q(\boldsymbol{\beta}) = 0$ should approximately hold even if we substitute $\widehat{\boldsymbol{\beta}}_{UR}$ for $\boldsymbol{\beta}$
- Wald statistic: Use asymptotic distribution of $\widehat{\beta}$ and representation of restrictions, properties of normal distribution to obtain form

$$W \equiv h(\widehat{\boldsymbol{\beta}}_{UR})' \left[\left(\frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{UR}} \right)' \widehat{\text{Var}(\widehat{\boldsymbol{\beta}}_{UR})} \left(\frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{UR}} \right) \right]^{-1} h(\widehat{\boldsymbol{\beta}}_{UR})$$

- The "meat" $\simeq \widehat{\text{Var}(h(\widehat{\boldsymbol{\beta}}_{UR}))}$ (Delta method)
- Choose any $\widehat{\text{Var}(\widehat{\boldsymbol{\beta}}_{UR})}$ as appropriate (e.g. Huber-White)
- We can show that $W \xrightarrow{d} \chi^2_Q$
- An important special case: $Q = 1$ and $H_0 : \beta = 0$
- In this case, we can use the z statistic:

$$z = W^{1/2} = \frac{\widehat{\boldsymbol{\beta}}_{UR}}{\text{s.e.}(\widehat{\boldsymbol{\beta}}_{UR})} \xrightarrow{d} \text{N}(0,1)$$

Hypothesis Testing — Wald Test

- Wald test: If true, the null $h_1(\boldsymbol{\beta}) = \cdots = h_Q(\boldsymbol{\beta}) = 0$ should approximately hold even if we substitute $\widehat{\boldsymbol{\beta}}_{UR}$ for $\boldsymbol{\beta}$
- Wald statistic: Use asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ and representation of restrictions, properties of normal distribution to obtain form

$$W \equiv h(\widehat{\boldsymbol{\beta}}_{UR})' \left[\left(\frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{UR}} \right)' \widehat{\mathsf{Var}(\widehat{\boldsymbol{\beta}}_{UR})} \left(\frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{UR}} \right) \right]^{-1} h(\widehat{\boldsymbol{\beta}}_{UR})$$

- The "meat" $\simeq \widehat{\mathsf{Var}(h(\widehat{\boldsymbol{\beta}}_{UR}))}$ (Delta method)
- Choose any $\widehat{\mathsf{Var}(\widehat{\boldsymbol{\beta}}_{UR})}$ as appropriate (e.g. Huber-White)
- We can show that $W \xrightarrow{d} \chi^2_Q$
- An important special case: $Q = 1$ and $H_0 : \beta = 0$
- In this case, we can use the z statistic:

$$z = W^{1/2} = \frac{\widehat{\beta}_{UR}}{\text{s.e.}(\widehat{\beta}_{UR})} \xrightarrow{d} \mathsf{N}(0,1)$$

Hypothesis Testing — Wald Test

- Wald test: If true, the null $h_1(\boldsymbol{\beta}) = \cdots = h_Q(\boldsymbol{\beta}) = 0$ should approximately hold even if we substitute $\widehat{\boldsymbol{\beta}}_{UR}$ for $\boldsymbol{\beta}$
- Wald statistic: Use asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ and representation of restrictions, properties of normal distribution to obtain form

$$W \equiv h(\widehat{\boldsymbol{\beta}}_{UR})' \left[\left(\frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{UR}} \right)' \widehat{\text{Var}(\widehat{\boldsymbol{\beta}}_{UR})} \left(\frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{UR}} \right) \right]^{-1} h(\widehat{\boldsymbol{\beta}}_{UR})$$

- The "meat" $\simeq \widehat{\text{Var}(h(\widehat{\boldsymbol{\beta}}_{UR}))}$ (Delta method)
- Choose any $\widehat{\text{Var}(\widehat{\boldsymbol{\beta}}_{UR})}$ as appropriate (e.g. Huber-White)
- We can show that $W \xrightarrow{d} \chi^2_Q$

- An important special case: $Q = 1$ and $H_0 : \beta = 0$
- In this case, we can use the z statistic:

$$z = W^{1/2} = \frac{\widehat{\beta}_{UR}}{\text{s.e.}(\widehat{\beta}_{UR})} \xrightarrow{d} \text{N}(0,1)$$

Hypothesis Testing — Wald Test

- Wald test: If true, the null $h_1(\boldsymbol{\beta}) = \cdots = h_Q(\boldsymbol{\beta}) = 0$ should approximately hold even if we substitute $\widehat{\boldsymbol{\beta}}_{UR}$ for $\boldsymbol{\beta}$
- Wald statistic: Use asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ and representation of restrictions, properties of normal distribution to obtain form

$$W \equiv h(\widehat{\boldsymbol{\beta}}_{UR})' \left[\left(\frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{UR}} \right)' \widehat{\mathrm{Var}(\widehat{\boldsymbol{\beta}}_{UR})} \left(\frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{UR}} \right) \right]^{-1} h(\widehat{\boldsymbol{\beta}}_{UR})$$

- The "meat" $\simeq \widehat{\mathrm{Var}(h(\widehat{\boldsymbol{\beta}}_{UR}))}$ (Delta method)
- Choose any $\widehat{\mathrm{Var}(\widehat{\boldsymbol{\beta}}_{UR})}$ as appropriate (e.g. Huber-White)
- We can show that $W \xrightarrow{d} \chi^2_Q$
- An important special case: $Q = 1$ and $H_0 : \beta = 0$
- In this case, we can use the z statistic:

$$z = W^{1/2} = \frac{\widehat{\beta}_{UR}}{\mathrm{s.e.}(\widehat{\beta}_{UR})} \xrightarrow{d} \mathrm{N}(0,1)$$

Hypothesis Testing — Wald Test

- Wald test: If true, the null $h_1(\boldsymbol{\beta}) = \cdots = h_Q(\boldsymbol{\beta}) = 0$ should approximately hold even if we substitute $\widehat{\boldsymbol{\beta}}_{UR}$ for $\boldsymbol{\beta}$
- Wald statistic: Use asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ and representation of restrictions, properties of normal distribution to obtain form

$$W \equiv h(\widehat{\boldsymbol{\beta}}_{UR})' \left[\left(\frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{UR}} \right)' \widehat{\text{Var}(\widehat{\boldsymbol{\beta}}_{UR})} \left(\frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{UR}} \right) \right]^{-1} h(\widehat{\boldsymbol{\beta}}_{UR})$$

- The "meat" $\simeq \widehat{\text{Var}(h(\widehat{\boldsymbol{\beta}}_{UR}))}$ (Delta method)
- Choose any $\widehat{\text{Var}(\widehat{\boldsymbol{\beta}}_{UR})}$ as appropriate (e.g. Huber-White)
- We can show that $W \xrightarrow{d} \chi^2_Q$
- An important special case: $Q = 1$ and $H_0 : \beta = 0$
- In this case, we can use the z statistic:

$$z = W^{1/2} = \frac{\widehat{\beta}_{UR}}{\text{s.e.}(\widehat{\beta}_{UR})} \xrightarrow{d} \text{N}(0,1)$$

Hypothesis Testing — Wald Test

- Wald test: If true, the null $h_1(\boldsymbol{\beta}) = \cdots = h_Q(\boldsymbol{\beta}) = 0$ should approximately hold even if we substitute $\widehat{\boldsymbol{\beta}}_{UR}$ for $\boldsymbol{\beta}$
- Wald statistic: Use asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ and representation of restrictions, properties of normal distribution to obtain form

$$W \equiv h(\widehat{\boldsymbol{\beta}}_{UR})' \left[\left(\frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{UR}} \right)' \widehat{\mathsf{Var}(\widehat{\boldsymbol{\beta}}_{UR})} \left(\frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{UR}} \right) \right]^{-1} h(\widehat{\boldsymbol{\beta}}_{UR})$$

- The "meat" $\simeq \widehat{\mathsf{Var}(h(\widehat{\boldsymbol{\beta}}_{UR}))}$ (Delta method)
- Choose any $\widehat{\mathsf{Var}(\widehat{\boldsymbol{\beta}}_{UR})}$ as appropriate (e.g. Huber-White)
- We can show that $W \xrightarrow{d} \chi^2_Q$
- An important special case: $Q = 1$ and $H_0 : \beta = 0$
- In this case, we can use the z statistic:

$$z = W^{1/2} = \frac{\widehat{\beta}_{UR}}{\mathsf{s.e.}(\widehat{\beta}_{UR})} \xrightarrow{d} \mathsf{N}(0,1)$$

Hypothesis Testing — Wald Test

- Wald test: If true, the null $h_1(\boldsymbol{\beta}) = \cdots = h_Q(\boldsymbol{\beta}) = 0$ should approximately hold even if we substitute $\widehat{\boldsymbol{\beta}}_{UR}$ for $\boldsymbol{\beta}$
- Wald statistic: Use asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ and representation of restrictions, properties of normal distribution to obtain form

$$W \equiv h(\widehat{\boldsymbol{\beta}}_{UR})' \left[\left(\frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{UR}} \right)' \widehat{\mathsf{Var}(\widehat{\boldsymbol{\beta}}_{UR})} \left(\frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{UR}} \right) \right]^{-1} h(\widehat{\boldsymbol{\beta}}_{UR})$$

- The "meat" $\simeq \widehat{\mathsf{Var}(h(\widehat{\boldsymbol{\beta}}_{UR}))}$ (Delta method)
- Choose any $\widehat{\mathsf{Var}(\widehat{\boldsymbol{\beta}}_{UR})}$ as appropriate (e.g. Huber-White)
- We can show that $W \xrightarrow{d} \chi^2_Q$
- An important special case: $Q = 1$ and $H_0 : \beta = 0$
- In this case, we can use the z statistic:

$$z = W^{1/2} = \frac{\widehat{\boldsymbol{\beta}}_{UR}}{\mathsf{s.e.}(\widehat{\boldsymbol{\beta}}_{UR})} \xrightarrow{d} \mathsf{N}(0,1)$$

```
> un_rest_reg$coef%*%solve(vcov(un_rest_reg))%*%un_rest_reg$coef
[,1]
[1,] 225.2437

> 1 - pchisq(225.2437, 3)
[1] 0
```

Hypothesis Testing — Score Test

- At the unrestricted MLE $\hat{\theta}_{UR}$, $\sum_{i=1}^{N} s_i(\widehat{\boldsymbol{\beta}}_{UR}) = s(\widehat{\boldsymbol{\beta}}) = 0$ by construction
- Score test: If the null is true, $s(\widehat{\boldsymbol{\beta}}_R)$ should also equal zero except for sampling variability
- Score statistic: Use asymptotic distribution and properties of normal distribution to "standardize" $s(\widehat{\boldsymbol{\beta}}_R)$

$$LM = s(\widehat{\boldsymbol{\beta}}_R)' \widehat{\text{Var}(\widehat{\boldsymbol{\beta}}_R)} s(\widehat{\boldsymbol{\beta}}_R) \xrightarrow{d} \chi^2_Q$$

- For $\widehat{\boldsymbol{\beta}}_{QMLE}$, the expression is more complicated
- Also known as the Lagrange multiplier (LM) test due to an alternative derivation

Hypothesis Testing — Score Test

- At the unrestricted MLE $\hat{\theta}_{UR}$, $\sum_{i=1}^{N} s_i(\widehat{\boldsymbol{\beta}}_{UR}) = s(\widehat{\boldsymbol{\beta}}) = 0$ by construction
- Score test: If the null is true, $s(\widehat{\boldsymbol{\beta}}_R)$ should also equal zero except for sampling variability
- Score statistic: Use asymptotic distribution and properties of normal distribution to "standardize" $s(\widehat{\boldsymbol{\beta}}_R)$

$$LM \;=\; s(\widehat{\boldsymbol{\beta}}_R)^{'}\widehat{\text{Var}(\widehat{\boldsymbol{\beta}}_R)}s(\widehat{\boldsymbol{\beta}}_R) \;\xrightarrow{d}\; \chi^2_Q$$

- For $\widehat{\beta}_{QMLE}$, the expression is more complicated
- Also known as the Lagrange multiplier (LM) test due to an alternative derivation

Hypothesis Testing — Score Test

- At the unrestricted MLE $\hat{\theta}_{UR}$, $\sum_{i=1}^{N} s_i(\widehat{\boldsymbol{\beta}}_{UR}) = s(\widehat{\boldsymbol{\beta}}) = 0$ by construction
- Score test: If the null is true, $s(\widehat{\boldsymbol{\beta}}_R)$ should also equal zero except for sampling variability
- Score statistic: Use asymptotic distribution and properties of normal distribution to "standardize" $s(\widehat{\boldsymbol{\beta}}_R)$

$$LM \;=\; s(\widehat{\boldsymbol{\beta}}_R)' \widehat{\text{Var}(\widehat{\boldsymbol{\beta}}_R)} s(\widehat{\boldsymbol{\beta}}_R) \;\xrightarrow{d}\; \chi^2_Q$$

- For $\widehat{\beta}_{QMLE}$, the expression is more complicated
- Also known as the Lagrange multiplier (LM) test due to an alternative derivation

Hypothesis Testing — Score Test

- At the unrestricted MLE $\hat{\theta}_{UR}$, $\sum_{i=1}^{N} s_i(\widehat{\boldsymbol{\beta}}_{UR}) = s(\widehat{\boldsymbol{\beta}}) = 0$ by construction
- Score test: If the null is true, $s(\widehat{\boldsymbol{\beta}}_R)$ should also equal zero except for sampling variability
- Score statistic: Use asymptotic distribution and properties of normal distribution to "standardize" $s(\widehat{\boldsymbol{\beta}}_R)$

$$LM \;=\; s(\widehat{\boldsymbol{\beta}}_R)^{\prime}\widehat{\text{Var}(\widehat{\boldsymbol{\beta}}_R)}s(\widehat{\boldsymbol{\beta}}_R) \;\xrightarrow{d}\; \chi^2_Q$$

- For $\widehat{\boldsymbol{\beta}}_{QMLE}$, the expression is more complicated
- Also known as the Lagrange multiplier (LM) test due to an alternative derivation

Hypothesis Testing — Score Test

- At the unrestricted MLE $\hat{\theta}_{UR}$, $\sum_{i=1}^{N} s_i(\widehat{\boldsymbol{\beta}}_{UR}) = s(\widehat{\boldsymbol{\beta}}) = 0$ by construction
- Score test: If the null is true, $s(\widehat{\boldsymbol{\beta}}_R)$ should also equal zero except for sampling variability
- Score statistic: Use asymptotic distribution and properties of normal distribution to "standardize" $s(\widehat{\boldsymbol{\beta}}_R)$

$$LM \;=\; s(\widehat{\boldsymbol{\beta}}_R)^{\prime}\,\widehat{\mathsf{Var}(\widehat{\boldsymbol{\beta}}_R)}\,s(\widehat{\boldsymbol{\beta}}_R) \;\xrightarrow{d}\; \chi^2_Q$$

- For $\widehat{\boldsymbol{\beta}}_{QMLE}$, the expression is more complicated
- Also known as the Lagrange multiplier (LM) test due to an alternative derivation

```
score_func<- function(coef, X, Y){
   y.tilde<- X%*%coef
   probs<- plogis(y.tilde)
   out<- t(Y - probs)%*%X
   return(out) }
```

```
> round(score_func(un_rest_reg$coef, X, once), 2)

[1,] 0 0 0 0

rest_score<- score_func(c(rest_reg$coef, 0, 0, 0), X, once)

> round(rest_score,2)
[1,] 0 -6.30 -128.92 129.51
```

```
hess_func<- function(coef, X, Y){
   y.tilde<- X%*%coef
   probs<- plogis(y.tilde)
   base<- matrix(0, nrow = len(coef), ncol = len(coef))
   for(z in 1:nrow(X)){
    base<- base + probs[z]*(1 - probs[z])* X[z,]%*%t(X[z,])
   }
   return(base)
}

rest_hess<- solve(hess_func(c(rest_reg$coef, 0, 0, 0), X, once))
>rest_score%*%rest_hess%*%t(rest_score)
[1,] 395.0382
> 1- pchisq(395, 3)
[1] 0
```

Comparing The Three Tests

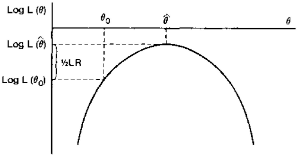


Figure 1. The Likelihood Ratio Test

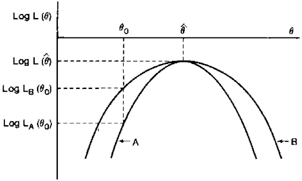


Figure 2. The Wald Test

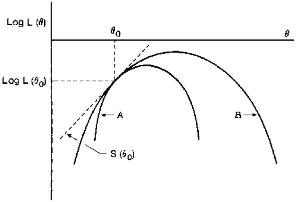


Figure 3. The Lagrange Multiplier Test

- All asymptotically equivalent
- But can be quite different in small samples

	Pros	*Cons*
LR	Most powerful (Neyman-Pearson)	Must compute both $\hat{\theta}_{UR}$ and $\hat{\theta}_R$ Cannot be easily robustified
W	Only need $\hat{\theta}_{UR}$ Easily robustified by sandwich	Not invariant to transformation (e.g. $\theta_1/\theta_2 = 1$ vs. $\theta_1 = \theta_2$)
LM	Only need $\hat{\theta}_R$	$\hat{\theta}_R$ often difficult to estimate

Comparing The Three Tests

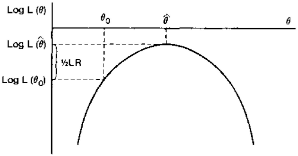


Figure 1. The Likelihood Ratio Test

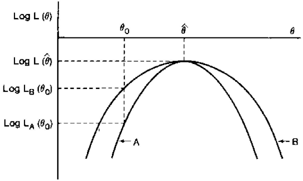


Figure 2. The Wald Test

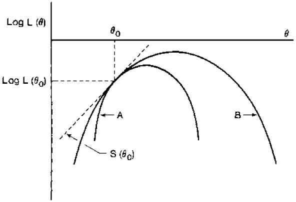


Figure 3. The Lagrange Multiplier Test

- All asymptotically equivalent
- But can be quite different in small samples

	Pros	*Cons*
LR	Most powerful (Neyman-Pearson)	Must compute both $\hat{\theta}_{UR}$ and $\hat{\theta}_R$ Cannot be easily robustified
W	Only need $\hat{\theta}_{UR}$ Easily robustified by sandwich	Not invariant to transformation (e.g. $\theta_1/\theta_2 = 1$ vs. $\theta_1 = \theta_2$)
LM	Only need $\hat{\theta}_R$	$\hat{\theta}_R$ often difficult to estimate

Comparing The Three Tests

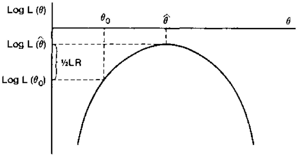


Figure 1. The Likelihood Ratio Test

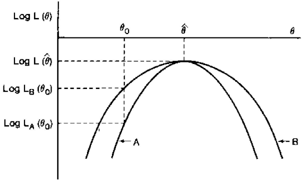


Figure 2. The Wald Test

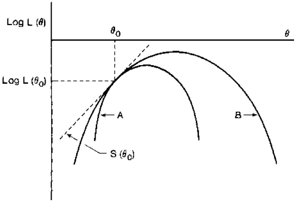


Figure 3. The Lagrange Multiplier Test

- All asymptotically equivalent
- But can be quite different in small samples

	Pros	*Cons*
LR	Most powerful (Neyman-Pearson)	Must compute both $\hat{\theta}_{UR}$ and $\hat{\theta}_R$ Cannot be easily robustified
W	Only need $\hat{\theta}_{UR}$ Easily robustified by sandwich	Not invariant to transformation (e.g. $\theta_1/\theta_2 = 1$ vs. $\theta_1 = \theta_2$)
LM	Only need $\hat{\theta}_R$	$\hat{\theta}_R$ often difficult to estimate

Comparing The Three Tests

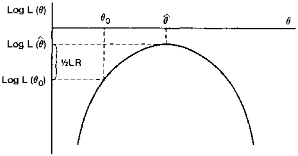


Figure 1. The Likelihood Ratio Test

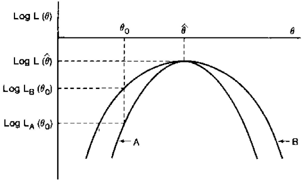


Figure 2. The Wald Test

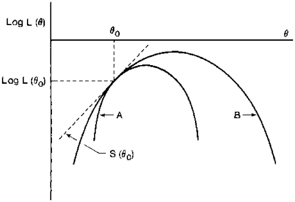


Figure 3. The Lagrange Multiplier Test

- All asymptotically equivalent
- But can be quite different in small samples

	Pros	*Cons*
LR	Most powerful (Neyman-Pearson)	Must compute both $\hat{\theta}_{UR}$ and $\hat{\theta}_R$ Cannot be easily robustified
W	Only need $\hat{\theta}_{UR}$ Easily robustified by sandwich	Not invariant to transformation (e.g. $\theta_1/\theta_2 = 1$ vs. $\theta_1 = \theta_2$)
LM	Only need $\hat{\theta}_R$	$\hat{\theta}_R$ often difficult to estimate

- Model diagnostics
- AIC/BIC
- Cross Validation
- Midterm