STA2201H Methods of Applied Statistics II

Monica Alexander

Week 3: GLM II and Survival Analysis

Overview

- Opioids (assignment 1)
- Binary and categorical data
- Survival analysis intro
- Lab: GLM

Opioid mortality in the US

- Huge increase in deaths involving opioids since 1990s, acceleration since 2010
- Around 70,000 deaths in 2017, mortality rate has increased 7x since 2000
- Essentially monotonic increase, but underlying patterns have changed
  - differences by geography
  - composition of opioids involved in death

Binary data

Binary Responses

We have n random variables $Z_1, \ldots, Z_n$ that are binary

$$Z_i = \begin{cases} 1 \text{ if outcome is a success} \\ 0 \text{ if outcome is a failure} \end{cases}$$

with

$$Pr(Z_1 = 1) = \pi_i$$

so

$$Pr(Z_1 = 0) = 1 - \pi_i$$

Logistic regression

We are interested in describing the probability of success π_i with a linear model

$$g(\pi_i) = \mathbf{x}^\mathsf{T}\beta$$

The **canonical link** is the logistic function, so

$$\text{logit}\ \pi_i = \log \frac{\pi_i}{1-\pi_i} = \mathbf{x}^\mathsf{T}\beta$$

Binomial distribution

Suppose now we are interested in groups of binary outcomes, where groups are defined in such a way that all individuals in a group have identical values of all covariates.

We are interested in the number of successes within that group $\sum_{i=1}^{n_i} Z_i = Y_i$ with group size n_i. This outcome follows a binomial distribution

$$Y_i \sim \text{Binomial}(n_i, \pi_i)$$

Logistic-binary regression

We can model this in the same way as before

$$\begin{aligned} Y_i &\sim \text{Binomial}(n_i, \pi_i) \\ \text{logit } \pi_i &= \mathbf{x}^{\mathbf{T}}\beta \end{aligned}$$

- Binary data can be thought of as a special case of the count data
- Count data can be thought of a special case of the binary data

Latent variable formulation

$$
\begin{aligned}
y_i &= \begin{cases} 1 \text{ if } z_i > 0 \\ 0 \text{ if } z_i < 0 \end{cases} \\
z_i &= X_i\beta + \epsilon_i \\
\epsilon_i &\sim f(.)
\end{aligned}
$$

Latent variable formulation

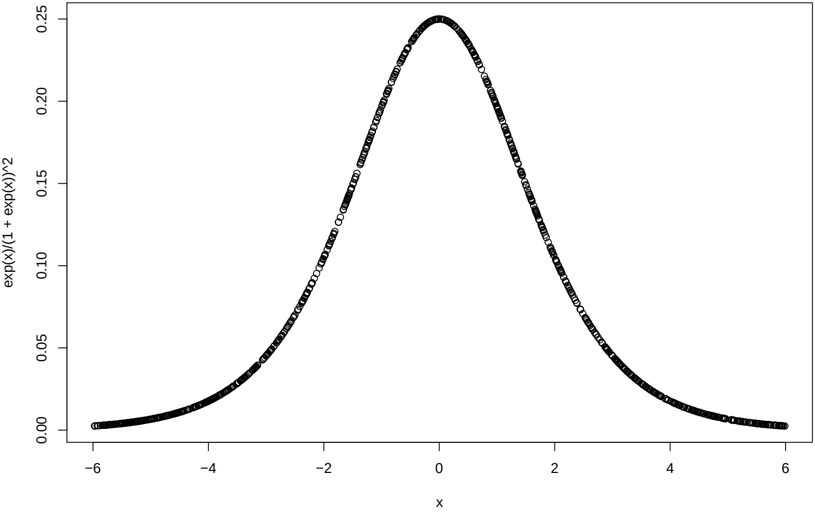
$$
\begin{aligned}
y_i &= \begin{cases} 1 \text{ if } z_i > 0 \\ 0 \text{ if } z_i < 0 \end{cases} \\
z_i &= X_i\beta + \epsilon_i \\
\epsilon_i &\sim f(.)
\end{aligned}
$$

For logistic regression, the errors ϵ have a *logistic* probability distribution

$$p(x) = \frac{e^x}{(1 + e^x)^2}$$

Latent variable formulation

The logistic pdf looks like



Latent variable formulation

Write $\eta_i = X_i\beta$.

Note that

$$\begin{aligned} \pi_i &= Pr(z_i > 0) \\ &= Pr(\epsilon_i > -\eta_i) \\ &= 1 - F(-\eta_i) \\ &= F(\eta_i) \end{aligned}$$

For the logistic, $F(\eta_i) = \frac{e^x}{(1+e^x)}$ so $\eta_i = F^{-1}(\pi_i) = \frac{\pi_i}{1-\pi_i}$ as before.

Probit regression

Any transformation that maps probabilities into the real line could be used to produce a generalized linear model, as long as the transformation is one-to-one, continuous and differentiable.

- We could also make errors normal

$$\epsilon \sim N(0, 1)$$

This implies

$$\pi_i = \Phi(\eta_i)$$

or

$$\Phi^{-1}(\pi_i) = \mathbf{X_i}\beta$$

where Φ is the standard normal cdf. This form is called **probit**. What's the interpretation of the β's?

Example: contracpetive use

Data set on contraceptive use in Fiji (source)

What the data look like:

age	education	wantsMore	notUsing	using
<25	low	yes	53	6
<25	low	no	10	4
<25	high	yes	212	52
<25	high	no	50	10
25-29	low	yes	60	14
25-29	low	no	19	10

Try a simple model: Using $\sim$ Age $+$ Desire

Example

Logit link:

```
##
## Call:
## glm(formula = cbind(using, notUsing) ~ age + wantsMore, family = binomial(link = "logit"),
##     data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7870  -1.3208  -0.3417   1.2346   2.4577
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.8698     0.1571  -5.536 3.10e-08 ***
## age25-29       0.3678     0.1754   2.097    0.036 *
## age30-39       0.8078     0.1598   5.056 4.27e-07 ***
## age40-49       1.0226     0.2039   5.014 5.32e-07 ***
## wantsMoreyes  -0.8241     0.1171  -7.037 1.97e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 165.772  on 15  degrees of freedom
## Residual deviance:  36.888  on 11  degrees of freedom
## AIC: 118.4
##
## Number of Fisher Scoring iterations: 4
```

What's the interpretation of the `wantsMore` coefficient?

Example

Probit link:

```
##
## Call:
## glm(formula = cbind(using, notUsing) ~ age + wantsMore, family = binomial(link = "probit"),
##     data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8352  -1.3411  -0.3773   1.2834   2.4893
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.51535    0.09178  -5.615 1.97e-08 ***
## age25-29      0.20861    0.10071   2.071   0.0383 *
## age30-39      0.46856    0.09267   5.056 4.27e-07 ***
## age40-49      0.60487    0.12207   4.955 7.23e-07 ***
## wantsMoreyes -0.49646    0.07102  -6.991 2.73e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 165.772  on 15  degrees of freedom
## Residual deviance:  38.261  on 11  degrees of freedom
## AIC: 119.77
##
## Number of Fisher Scoring iterations: 4
```

What's the interpretation of the `wantsMore` coefficient?

Comparison

- $\mathbf{X}\beta$ refers to change in z-score
- Not overly intuitive, but then again what are odds ratios
- Can convert between the two: divide by $\pi/\sqrt{3}$
- ... in both cases might be better off converting to the original (probability) scale

Categorical data

Categorical data/multinomial responses

- Extension of binomial / binary outcomes.
- Now Y_i make take one of several discrete values, $1, 2, \ldots, J$.
- Now the probability is

$$\pi_{ij} = Pr(Y_i = j)$$

with

$$\sum_j \pi_{ij} = 1$$

- As before, for grouped data, n_i is the number of cases in the ith group and y_{ij} is the number of responses that fall in jth category, so the vector of categories $\mathbf{y_i}$ is a of counts that add up to n_i.
- For individual data, $n_i = 1$ and y_{ij} is 0 or 1, so the vector of categories $\mathbf{y_i}$ is a vector of 0s or 1s.

Multinomial distribution

The probability distribution of the counts Y_{ij} given the total n_i is given by the multinomial distribution

$$Pr\{Y_{i1} = y_{i1}, \dots, Y_{iJ} = y_{iJ}\} = \binom{n_i}{y_{i1}, \dots, y_{iJ}} \cdot \pi_{i1}^{y_{i1}} \dots \pi_{ij}^{y_{iJ}}$$

Can this be represented as exponential family?

Conditional distribution

- Let $Y_1, \dots Y_J$ be Poisson with rate λ_j
- Let $n = \sum Y_j$, which is Poisson with rate $\sum_j \lambda_j$
- Multinomial distribution is joint distribution of Poisson, conditional on sum.

Multinomial regression

- Easy extension to binomial model if we model with respect to a reference category J

$$\eta_{ij} = \log \frac{\pi_{ij}}{\pi_{iJ}} = \mathbf{x_i}^{\mathbf{T}} \beta$$

for $j = 1, \dots J - 1$.

- Note that if $J = 2$ we have the usual logistic regression
- Coefficients can be interpreted as before, but OR are in relation to reference category

Convert to probabilities

$$\pi_{ij} = \frac{\exp(\eta_{ij})}{\sum_k \exp\left(\eta_{ik}\right)} = \mathsf{softmax}(\eta)_i$$

- Choice of reference category would affect βs but not probabilities

Ordered response

What if our categories are ordered? e.g. survey responses are often on an ordinal scale. As before,

$$\pi_{ij} = Pr(Y_i = j)$$

Now consider cumulative probability

$$\gamma_{ij} = Pr(Y_i < j)$$

so

$$\gamma_{ij} = \pi_{i1} + \pi_{i2} + \cdots + \pi_{ij}$$

Model is of the form

$$g(\gamma_{ij}) = \theta_j + \mathbf{x_i}^{\mathbf{T}} \beta$$

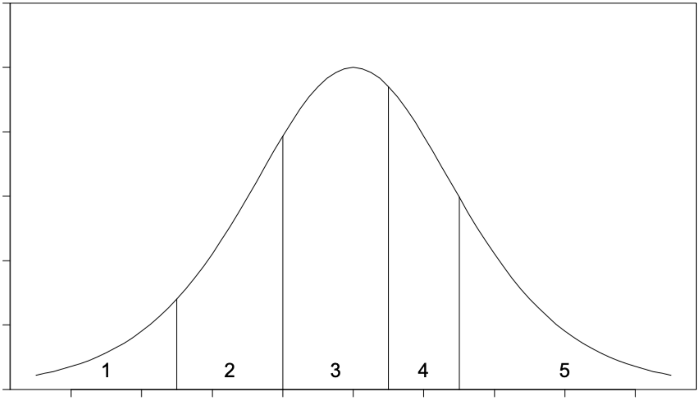
Here θ_j is a constant representing the baseline value of the transformed cumulative probability for category j.

Ordinal regression

Alternatively, can think of a latent variable set-up with cut-points $\theta_1, \ldots, \theta_J$

$$\begin{aligned} y_i &= \begin{cases} 1 \text{ if } z_i < \theta_1 \\ 2 \text{ if } z_i \in (\theta_1, \theta_2) \\ \ldots \\ J \text{ if } z_i > \theta_{J-1} \end{cases} \\ z_i &= X_i \beta + \epsilon_i \\ \epsilon_i &\sim f(.) \end{aligned}$$

Ordinal regression



Ordinal regression

From the latent formulation

$$\begin{aligned} \gamma_{ij} &= Pr(Y_i < j) \\ &= Pr(z_i < \theta_j) \\ &= Pr(e_i < \theta_j - X_i\beta) \\ &= F(\theta_j - \mathbf{x_i^T}\beta) \end{aligned}$$

so

$$g(\gamma_{ij}) = F^{-1}(\theta_j - \mathbf{x_i^T}\beta)$$

as before.

Proportional odds model

Like a logistic regression, but applied to the cumulative probabilities

$$\log \frac{\gamma_{ij}}{1-\gamma_{ij}} = \theta_j + \mathbf{x_i^T}\beta$$

or

$$\frac{\gamma_{ij}}{1-\gamma_{ij}} = \lambda_j \exp(\mathbf{x_i^T}\beta)$$

λ_j is baseline odds of response being in category j.

Pretty strong assumption of proportional odds!

Example

Housing Conditions in Copenhagen

housing	influence	contact	satisfaction	n
tower	low	low	low	21
tower	low	low	medium	21
tower	low	low	high	28
tower	low	high	low	14
tower	low	high	medium	19
tower	low	high	high	37

Example

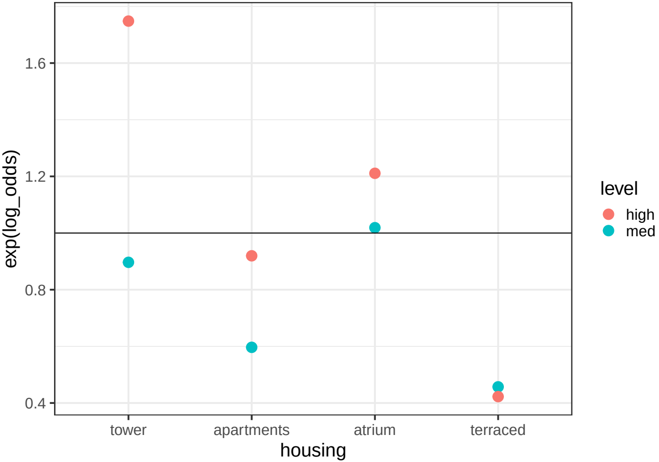
First let's do a multinomial regression, with just housing and contact:

```
## # weights:  18 (10 variable)
## initial  value 1846.767257
## iter  10 value 1793.932058
## final  value 1789.600661
## converged
```

```
## Call:
## nnet::multinom(formula = Y ~ housing + contact, data = copen_wide)
##
## Coefficients:
##            (Intercept) housingapartments housingatrium housingterraced
## sat_medium  -0.1091063         -0.407446     0.1278116      -0.6738718
## sat_high     0.5586042         -0.642400    -0.3672630      -1.4199239
##            contacthigh
## sat_medium   0.3005283
## sat_high     0.3334568
##
## Std. Errors:
##            (Intercept) housingapartments housingatrium housingterraced
## sat_medium   0.1524817         0.1713221     0.2217222       0.2051505
## sat_high     0.1330480         0.1501078     0.2048673       0.1947044
##            contacthigh
## sat_medium   0.1306991
## sat_high     0.1190333
##
## Residual Deviance: 3579.201
## AIC: 3599.201
```

Multinomial regression

Plot the result odds ratios (cf low satisfaction, for low contact)



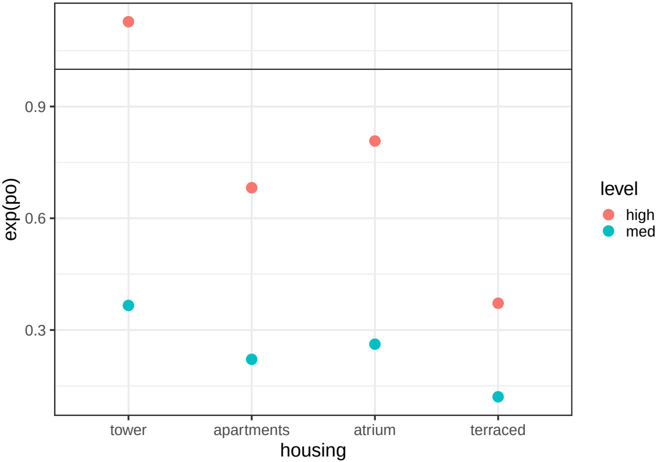
Proportional odds model

Now fit the same idea but with a proportional odds model (ordinal)

```
## Call:
## MASS::polr(formula = satisfaction ~ housing + contact, data = copen,
##     weights = n)
##
## Coefficients:
##                     Value Std. Error t value
## housingapartments -0.5030     0.1169  -4.304
## housingatrium     -0.3341     0.1518  -2.201
## housingterraced   -1.1093     0.1493  -7.428
## contacthigh        0.2540     0.0934   2.720
##
## Intercepts:
##             Value   Std. Error t value
## low|medium  -1.0053  0.1077    -9.3325
## medium|high  0.1202  0.1048     1.1465
##
## Residual Deviance: 3587.389
## AIC: 3599.389
```

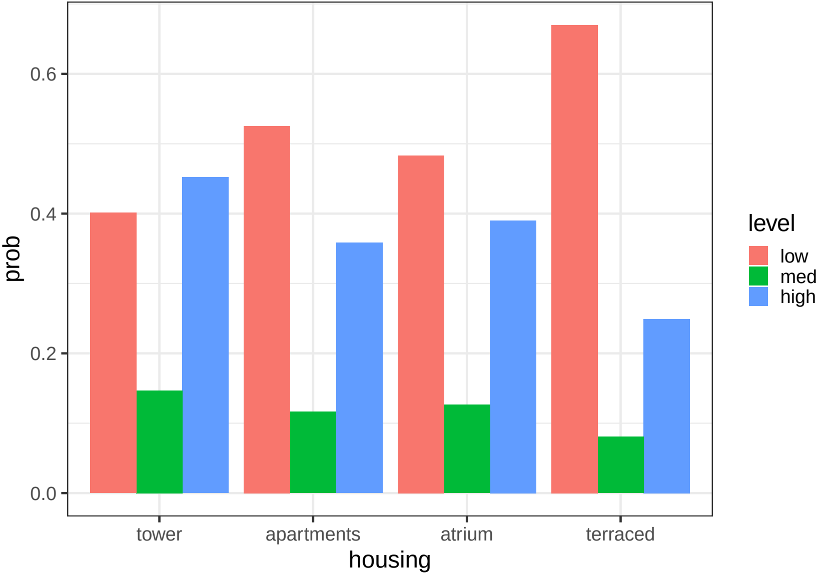
Ordinal regression

Plot of odds ratios



Ordinal regression

Convert log-odds to probabilities:



Summary

- Multinomial models are a natural extension to binomial models
- Looked at logistic forms, but easy to go probit (or other)
- Interpretation is often easiest when we convert to the natural scale

Survival analysis

Introduction

- Interested in the **waiting time** to an event / outcome
- Terminology is all around survival and death, but can be used to study any sort of waiting time
  - time to first birth
  - time to leaving home
  - time to finishing PhD :)
- Increasing amount of information considered (not just looking at end outcome)

Goals:

- Analyse waiting times wrt covariates
- Adjust for potential censoring or truncation

Survival analysis is a suite of methods to do this including parametric, semi-parametric and non-parametric methods.

Definitions

Let T be a non-negative random variable representing the waiting time to an event of interest.

- Assume T is continuous
- Define the pdf of T as $f(t)$
- cdf is $P(T < t) = F(t)$
- **Survival function** $P(T \geq t) = 1 - F(t) = S(t)$

Definitions

The **hazard rate** $\lambda(t)$ is the instantaneous rate of occurrence

$$\lambda(t) = \lim_{dt \to 0} \frac{Pr(t \leq T < t + dt | T \geq t)}{dt}$$

which is

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log(S(t))$$

Definitions

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\log(S(t))$$

implies

$$S(t) = \exp\left(\int_0^t \lambda(x)dx\right) = \exp\left(\Lambda(t)\right)$$

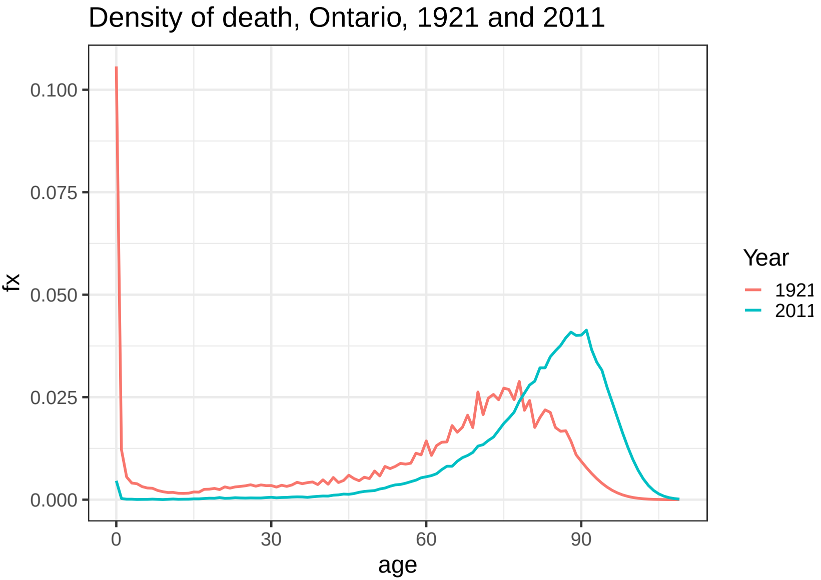
where $\Lambda(t)$ is the cumulative hazard = the sum of risks one faces up to time t.

Given either the hazard rate or survival function, you can get everything else.

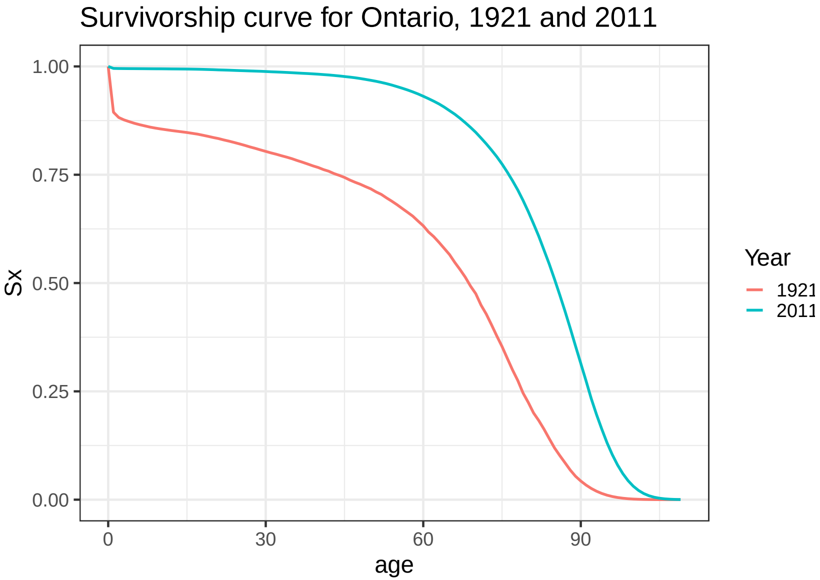
- What is $E(T)$?

Characteristics shapes

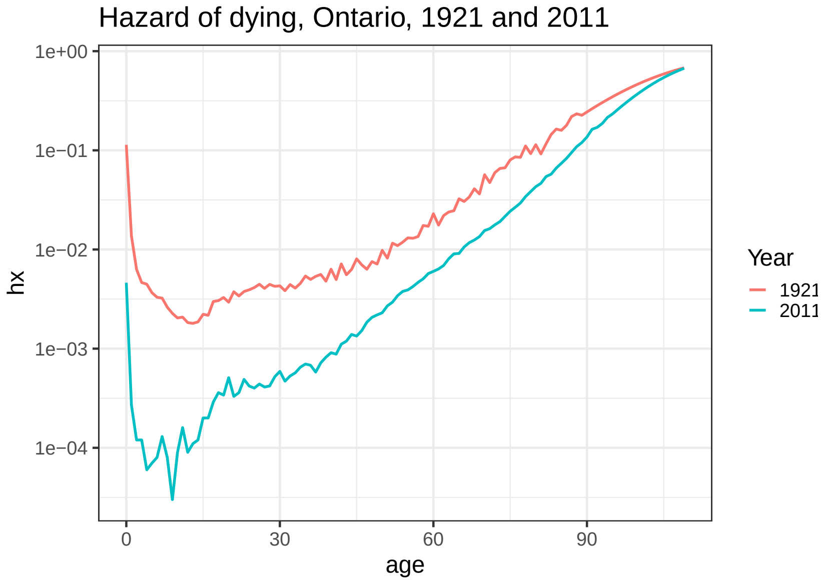
Characteristics shapes



Characteristics shapes



Characteristics shapes



Example: constant hazards

$$S(t) = \exp\left(\int_0^t \lambda(x)dx\right) = \exp\left(\Lambda(t)\right)$$

- The simplest case is if $\lambda(t) = \lambda$ for all t
- Constant hazard of dying/event occurring

This implies

$$S(t) = \exp(-\lambda t)$$

and

$$f(t) = \lambda \exp(-\lambda t)$$

What is this distribution?

Likelihood

Individuals i, $i = 1, \ldots, n$

- By any particular time t_i, i is either alive or dead
- If they are alive, they are **censored**
- Contribution to likelihood if died: $f(t_i) = \lambda(t_i)S(t_i)$
- Contribution to likelihood if alive: $S(t_i)$

Likelihood is then

$$L = \prod_i L_i = \prod_i \lambda(t_i)^{d_i} S(t_i)$$

d_i is indicator of whether or not individual died.

and LL is

$$\log L = \sum_i d_i \log(\lambda(t_i)) - \Lambda(t_i)$$

Example: constant hazards

If $\lambda(t) = \lambda$ then

$$\log L = \sum_i d_i \log \lambda - \sum_i t_i \lambda$$

So MLE for λ is ?

Example: constant hazards

$$\hat{\lambda} = \frac{\sum d_i}{\sum t_i}$$

- if nothing is censored this is just Exponential

Example: constant hazards

Instead of looking at waiting times, look at deaths $\sum_i d_i = D$ and assume

$$D \sim \text{Poisson}(\lambda T)$$

What's the MLE?

$$\log L = \sum_i d_i \log \lambda - \sum_i t_i \lambda + \text{constant}$$

- Waiting times Exponential $==$ deaths are Poisson
- Helps to use GLM!

Lab