GOV 2001/ 1002/ E-2001 Section 8[1]
Overdispersion and Introduction to Duration Models

Mayya Komisarchik

Harvard University

March 30, 2016

[1] Heartfelt thanks to all of the Gov 2001 TFs of yesteryear; this section draws heavily upon materials created by Stephen Pettigrew, Solé Prillaman, and Patrick Lam.

Outline

Logistics

- Readings for this week
  - "Presidential Appointments to the Supreme Court: Adding Systematic Explanation to Probabilistic Description". Gary King. 1987.
  - "Event Count Models for International Relations: Generalizations and Applications". Gary King. 1989.
  - "A Unified Model of Cabinet Dissolution in Parliamentary Democracies". Gary King, James Alt, Nancy Burns, Michael Laver. 1990.
  - "Time Is of the Essence: Event History Models in Political Science". Janet M. Box-Steffensmeier and Bradford S. Jones. 1997. (Optional, but fun!)
- Assignment Memo
  - 1 page
  - Look for evidence of model dependence in the paper that you are replicating for your final project
    - Don't worry if you don't end up finding any! Just go through the exercise!

What is Dispersion?

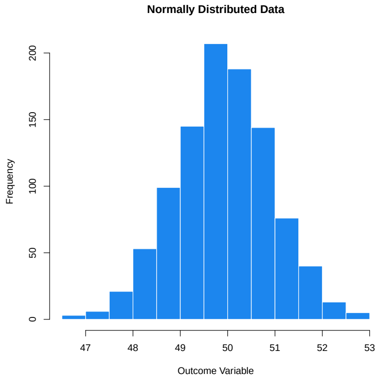
- Dispersion, in this context, refers to the relationship between the mean and variance of our chosen probability distribution.
- **Overdispersion** occurs when there is greater variability in the data than expected given the distribution assumed to produce the data
  - $E[Y|X] < Var[Y|X]$
- **Underdispersion** occurs when there is less variability in the data than expected given the distribution assumed to produce the data
  - $E[Y|X] > Var[Y|X]$
- Overdispersion and underdispersion (though that's much rarer) represent violations of your modeling assumptions

When do we Encounter Overdispersion?

- Overdispersion isn't unique to the Poisson.
- We run into overdispersion problems whenever we select a model that restricts the variance of our data to a level below what it actually is.

WHEN DO WE ENCOUNTER OVERDISPERSION?

For instance, imagine that our data looked like this:

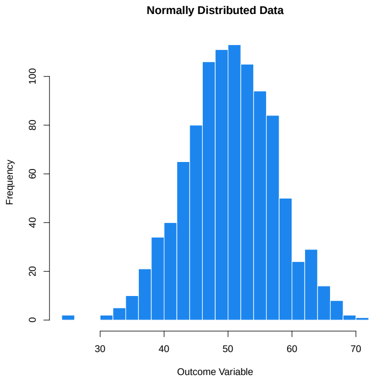


Let's say that $Y \sim N(\mu, 1)$, where $\mu \approx 50$. Could we summarize this data using a stylized normal using just one parameter, μ?

In this case, yes.

WHEN DO WE ENCOUNTER OVERDISPERSION?

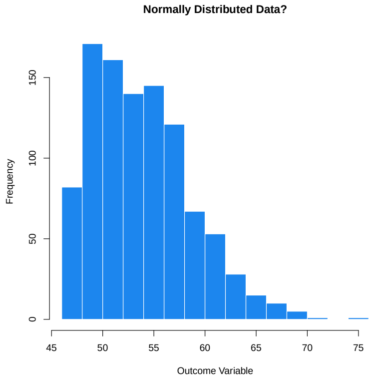
But what if our data looked like this:



Could we summarize this data using a stylized normal using just one parameter, μ? Why? No. Compare this to the last histogram. The variance here is clearly > 1.

Other Violations of Modeling Assumptions

What if our data looked like this:



What assumptions might we violate if we modeled this as a normal distribution with $N(\mu, \sigma^2)$?By definition, the skewness of a normal distribution = 0.

How do we Address Overdispersion?

- Generally, correcting for overdispersion requires you to relax the assumptions you're making about the variance.
- Specifically, you'll want to choose another model that adds the parameters which describe the features of your data your original model missed.
  - In the case we just saw, we could model our data using the skew normal distribution, which is the normal distribution with an additional parameter for skewness.
- Canonical model corrections for overdispersion:
  - Poisson $\rightarrow$ Negative Binomial
  - Binomial $\rightarrow$ Beta Binomial

How do we Address Overdispersion?

WARNING

Remember that you can add parameters to your model until you're representing every individual data point, but that's about as helpful as handing your readers a spreadsheet with your data on it. Strike a balance between underfitting and overfitting.

Binary Variables

Incidentally, why don't we run into overdispersion problems with binary outcome variables (i.e. models with Bernoulli stochastic components)?

That is, why is it that binary outcome variables will have the correct variance with only one parameter, π, when exponential or Poisson models will not?

Because π fully characterizes the distribution! Knowing π gives you the mean and the variance of the distribution. There is no other information you can get out of your data, so adding additional parameters doesn't actually give you additional information.

Count Models

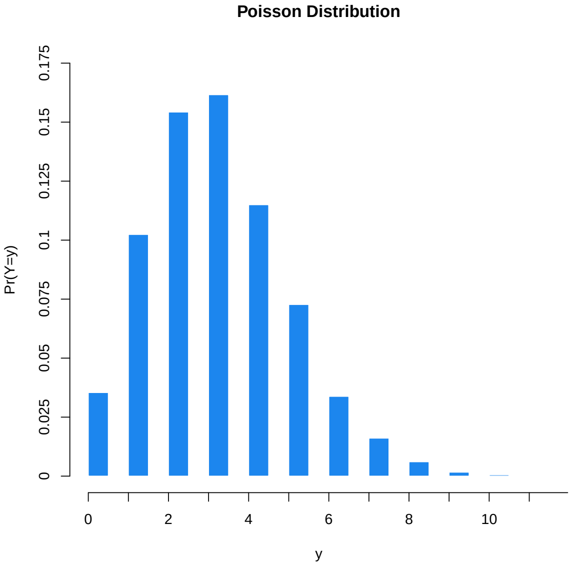
I count the spiders on the wall
I count the cobwebs in the hall
I count the candles on the shelf
When I'm alone, I count myself!

Poisson

Let's look at a specific example of overdispersion through the lens of a Poisson example. First, some Poisson review:

- The Poisson is a discrete distribution that gives the probability that some number of events will occur in a fixed period of events or space.
- PMF: $P(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}$
- Suppose $Y \sim \text{Pois}(3)$ What is $P(Y = 4)$?
- $P(Y = 4) = \frac{3^4}{4!} e^{-3} = 0.168$

Poisson



Poisson Distribution
Pr(Y=y)
0
0.025
0.05
0.075
0.1
0.125
0.15
0.175
0
2
4
6
8
10
y

Poisson: Mean

- $E[Y] = \sum_{y=0}^{\infty} y \cdot P(Y = y)$
- $E[Y] = \sum_{y=0}^{\infty} y \frac{\lambda^y}{y!} e^{-\lambda}$
- $E[Y] = \sum_{y=0}^{\infty} \frac{\lambda^y}{(y-1)!} e^{-\lambda}$
- $E[Y] = \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^{y-1}}{(y-1)!}$; let $j = y - 1$
- $E[Y] = \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^j}{(j)!}$
- $E[Y] = \lambda e^{-\lambda} e^{\lambda} = \lambda$

$$E[Y] = \lambda$$

Poisson: Variance

- $Var[Y] = E[Y^2] - E[Y]^2$
- $E[Y^2] = \sum_{y=0}^{\infty} y^2 \frac{\lambda^y}{y!} e^{-\lambda}$
- $E[Y^2] = \sum_{y=0}^{\infty} y(y-1) \frac{\lambda^y}{y!} e^{-\lambda} + \sum_{y=0}^{\infty} y \frac{\lambda^y}{y!} e^{-\lambda}$
- $E[Y^2] = \sum_{y=0}^{\infty} y(y-1) \frac{\lambda^y}{y!} e^{-\lambda} + \lambda$
- $E[Y^2] = e^{-\lambda} \lambda^2 \sum_{y=0}^{\infty} \frac{\lambda^{y-2}}{(y-2)!} + \lambda$
- $E[Y^2] = e^{-\lambda} \lambda^2 \sum_{y=0}^{\infty} \frac{\lambda^{y-2}}{(y-2)!} + \lambda$; let $k = y - 2$
- $E[Y^2] = e^{-\lambda} \lambda^2 \sum_{y=0}^{\infty} \frac{\lambda^k}{k!} + \lambda = e^{-\lambda} \lambda^2 e^{\lambda} + \lambda = \lambda^2 + \lambda$
- $Var[Y] = (\lambda^2 + \lambda) - \lambda^2$

$$Var[Y] = \lambda$$

Poisson: Assumptions

- As we see from the previous two slides, the Poisson assumes that $E[Y_i|U_i] = Var[Y_i|U_i] = \lambda_i$
- Note that the Poisson also assumes that events are independent, so the probability that some individual event i occurs does not affect the probability that another individual event $i+1$ occurs.
- That's what a constant λ implies. It varies with covariates, but within a unit given those covariates individual events occur at a constant underlying rate.

Poisson: Terrorism Example

Let's look at a univariate Poisson example:

Outcome Variable: The number of deaths from terrorist attacks in a particular country over a 30 year period.

Explanatory Variable: The unemployment rate in that country for each month in the 30 year period.

Poisson: Terrorism Example

First, let's write down the model.

Stochastic Component: $Y_i \sim Pois(\lambda_i)$
Where Y is the number of terrorist attacks in a month.

Systematic Component: $\lambda_i = E(Y_i|U_i) = e^{\beta_0+\beta_1 U_i}$
Where U_i is the unemployment rate in a given month.

Link Function: $\lambda_i = E(Y_i|U_i) = e^{U_i \beta}$

Remember that λ_i is a rate parameter that has to be > 0, and $e^{U_i \beta}$ bounds λ_i to be > 0.

Poisson: Terrorism Example

Now, let's derive the log-likelihood of the Poisson:

If the PMF of the Poisson is:

$$\Pr(Y = y) = \frac{\lambda^y}{y!}e^{-\lambda}$$

then the likelihood is:

$$L(\lambda|y) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i}}{y_i!}e^{-\lambda_i}$$

And the log-likelihood is:

$$\begin{aligned}
\ell(\lambda|y) &= \log\left[\prod_{i=1}^{n} \frac{\lambda_i^{y_i}}{y_i!}e^{-\lambda_i}\right] \\
&= \sum_{i=1}^{n} \log\left[\frac{\lambda_i^{y_i}}{y_i!}e^{-\lambda_i}\right]
\end{aligned}$$

POISSON: TERRORISM EXAMPLE

Now, let's derive the log-likelihood of the Poisson:

$$
\begin{aligned}
&= \sum_{i=1}^{n} \log \left[\frac{\lambda_i^{y_i}}{y_i!} \mathrm{e}^{-\lambda_i} \right] \\
&= \sum_{i=1}^{n} y_i \log \lambda_i - \log(y_i!) - \lambda_i \\
&\propto \sum_{i=1}^{n} y_i \log \lambda_i - \lambda_i \\
\ell(\beta|y,U) &\propto \sum_{i=1}^{n} y_i \log(exp(U_i\beta) - exp(U_i\beta) \\
&\propto \sum_{i=1}^{n} y_i U_i\beta - exp(U_i\beta)
\end{aligned}
$$

Poisson: Terrorism Example

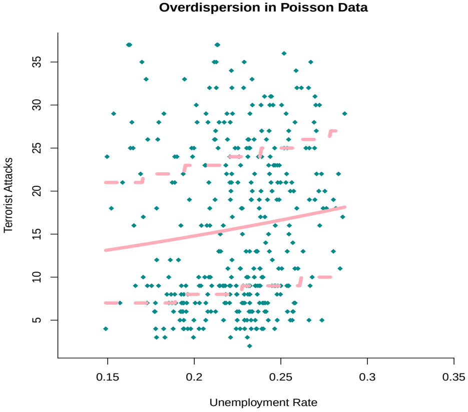
And estimate our MLE for β using optim:

```
pois.ll <- function(par, y, x){
  x <- as.matrix(cbind(1, x))
  xb <- x %*% par
  sum(y * xb - exp(xb))
}

opt.pois <- optim(par = rep(0,2),
                  fn = pois.ll,
                  x = attacks$unemploy,
                  y = attacks$deaths,
                  hessian = T,
                  method = "BFGS",
                  control = list(fnscale = -1))
```

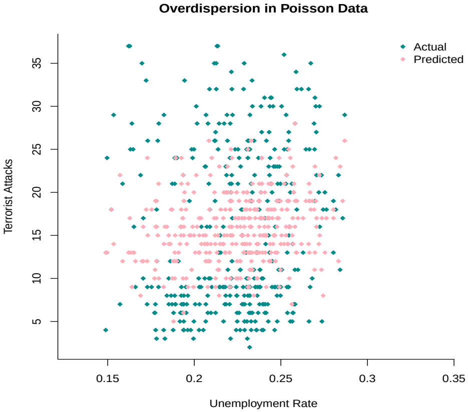
Poisson: Terrorism Example

We know that the Poisson restricts $E[Y_i|U_i] = Var[Y_i|U_i]$, which is problematic if our variance is actually greater than our mean (overdispersion). Let's diagnose that. Does it look like 95% of our data falls within our confidence interval for points?



Poisson: Terrorism Example

When you're relying on more than 1 covariate, you can generate predicted values from your model, plot them against your actual data, and see whether there is greater spread in your data than in your predicted values:



Poisson: Terrorism Example

You can also do some other formal tests:

Ratio of Residual Deviance to Degrees of Freedom:

```
# You can also run your model as glm() object and check the ratio of your residual deviance
# (data variance not explained by your model) to your
# degrees of freedom (reduced by # params you've calculated)

poismod <- glm(deaths ~ unemploy, family = "poisson", data = attacks)
deviance(poismod)/poismod$df.residual
# a value much > 1 here suggests overdispersion, and we have that problem
[1] 5.586653
```

AER Package Dispersion Test:

```
require(AER)
dispersiontest(poismod)
Overdispersion test

data:  poismod
z = 15.263, p-value < 2.2e-16
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
5.560199
```

Poisson: Terrorism Example

Now we know we have overdispersion in our data. What do we do now?

We know that we want to relax the assumption that $E[Y_i|U_i] = Var[Y_i|U_i] = \lambda_i$, so our next task is to derive a different distribution that allows λ_i to vary with some other parameter.

Let's just introduce a new parameter, ζ, and let λ_i vary with ζ:

Poisson: Terrorism Example

Let's derive the negative binomial model by adding a parameter to the Poisson.

Here's the new stochastic component:

$$
\begin{aligned}
Y_i|\lambda_i, \zeta_i &\sim Poisson(\zeta_i \lambda_i) \\
\zeta_i &\sim Gamma\left(\frac{1}{\sigma^2 - 1}, \frac{1}{\sigma^2 - 1}\right)
\end{aligned}
$$

Note that this Gamma distribution has a mean of 1. Therefore, $Poisson(\zeta_i \lambda_i)$ has mean λ_i.

Note that the variance of this new Poisson distribution is $\sigma^2 - 1$.

This means that as σ^2 goes to 1, the distribution of ζ_i collapses to a spike over 1.

Poisson: Terrorism Example

Using a similar approach to that described in UPM pgs. 51-52 we can derive the marginal distribution of Y as:

$$Y_i \sim Negbin(\lambda_i, \sigma^2)$$

where the PMF is:

$$f_{nb}(y_i|\lambda_i, \sigma^2) = \frac{\Gamma(\frac{\lambda_i}{\sigma^2-1} + y_i)}{y!\Gamma(\frac{\lambda_i}{\sigma^2-1})} \left(\frac{\sigma^2 - 1}{\sigma^2}\right)^{y_i} (\sigma^2)^{-\frac{\lambda_i}{\sigma^2-1}}$$

Notes:

1. $\lambda_i > 0$ and $\sigma^2 > 1$
2. $\mathrm{E}[Y_i] = \lambda_i$ and $\mathrm{Var}[Y_i] = \lambda_i \sigma^2$. What value of σ^2 would be evidence *against* overdispersion?
3. We still have the same old systematic component: $\lambda_i = exp(U_i \beta)$.

Poisson: Terrorism Example

How do we implement this?

If you said "find the log likelihood and then use `optim()`", then you win!

So what is the log likelihood of this thing?

Poisson: Terrorism Example

It looks like this:

$$
\begin{aligned}
L(\lambda, \sigma^2 | y_i) &= \prod_i^N \frac{\Gamma(\frac{\lambda}{\sigma^2-1} + y_i)}{y_i!\Gamma(\frac{\lambda}{(\sigma^2-1)})} (\frac{\sigma^2 - 1}{\sigma^2})^{y_i} (\sigma^2)^{\frac{-\lambda}{\sigma^2-1}} \\
\ell(\phi, \sigma^2 | y_i) &= \sum_i^N log\left(\frac{\Gamma(\frac{\lambda}{\sigma^2-1} + y_i)}{y_i!\Gamma(\frac{\lambda}{(\sigma^2-1)})} (\frac{\sigma^2 - 1}{\sigma^2})^{y_i} (\sigma^2)^{\frac{-\lambda}{\sigma^2-1}} \right) \\
&= \sum_i^N log\Gamma(\frac{\lambda}{\sigma^2 - 1} + y_i) - logy_i! - log\Gamma(\frac{\lambda}{(\sigma^2 - 1)}) + \\
&\quad y_i log(\sigma^2 - 1) - y_i log(\sigma^2) - \frac{\lambda}{\sigma^2 - 1} log(\sigma^2)
\end{aligned}
$$

Poisson: Terrorism Example

In R:

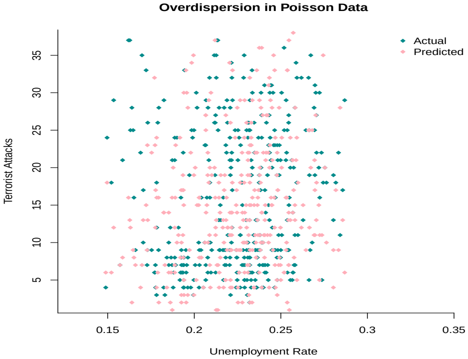
```
ll.nb <- function(par, y, X){
  beta <- par[1:ncol(X)]
  lambda <- exp(X%*%beta)
  sigma2 <- exp(par[ncol(X)+1]) + 1
  sum(lgamma(lambda/(sigma2 - 1) + y) -
  lgamma(lambda/(sigma2-1)) +
  y*log((sigma2-1)/sigma2) -
  (lambda/(sigma2-1))*log(sigma2))
}
```

OR:

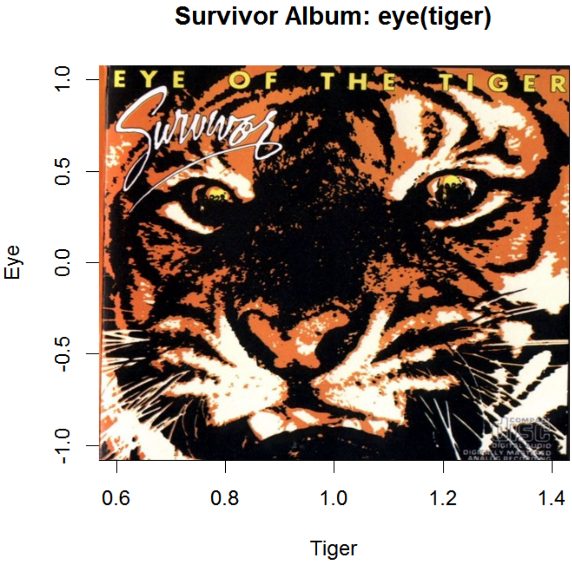
```
library(MASS)
mod.nb <- glm.nb(deaths ~ unemploy, data = attacks)
```

Poisson: Terrorism Example

Did we solve the problem?



Looks that way! If we drew from a negative binomial distribution with the parameters we estimated from our data, then we'd be predicting as much variation as we actually have. That's good! We're not overdispersed.


Survivor Album: eye(tiger)
EYE OF THE TIGER
Survivor
Eye
1.0
0.5
0.0
-0.5
-1.0
0.6
0.8
1.0
1.2
1.4
Tiger

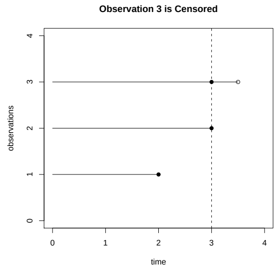
When Do We Use Duration Models?

- First, note that the terms "survival models", "duration models", and "event history models" all refer to the same class of models.
- In a duration model, the dependent variable, Y, represents the duration of time units spend in a particular state (peace, health, negotiations, etc.) before an event occurs (war, death, treaty).
- These are used widely across different disciplines:
  - Biostatistics: how long until a sick patient dies of a disease?
  - Engineering: how long until a mechanical component fails?
  - Political Science: How long do coalition governments last? How long will a regime stay in power? How long until a legislator leaves office?
- Your observations should be in the same temporal units (that is, don't list event times for some observations in days and others in years).

Why Not Just Use OLS?

There are several reasons we don't want to use OLS to estimate durations:

- OLS assumes that $Y|x \sim$ Normal. But if your dependent variable Y is actually a duration, then $Y|x > 0$ since it's some number of years, days, hours, etc.
- Duration models can handle censoring in your observations



Here, observation 3 is censored, which means that it has not experience the event at the time you stop collecting data (like a graduate student who hasn't graduated by 2000 in our class example). We don't know its true duration.

Why Not Just Use OLS?

- Duration models can handle time-varying covariates.
  - Imagine that Y is the duration of a political regime. The GDP for the country ruled by that political regime may change throughout the duration of the regime.
  - But OLS does not allow for multiple values of GDP per observation.
  - Note that you can set your data up in a way that lets you accommodate time-varying covariates (we're not getting into this today, but it's relies on some of the same principles as censoring observations).

Survival Models: Basic Structure

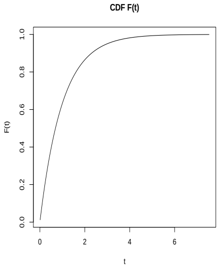
- Let T be a continuous, positive random variable that represents duration or survival times, such that $T = Y$
- T has some underlying probability density function (PDF), $f(t)$
  - This represents roughly the probability that an event occurs at time $T = t$ (since $f(t)$ is continuous, we'd have to integrate it over some infinitely small interval $[a, b]$ to get the probability that an event occurs in the time window $[a, b]$).

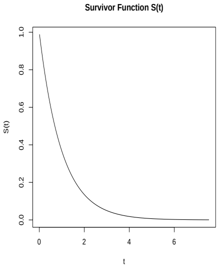
SURVIVAL MODELS: BASIC STRUCTURE

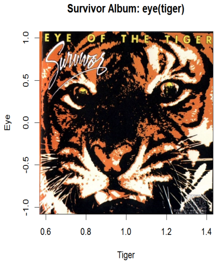
F(t): the CDF of $f(t)$, $\int_0^t f(u)du = P(T \leq t)$, which is the probability of an event occurring before (or at exactly) time t

Survivor function: The probability of surviving (i.e. no event occuring) until at least time t: $S(t) = 1 - F(t) = P(T > t)$

Eye of the Tiger: 1982 album by the band Survivor, which reached number 2 on the US Billboard 200 chart.







Hazard Rate

- **Hazard Rate**: (or hazard function). The hazard function for T, or $h(t)$, is roughly the probability of an event occurring at time t given that the unit has survived to time t
- $h(t) = P(\text{event at t}|\text{survival through t})$
- $h(t) = \frac{P(\text{survival through t}|\text{event at t}) \cdot P(\text{event at t})}{P(\text{survival through t})}$
- $h(t) = \frac{P(\text{event at t})}{P(\text{survival up to t})}$
- $h(t) = \frac{f(t)}{S(t)}$

Relating the Density, Survival, and Hazard Functions

$$\underbrace{f(t)}_{\text{density function}} = \underbrace{h(t)}_{\text{hazard function}} \cdot \underbrace{S(t)}_{\text{survival function}}$$

Modeling Duration Times with Covariates

We can model the mean of the duration times in our data as a function of covariates by using a link function, $g(\cdot)$, where:

$$g(E[T_i]) = X_i\beta$$

Then we can estimate our βs using the same maximum likelihood estimation process we have been doing throughout this course!

Modeling Duration Times with Covariates

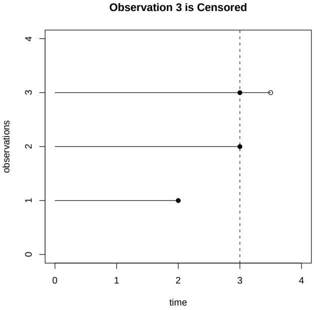
This is the general procedure for performing MLE with a duration model:

- Make the assumption that T follows a specific distribution, $f(t)$ (that is, choose your stochastic component. Common selections include: exponential, Cox proportional hazard, Weibull, etc.)
- Use your covariates to model the hazard rate (that is, specify your systematic component)
- Derive the log-likelihood function for the model and get the values of your parameters that maximize your likelihood.
- Interpret your quantities of interest (hazard ratios, expected survival times, etc.)

What IS Different About Survival Models?

Censoring:

... it makes modeling a little tricky. But not too tricky



Observation 3 is censored because it had not experienced the event when we collected the data, so we don't know its true duration.

Censoring

- Observations that are censored still give us information about how long they survive.
- For censored observations, we know that they survived until at least some observed time, t^c. We also know that their true duration, t, must be $t \geq t^c$.
- So how do we approach data that has censored observations?
  - First, let's create an indicator variable, c_i, such that:

$$c_i = \begin{cases} 1 & \text{if not censored} \\ 0 & \text{if censored} \end{cases}$$

Censoring

- Now, we can use that indicator variable to incorporate information from censored observations into our likelihood function:

$$\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} [f(t_i)]^{c_i} \; [P(T_i \geq t_i^c)]^{1-c_i} \\
&= \prod_{i=1}^{n} [f(t_i)]^{c_i} \; [1 - F(t_i)]^{1-c_i} \\
&= \prod_{i=1}^{n} [f(t_i)]^{c_i} \; [S(t_i)]^{1-c_i}
\end{aligned}$$

- So, uncensored observations contribute to the density function and censored observations contribute to the survivor function in the likelihood.

Next Week

Next week we'll talk more about specific types of duration models and how to work with them.

Questions?

Section's over! You're free!