Count and Duration Models

Stephen Pettigrew

April 2, 2015

Outline

Outline

Logistics

- Final papers due Wednesday April 29 (last day of class before reading period)
- If you need an extension on the paper, you don't have to ask permission. We'll accept papers until Monday, May 4 at 5pm
- By tonight you should receive memos from the group re-replicating your paper. If you haven't sent your memo to the other group, do this as soon as possible
- Write a draft of the abstract for your final paper and post it to the discussion board for feedback (due April 15)
- Problem set 6 will be posted tonight. It's due a week from today.

Outline

Data for assessment question

The adjacency matrix:

$$\mathbf{Y} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

The party of the five senators:

$$\mathbf{x} = (0, 0, 1, 1, 1)$$

The upper triangle of the adjacency matrix, in vector form:

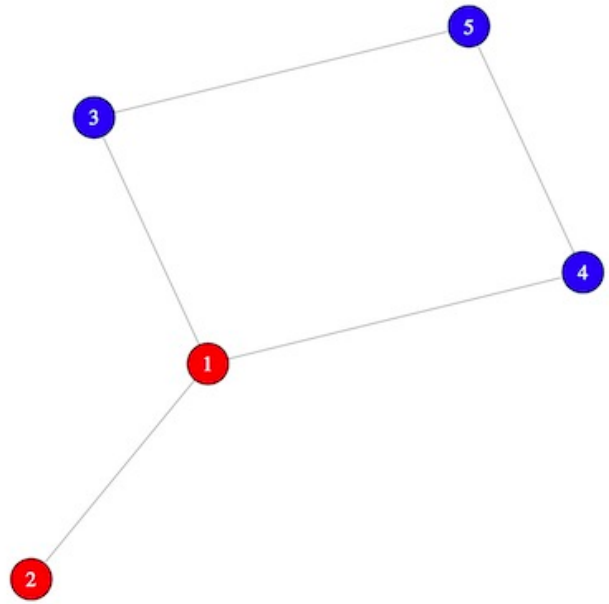
$$Y_{ij} \ \forall \ i < j = (1, 1, 1, 0, 0, 0, 0, 0, 1, 1)$$

Indicator variables for whether each of $\binom{5}{2}$ pairs of senators are in the same party

$$|X_i - X_j| \ \forall \ i < j = (0, 1, 1, 1, 1, 1, 1, 0, 0, 0)$$

Data for assessment question

Colors indicate party of each of the 5 senators.
Lines connecting a pair of senators indicates that they cosponsored



The model

We're told that the model for our data is:

$$Y_{ij} = Bern(\pi_{ij})$$

$$\pi_{ij} = \begin{cases} p_s & \text{if } X_i = X_j \\ p_d & \text{if } X_i \neq X_j \end{cases}$$

Question 1A

Rewrite the systematic component as a one-line equation, where π_{ij} is on the left-hand side and the right-hand side is expressed in terms of p_s, p_d, and $|X_i - X_j|$. (Hint: $|X_i - X_j|$ is 0 when $X_i = X_j$ and 1 when $X_i \neq X_j$). Note also that you should not reparametrize the parameters.

How can we rewrite the systematic component

$$\pi_{ij} = \begin{cases} p_s & \text{if } X_i = X_j \\ p_d & \text{if } X_i \neq X_j \end{cases}$$

so that it's a single equation?

Since $|X_i - X_j|$ can only be either 0 or 1, we can use it as an indicator variable to switch between p_s and p_d:

$$\pi_{ij} = p_s \cdot (1 - |X_i - X_j|) + p_d \cdot |X_i - X_j|$$

$$\pi_{ij} = p_s + (p_d - p_s) \cdot |X_i - X_j|$$

or

$$\pi_{ij} = p_d^{|X_i - X_j|} p_s^{1-|X_i - X_j|}$$

Question 1B

Write the log likelihood.

$$L(\pi_{ij}|\mathbf{y},\mathbf{x}) = \prod_{i<j}(\pi_{ij})^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}}$$

Now substitute in the systematic component you solved for in 1A:

$$\sum_{i<j}[y_{ij}\ln(p_s + (p_d - p_s)\cdot|X_i - X_j|) + (1-y_{ij})\ln(1 - p_s - (p_d - p_s)\cdot|X_i - X_j|)]$$

or

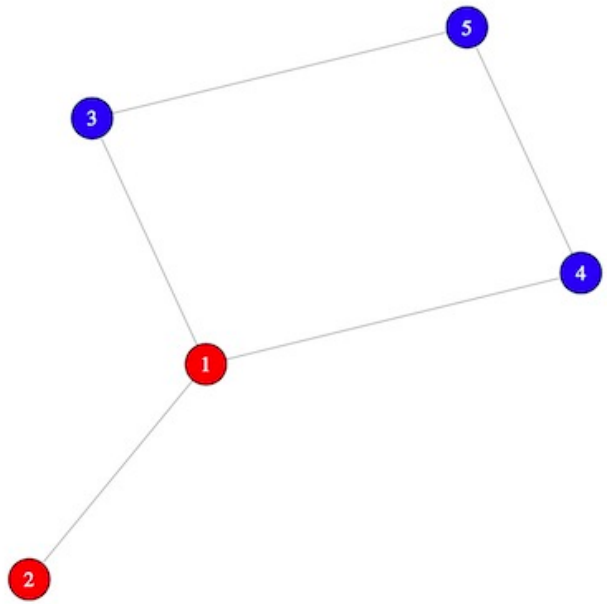
$$\sum_{i<j}\left[y_{ij}\ln(p_d^{|X_i-X_j|}p_s^{1-|X_i-X_j|}) + (1-y_{ij})\ln(1 - p_d^{|X_i-X_j|}p_s^{1-|X_i-X_j|})\right]$$

Question 1D

Optimize and find the MLE.

$\widehat{p_s} = 0.75$
$\widehat{p_d} = 0.33$

These results should be intuitive from the graph:



Outline

Counts

↑ Arithmomaniacs love count models

Any time your outcome variable is a count of the number of times as event happened, you'll want to use a count model.

If you know the number of trials and the probability of success is bigger than a really tiny decimal, you'll want to use a Bernoulli or binomial model

If the number of "trials" is too large (or impossible) to easily count and the rate of success is extremely low (or zero), you'll want to use another type of count model

Examples:

1. number of raindrops that hit a 1" x 1" square on the sidewalk during a rainstorm
2. number of publications by a professor in a career
3. number of times word "hope" is used in a Barack Obama speech
4. number of terrorist attacks in a given year
5. number of bear attacks in a year

The Poisson Distribution

The Poisson distribution is a discrete probability distribution which gives the probability that some number of events will occur in a fixed period of time.

Here's the probability density function (PDF) for a random variable Y that is distributed Pois(λ):

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

The Poisson Distribution

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

Using a little bit of geometric series magic, it isn't too hard to show that

$$\mathrm{E}[Y] = \sum_{y=0}^{\infty} y \cdot \frac{\lambda^y}{y!}\mathrm{e}^{-\lambda} = \lambda$$

It also turns out that $\mathrm{Var}(Y) = \lambda$, a feature of the model we will discuss later on.

The Poisson Distribution

Poisson data arises when there is some discrete event which occurs (possibly multiple times) at a constant rate for some fixed time period.

Another way to state this constant rate assumption is that the probability of an event occurring at any moment is independent of whether an event has occurred at any other moment.

So whether or not raindrop i hits our 1" x 1" square on the sidewalk doesn't affect the probably that raindrop $i + 1$ hits the within the square

Whether there's a civil war in country i at time t doesn't affect whether there's a civil war in country j at time $t + 1$

The Poisson Model for Event Counts

So what's the model?

1. The stochastic component:

$$Y_i \sim Pois(\lambda_i)$$

2. The systematic component:

$$\lambda_i = exp(X_i\beta)$$

In the Generalized Linear Models framework, $exp(X_i\beta)$ is our link function. Why do we use this?

Because λ_i is the rate parameter which must be greater than zero, and $exp(X_i\beta) > 0$

Deriving the log likelihood of the Poisson Model

If the PMF of the Poisson is:

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

then the likelihood is:

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

And the log-likelihood is:

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

Deriving the log likelihood of the Poisson Model

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

Example: Terrorist attacks

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

Explanatory variable: the unemployment rate in each month during the 30 year period.

Example: Terrorist attacks

The model: Let $Y_i = \#$ of terrorist attack deaths in a month. Our sole predictor for the moment will be: $U =$ the unemployment rate during that month.

Our model is:

$$Y_i \sim Pois(\lambda_i)$$

and

$$\lambda_i = E(Y_i|U_i) = exp(\beta_0 + \beta_1 \cdot U_i)$$

or more generally:

$$\lambda_i = E(Y_i|X_i) = exp(X_i\beta)$$

Let's estimate this thing!

We could estimate it by coding up the log likelihood function:

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

And then we could use `optim()`:

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

Let's estimate this thing!

Or we could use Zelig, since the Poisson model is already incorporated into that package:

```
require(Zelig)
zel.pois <- zelig(deaths ~ unemploy,
                  data = attacks,
                  model = "poisson")

summary(zel.pois)$coefficients
```

```
             Estimate Std. Error  z value      Pr(>|z|)
(Intercept)  1.230495 0.04203808 29.27095 2.430411e-188
unemploy    15.668797 0.17671220 88.66845  0.000000e+00
```

Checking our model

We now have estimates of our coefficients. Are we ready to move on to generating quantities of interest and pretty graphs?

Not yet

First we need to check our assumption that the data generation process was Poisson. How can we do this? What might we look for?

Dispersion

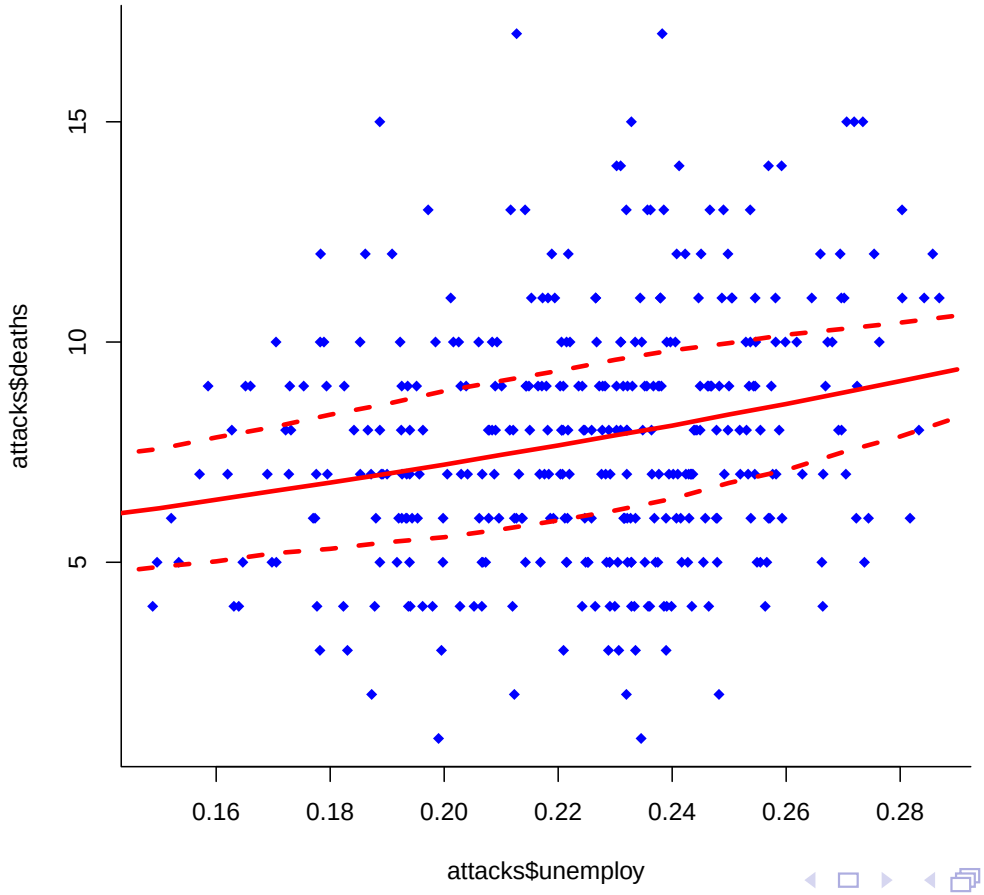
Recall from earlier that with the Poisson distribution, we assume that $E(Y|X) = Var(X|X) = \lambda$

If $E(Y|X) < Var(X|X)$ then our data are overdispersed.
If $E(Y|X) > Var(X|X)$ then our data are underdispersed.

How could we check for over or underdispersion?

Overdispersion

A lot more than 95% of our data is falling outside the 95% confidence intervals, so it looks like we have overdispersion



Overdispersion is a problem with the Poisson model in part because the model only has one parameter, λ, which restricts the variance to be a specific value

Recall from the Learning Catalytics question from Monday's lecture that this is a similar problem with exponential duration models, stylized normal models, and other models with only one parameter.

But it's not a problem with Bernoulli models.

Before we figure out why the heck the Bernoulli model is different from Poissons, exponentials, and others, let's take a step back.

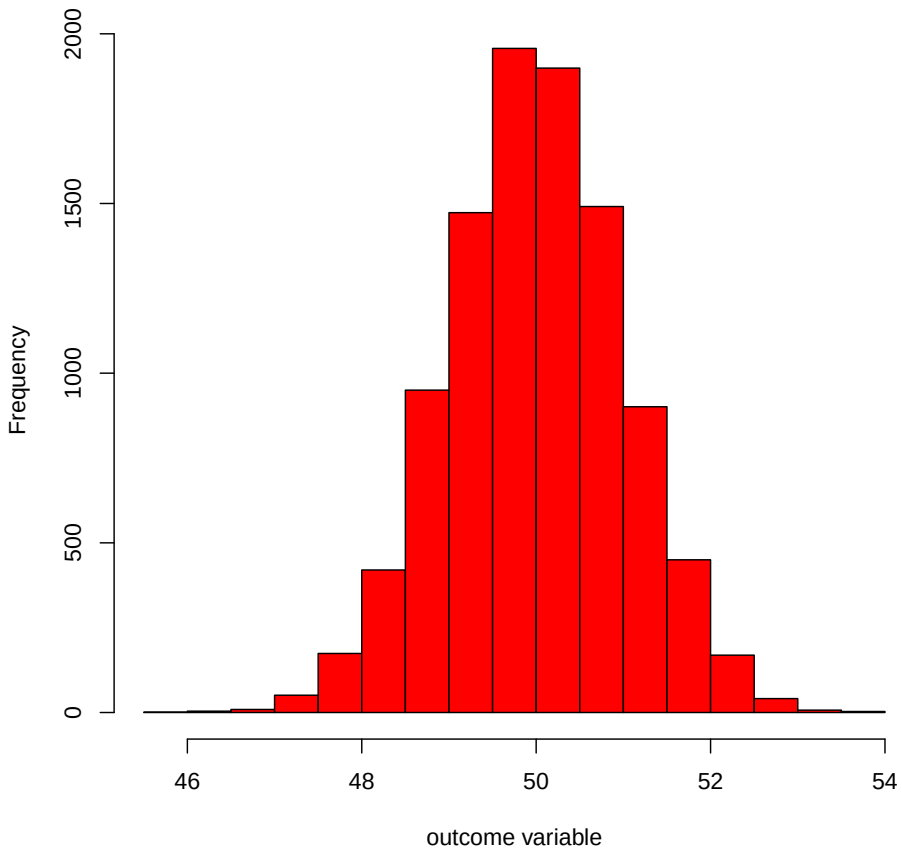
Big picture

Remember that when we're modeling, our big goal is to take data that's messy and complicated and summarize it using a distribution

Then we take that distribution and its parameters and make assessments about our original messy data.

When we do this, we're essentially collapsing our big messy data down to a few numbers (the parameters)

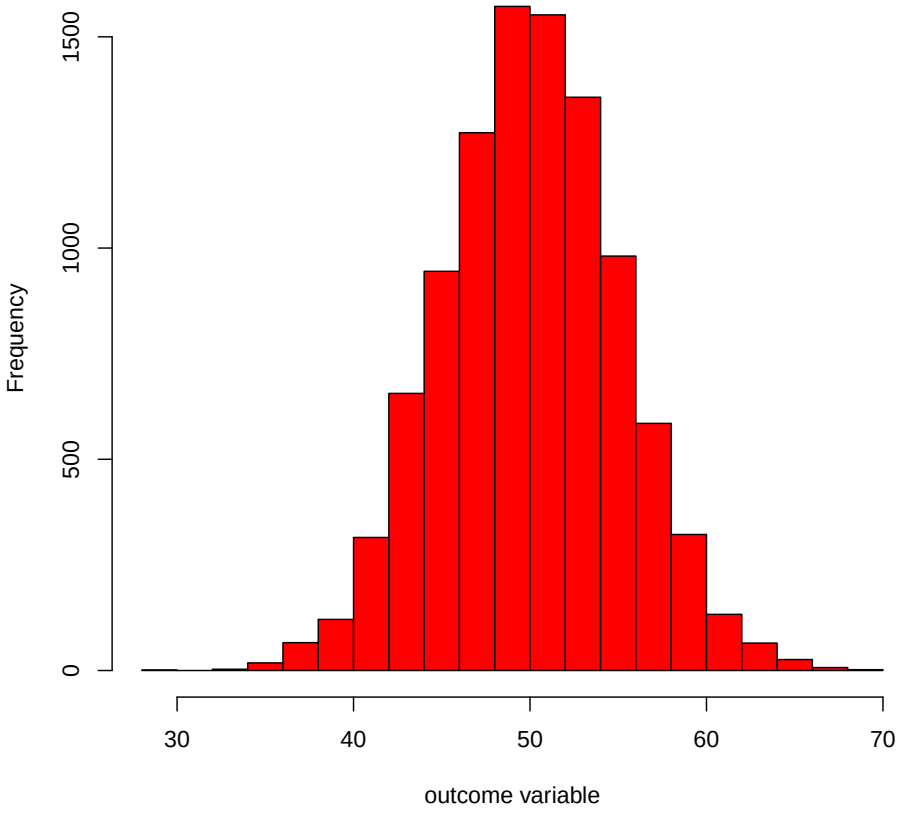
Let's look at an example of some data. Here's a histogram of an outcome variable you might be interested in modeling:



It appears that this data was generated from a stylized normal distribution, $N(\mu, 1)$.

If we used that as our model, we'd be summarizing our dataset using just one number, μ

What if our data looked like this? Would we want to use a stylized normal distribution?

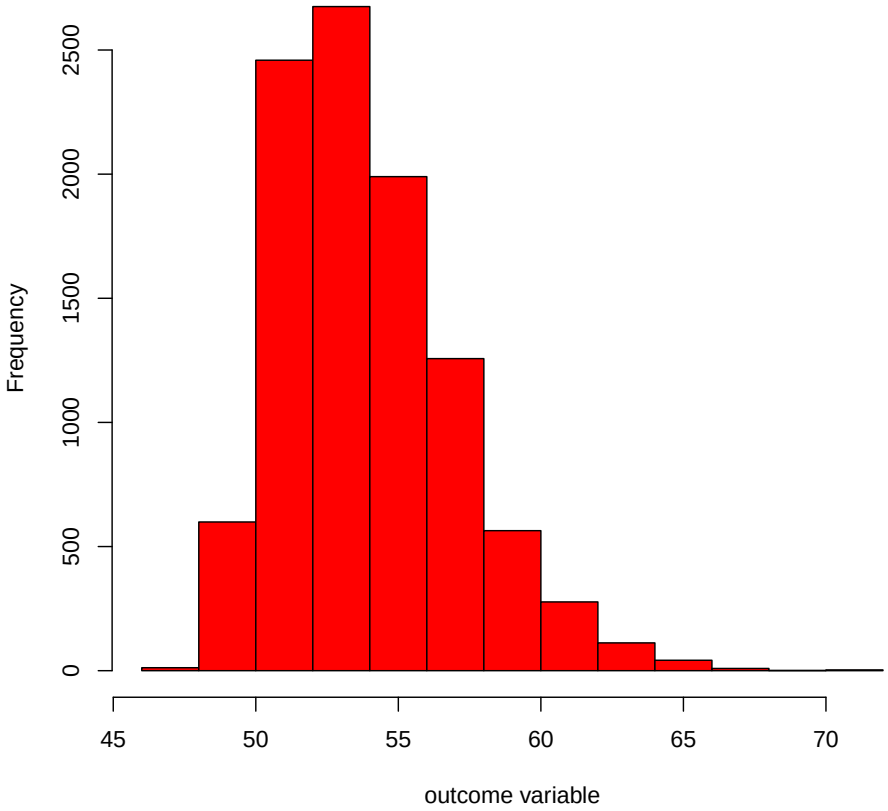


No we wouldn't, because if we did we'd be making the incorrect assumption that the variance was 1.

To properly describe these data we'd want to use two parameters, one for the mean and one for the variance.

Are we now fully characterizing the distribution of these data? Is there any information we could add to our model to better summarize the data?

What if our outcome looked like this? What assumption would we be violating if we used a $N(\mu, \sigma^2)$ model?



The data are skewed and not symmetric, as the normal distribution assumes.

But we could model this by using the skew-normal distribution, which is a normal distribution with a third parameter to model the skewness.

Why does this matter?

So then what's the point of all of this?

When you choose to model your data using just about any distribution, you're making some assumptions about the data.

Generally you could relax those assumptions by adding another parameter which helps you to better characterize the attributes of the data, like the mean, variance, skewness, kurtosis, etc.

Why does this matter?

Then why don't we do this? Or put another way, what's one way to perfectly capture all the information in our dataset?

Print off a spreadsheet with all your observations and variables and give it to your reader to interpret. Obviously that's not helpful

The more and more parameters we add to our model, the closer we get to this extreme and the more our estimates become tightly fit to our specific dataset.

There's a tradeoff between summarizing too little and summarizing too much. You have to strike a balance between underfitting your model and making incorrect assumptions about the data, and overfitting your model and not being able to draw conclusions about the world outside of your dataset

Why are binary outcome variables different?

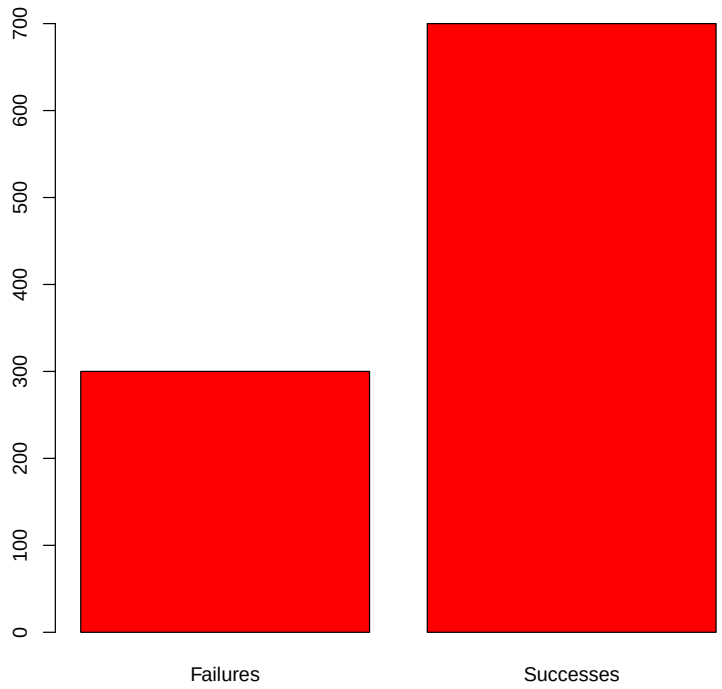
So adding parameters to models can help to relax assumptions and better fit the data.

Why don't we have to do this to get the variance correct in binary outcome models (i.e. models with Bernoulli stochastic components)?

Why is it that binary outcome models will have the correct variance with only one parameter, π, but exponential or Poisson models often won't?

Why are binary outcome variables different?

If the figure below shows binary outcome data for 1000 observations, how many numbers do you need to fully describe the data generation process for each of those 1000 observations?



You can fully characterize the underlying data generation process for this outcome by just knowing that $\pi = 0.7$.

With binary outcome models no information squeezed out of your data except this proportion, so adding another parameter for variance (or anything else) can't help to better describe the data generation process.

Outline

The Negative Binomial Model

Recall that with our Poisson model we found evidence of overdispersion, meaning that the assumption that $Var(Y) = E(Y) = \lambda$ was violated in the data.

Just like we saw with the jump from stylized normal to normal (or normal to skewed normal) we can add a parameter to relax this assumption.

In particular, we can derive a different distribution which relaxes the constant rate assumption and independence of events assumption of the Poisson

The trick is to assume that λ varies according to a new parameter we will introduce called zeta, ζ.

Deriving the Negative Binomial

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.
The variance of this new Poisson distribution is $\sigma^2 - 1$. This means that as σ^2 goes to 1, ζ_i becomes a spike over 1.

Deriving the Negative Binomial

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 PDF is:

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

Notes:

1. $\lambda_i > 0$ and $\sigma > 1$
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: $\lambda_i = exp(X_i\beta)$.

Negative binomial

In the problem set, you'll be asked to take this count model and use it to find the MLE for some data.

One thing that will help code this likelihood in R is knowing that $\Gamma(\cdot)$ is the gamma function, which is a generalization of factorials for non-integers. Note: The gamma function is a completely separate from the gamma distribution.

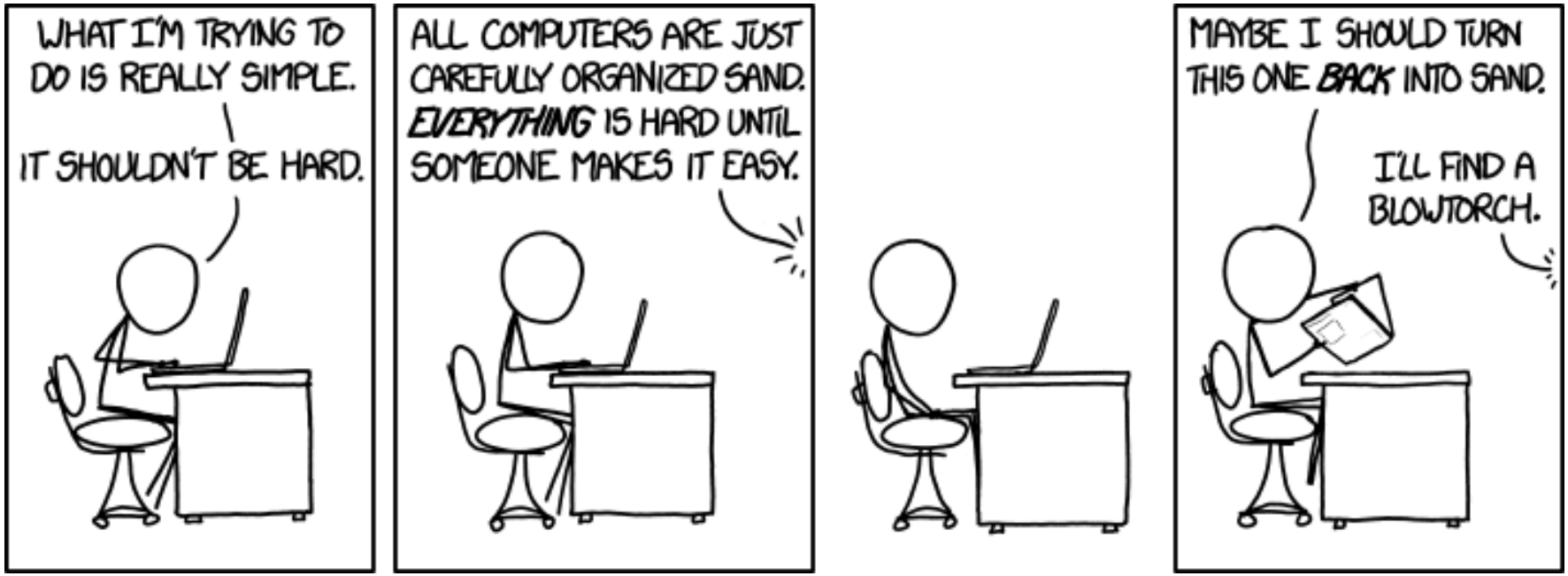
In R, the function is `gamma()`.

Other Models

Note that there are many other count models:

- Generalized Event Count (GEC) Model
- Zero-Inflated Poisson
- Zero-Inflated Negative Binomial
- Zero-Truncated Models
- Hurdle Models

Questions so far?



Outline

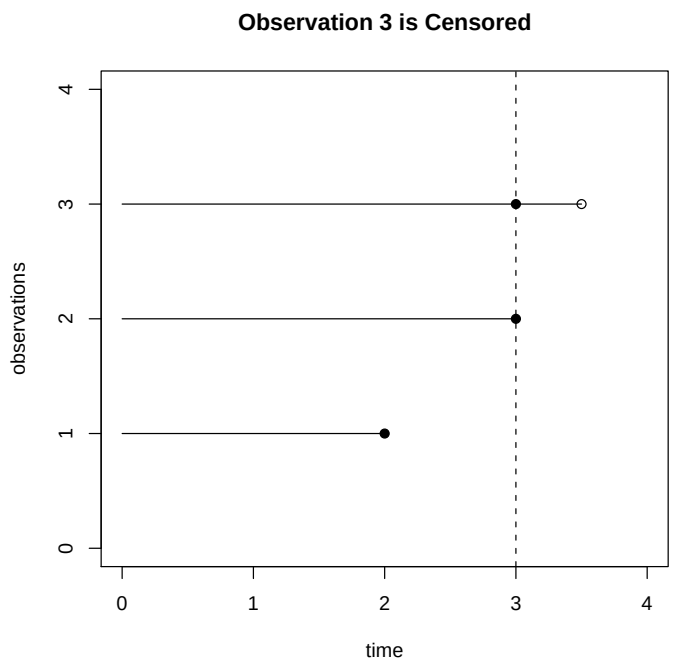
What are duration models used for?

- Survival models = duration models = event history models
- Dependent variable Y is the duration of time that observations spend in some state before experiencing an event (aka failure, death)
- Used in biostatistics and engineering: i.e. how long until a patient dies
- Models the relationship between duration and covariates (how does an increase in X affect the duration Y)
- In social science, used in questions such as how long a coalition government lasts, how long a war lasts, how long a regime stays in power, or how long until a legislator leaves office
- Observations should be measured in the same (temporal) units, i.e. don't have some units' duration measured in days and others in months

Why not just use OLS?

Three reasons:

1. OLS assumes Y is Normal but duration dependent variables are always positive (number of years, number of days. etc.)
2. Duration models can handle censoring



Observation 3 is censored in that it has not experienced the event at the time we collected the data, so we don't know its true duration

Why not use OLS?

3. Duration models can handle time-varying covariates
   - If Y is duration of a regime, GDP may change during the duration of the regime
   - OLS cannot handle multiple values of GDP per observation
   - You can set up data in a special way with duration models such that you can accomodate time-varying covariates. We won't cover this today but it's the same principle as censoring

Duration/Survival Model Jargon

Let T denote a continuous positive random variable representing the duration/survival times ($T = Y$)

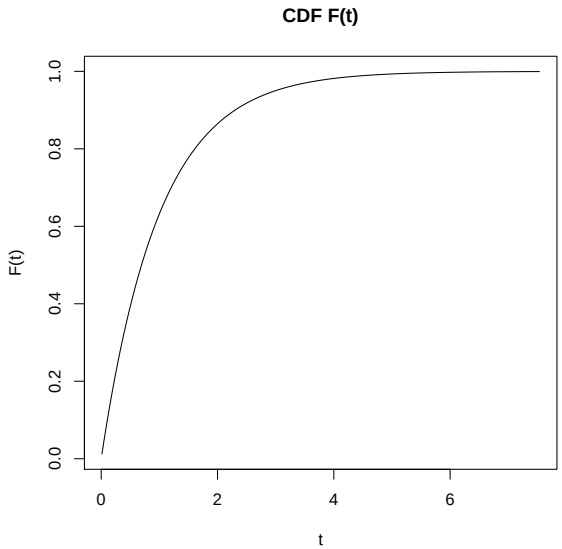
T has a probability density function $f(t)$

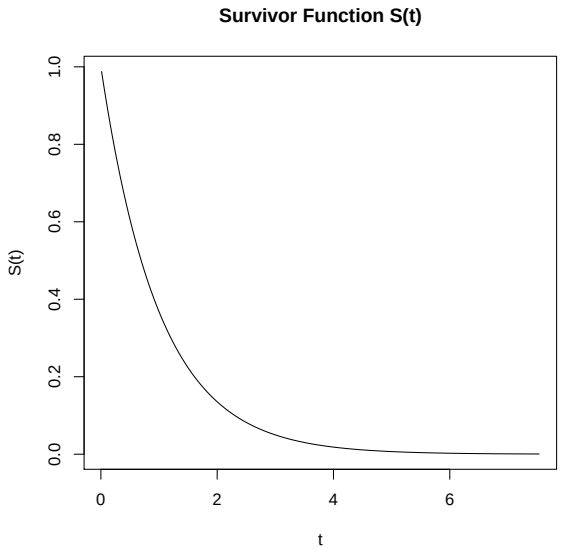
Duration/Survival Model Jargon

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: $S(t) = 1 - F(t) = P(T > t)$

Survivor: A 2001 Destiny's Child album that debuted at #1 on the Billboard chart and went quadruple platinum





Duration/Survival Model Jargon

Hazard rate (or hazard function): $h(t)$ is roughly the probability of an event at time t given survival up to time t

$$
\begin{aligned}
h(t) &= P(t \leq T < t + \tau | T \geq t) \\
&= P(\text{event at } t | \text{survival up to } t) \\
&= \frac{P(\text{survival up to } t | \text{event at } t) P(\text{event at } t)}{P(\text{survival up to } t)} \\
&= \frac{P(\text{event at } t)}{P(\text{survival up to } t)} \\
&= \frac{f(t)}{S(t)}
\end{aligned}
$$

Relating the Density, Survival, and Hazard Functions

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

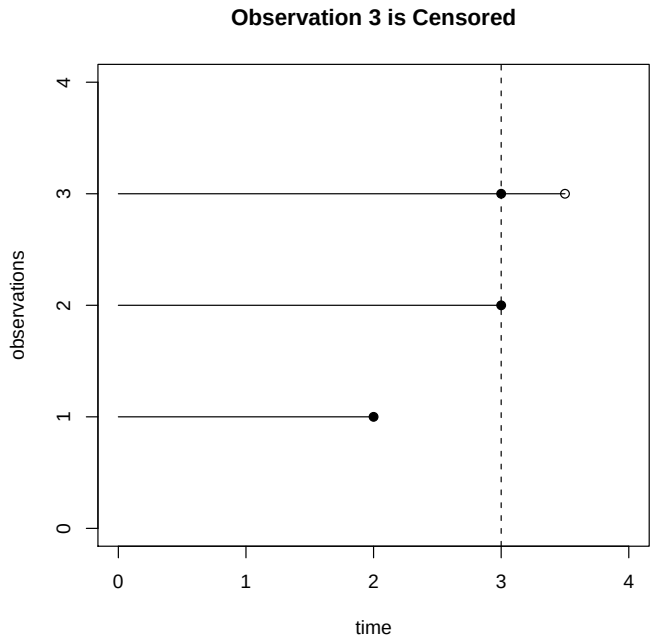
How to estimate parametric survival models

They might seem fancy and complicated, but we estimate these models the same as every other model!

1. Make an assumption that T_i follows a specific distribution $f(t)$ (i.e. choose the stochastic component).
2. Model the hazard rate with covariates (i.e. specify the systematic component).
3. Estimate via maximum likelihood.
4. Interpret quantities of interest (hazard ratios, expected survival times).

What's Special About Survival Models?

Censoring:



... it makes modeling a little tricky, but not too much

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 give us information about how long they survive.

For censored observations, we know that they survived at least until some observed time, t^c, and that the true duration, t is greater than or equal to t^c.

For each observation, let's create a censoring indicator variable, c_i, such that

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

Censoring

We can incorporate the information from the censored observations into the 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 go through a couple specific examples of duration models, as well as some stuff about assessing model fit

Questions?