Genera*lized* Linear Models for Other Not-Normal Outcomes

PSYC 943 (930): Fundamentals of Multivariate Modeling

Lecture 9: September 26, 2012

Today's Class

- A taxonomy of (conditionally) not-normal outcomes
  - "Discrete"
  - "Continuous"

- A (brief) tour of models for discrete outcomes
  - Poisson and its cousins
  - Real data example

- An (even briefer) tour of models for continuous outcomes
  - Y = Any positive value, but not normally distributed

3 Parts of a General*ized* Linear Model

- Link Function (main difference from GLM):
  - How a non-normal **outcome gets transformed** into something we can predict that is more continuous (unbounded)
  - **Why?** To keep outcome **predictions** within its sample space (slopes shut off)
  - So far we've seen *logit* links for binary outcomes, *cumulative logit* links for ordinal outcomes, and *generalized logit* links for nominal outcomes

- Model for the Means ("Structural Model"):
  - How predictors **linearly** relate to the link-transformed outcome
  - **New link-transformed** $\boldsymbol{Y_p = \beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p}$

- Model for the Variance ("Sampling/Stochastic Model"):
  - Many alternative distributions that map onto what the distribution of **residuals** could possibly look like (and kept within sample space)
  - **Why?** To get the most correct **standard errors** (which come from variance)
  - We've seen binomial and multinomial distributions for categorical outcomes so far, but there are many more available…

A Taxonomy of (Conditionally) Not-Normal Outcomes

- **“Discrete” outcomes**—all responses are **whole** numbers
  - **Categorical variables** in which **values are labels**, not amounts
    - Binomial (2 options) or multinomial (3+ options) distributions
    - Question: Are the values ordered → which link?
  - **Count of things that happened**, so values < 0 cannot exist
    - Sample space goes from 0 to positive infinity
    - Poisson or Negative Binomial distributions (usually)
    - Log link (usually) so predicted outcomes can’t go below 0
    - Question: Are there *extra* 0 values? What to do about them?

- **“Continuous” outcomes**—responses can be **any** number
  - Question: What does the residual distribution look like?
    - Normal-ish? Skewed? Cut off? Mixture of different distributions?

A BRIEF TOUR OF MODELS FOR DISCRETE OUTCOMES

Models for Count Outcomes

- **Counts: non-negative integer unbounded responses**
  - e.g., how many cigarettes did you smoke this week?
  - Traditionally uses natural log link so that predicted outcomes stay ≥ 0

- $g(\bullet)$ $\quad Log\left(E\left(Y_p\right)\right) = Log\left(\mu_p\right) = model$ → predicts mean of Y_p
- $g^{-1}(\bullet)$ $\quad E\left(Y_p\right) = exp(model)$ → to un-log it, use $exp(model)$

- e.g., if $Log\left(\mu_p\right) = model$ provides predicted $\mu_p = 1.098$, that translates to an actual predicted count of $\exp(1.098) = 3$
- e.g., if $Log\left(\mu_p\right) = model$ provides predicted $\mu_p = -5$, that translates to an actual predicted count of $\exp(-5) = 0.006738$

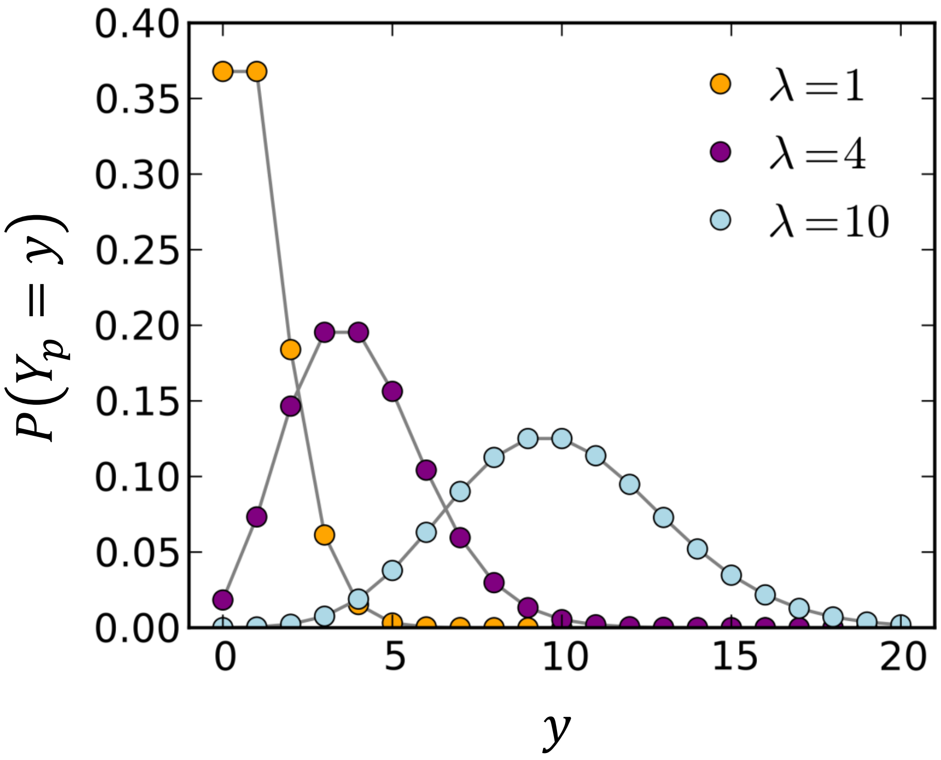
- So that’s how the model for the means predicts μ_p, the expected count for Y_p, but what about the model for the variance?

Poisson Distribution for Residuals in Count Outcomes

- Poisson distribution has one parameter, λ, which is both its mean and its variance (so λ = mean = variance in Poisson distribution)

- $f(Y_p|\lambda) = \text{Prob}(Y_p = y) = \frac{\lambda^y * exp(-\lambda)}{y!}$

$y!$ is factorial of y



The dots indicate that only integer values are observed.

Distributions with a small expected value (mean or λ) are predicted to have a lot of 0's.

Once $\lambda > 6$ or so, the shape of the distribution is close to a that of a normal distribution.

Poisson Distribution for Residuals in Count Outcomes

- Just as for other discrete outcomes, maximum likelihood is used to estimate model parameters to predict the expected Y_p using the response distribution we picked—here, Poisson (not multinomial)

- Model: $Log\left(E(Y_p)\right) = Log\left(\mu_p\right) = \beta_0 + \beta_1 X_p + \beta_2 Z_p$

- PDF: $\text{Prob}\left(Y_p = y|\beta_0, \beta_1, \beta_2\right) = \frac{\mu_p^y * exp(-\mu_p)}{y!}$

Note what's *not* in the model…

- Creates this log-likelihood function for each person's count Y_p:

$$Log\ L(\beta_0, \beta_1, \beta_2 | Y_1, \dots, Y_N)$$

$$= Log[L(\beta_0, \beta_1, \beta_2 | Y_1) \times L(\beta_0, \beta_1, \beta_2 | Y_2) \times \cdots \times L(\beta_0, \beta_1, \beta_2 | Y_N)]$$

$$= \Sigma_{p=1}^{N}\left[Y_p * Log\left(\mu_p\right) - \mu_p - Log\left(Y_p!\right)\right]$$

- in which μ_p is the predicted (mean) count per person

What could go wrong? 3 potential problems for Poisson…

- The standard Poisson distribution is rarely sufficient, though

- **Problem #1: When mean ≠ variance**
  - If variance < mean, this leads to "under-dispersion" (not that likely)
  - If variance > mean, this leads to "over-dispersion" (happens frequently)

- **Problem #2: When there are *no* 0 values**
  - Some 0 values are expected from count models, but in some contexts $Y_p > 0$ always (but subtracting 1 won't fix it; need to adjust the model)

- **Problem #3: When there are *too many* 0 values**
  - Some 0 values are expected from the Poisson and Negative Binomial models already, but many times there are even more 0 values observed than that
  - To fix it, there are two main options, depending on what you do to the 0's

- Each of these problems requires a model adjustment to fix it…

Problem #1: Variance > mean = over-dispersion

- To fix it, we must add another parameter that allows the variance to exceed the mean… becomes a **Negative Binomial distribution**
  - Says residuals are a mixture of Poisson and gamma distributions
- Model: $Log(Y_p) = Log(\mu_p) = \beta_0 + \beta_1 X_p + \beta_2 Z_p + e_p^G$
- Poisson PDF was: $\text{Prob}(Y_p = y|\beta_0, \beta_1, \beta_2) = \frac{\mu_p^y * exp(-\mu_p)}{y!}$
- Negative Binomial PDF with a new ***k*** **dispersion** parameter is now:
  - $\text{Prob}(Y_p = y|\beta_0, \beta_1, \beta_2) = \frac{\Gamma\left(y+\frac{1}{\boldsymbol{k}}\right)}{\Gamma(y+1)*\Gamma\left(\frac{1}{\boldsymbol{k}}\right)} * \frac{(\boldsymbol{k}\mu_p)^y}{(1+\boldsymbol{k}\mu_p)^{y+\frac{1}{\boldsymbol{k}}}}$

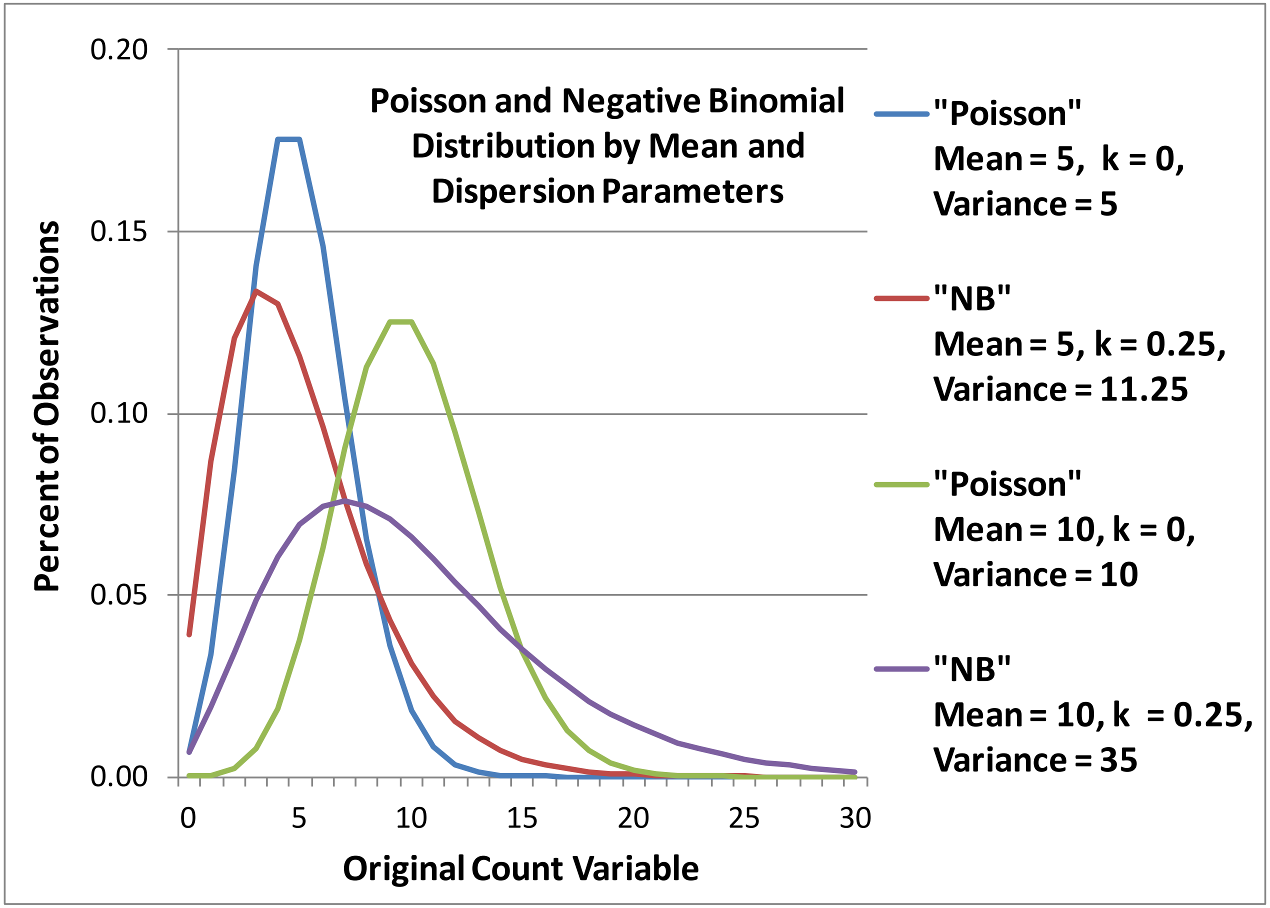
DIST = NEGBIN
in SAS GENMOD

  - ***k*** **is dispersion**, such that $Var(Y_p) = \mu_p + k\mu_p^2$

So is Poisson if $k = 0$

  - Non-Poisson related $e_p^G \sim Gamma(mean = 1, variance = k)$
    - Since Log(1) = 0, the extra 0's won't add to the predicted log count, and if there is no extra dispersion, then variance of $e_p^G \sim 0$

Negative Binomial (NB) = "Stretchy" Poisson…



$\text{Mean} = \lambda$
$\text{Dispersion} = k$

$$Var(Y_p) = \lambda + k\lambda^2$$

A Negative Binomial model can be useful for count outcome residuals that have some extra skewness, but otherwise follow a Poisson distribution.

- Because its k dispersion parameter is fixed to 0, the Poisson model is nested within the Negative Binomial model—to test improvement in fit:
- Is $-2(LL_{Poisson} - LL_{NegBin}) > 3.84$ for $df = 1$? Then $p < .05$, keep NB

Problem #2: There are no 0 values

- "**Zero-Altered**" or **"Zero-Truncated"** Poisson or Negative Binomial
  - ZAP/ZANB or ZTP/ZTNB
  - Is usual count distribution, just not *allowing* any 0 values
  - Poisson version is readily available within SAS PROC FMM using DIST=TRUNCPOISSON (next version should have TRUNCNEGBIN, too)

- Poisson PDF was: $\text{Prob}\left(Y_p = y|\mu_p\right) = \frac{\mu_p^y * exp(-\mu_p)}{y!}$

- Zero-Truncated Poisson PDF is:
  - $\text{Prob}\left(Y_p = y|\mu_{p,} Y_p > 0\right) = \frac{\mu_p^y * exp(-\mu_p)}{y![1 - exp(-\mu_p)]}$
  - $Prob\left(Y_p = 0\right) = exp\left(-\mu_p\right)$, so $Prob\left(Y_p > 0\right) = 1 - exp\left(-\mu_p\right)$
  - Divides by probability of non-0 outcomes so total probability still sums to 1

Problem #3: There are too many 0 values… Option #1

- "**Zero-Inflated**" Poisson (DIST=ZIP) or Negative Binomial (DIST=ZINB)
  - Readily available within SAS PROC GENMOD (and Mplus)
  - Distinguishes **two kinds of 0 values**: **expected** and **inflated** ("structural") through a mixture of distributions (Bernoulli + Poisson/NB)
  - Creates two submodels to predict "if *extra* 0" and "if not, how much"?
    - Does not readily map onto most hypotheses (in my opinion)
    - But a ZIP example would look like this…

- Submodel 1: $Logit\left(Y_p = extra\ 0\right) = \ \beta_{01} + \beta_{11}X_p + \beta_{21}Z_p$
  - Predict being an extra 0 using Link = Logit, Distribution = Bernoulli
  - Don't have to specify predictors for this part, can simply allow an intercept (but need ZEROMODEL option to include predictors in SAS GENMOD)

- Submodel 2: $Log\left(E\left(Y_p\right)\right) = \beta_{02} + \beta_{12}X_p + \beta_{22}Z_p$
  - Predict rest of counts (including 0's) using Link = Log, Distribution = Poisson

- Same idea for ZINB, just adds the k dispersion parameter, too

Example of Zero-Inflated Outcomes

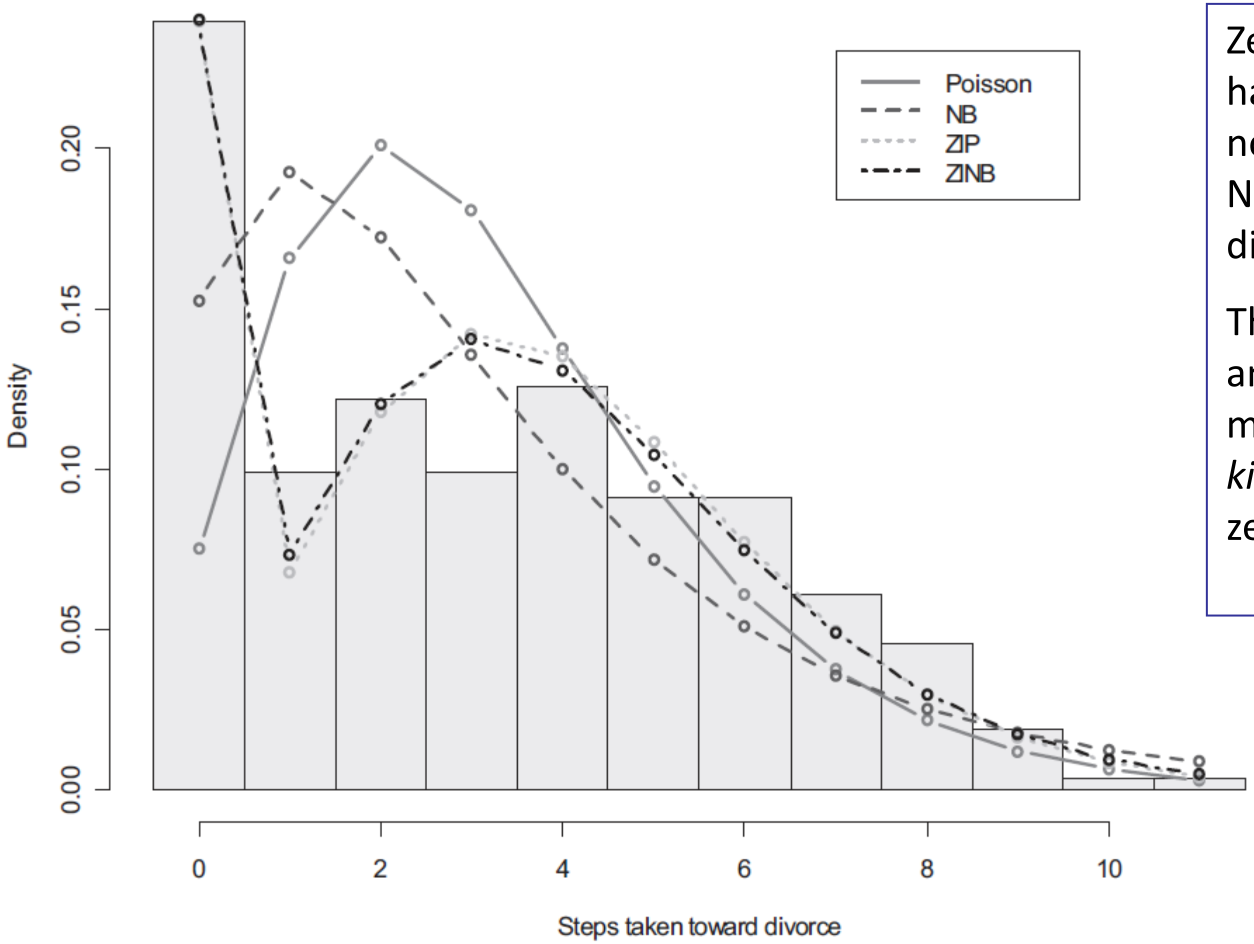


Figure 1. Histogram of Marital Status Inventory with predicted probabilities from regressions. NB = negative binomial; ZIP = zero-inflated Poisson; ZINB = zero-inflated negative binomial.

Zero-inflated distributions have extra “structural zeros” not expected from Poisson or NB (“stretched Poisson”) distributions.

This can be tricky to estimate and interpret because the model distinguishes between *kinds of zeros* rather than zero or not...

Image borrowed from Atkins & Gallop, 2007

Problem #3: There are too many 0 values… Option #1

- The Zero-Inflated models get put back together again as follows:
  - ω_p is the predicted probability of being an extra 0, from:

$$\omega_p = \frac{exp\left[Logit\left(Y_p = extra\ 0\right)\right]}{1 + exp\left[Logit\left(Y_p = extra\ 0\right)\right]}$$

  - μ_p is the predicted count for the rest of the distribution, from:

$$\mu_p = exp\left[Logit\left(Y_p\right)\right]$$

  - ZIP: $Mean\ \left(original\ Y_p\right) = \left(1 - \omega_p\right)\mu_p$
  - ZIP: $Variance\left(original\ Y_p\right) = \mu_p + \frac{\omega_p}{\left(1-\omega_p\right)}\mu_p^2$

  - ZINB: $Mean\ \left(original\ Y_p\right) = \left(1 - \omega_p\right)\mu_p$
  - ZINB: $Variance\left(original\ Y_p\right) = \mu_p + \left[\frac{\omega_p}{\left(1-\omega_p\right)} + \frac{k}{1-\omega_p}\right]\mu_p^2$

Problem #3: There are too many 0 values… Option #2

- "**Hurdle**" models for Poisson or Negative Binomial
  - PH or NBH: Explicitly **separates 0 from non-0 values** through a mixture of distributions (Bernoulli + Zero-Altered Poisson/NB)
  - Creates two submodels to predict "if *any* 0" and "if not 0, how much"?
    - Easier to think about in terms of prediction (in my opinion)

- Submodel 1: $Logit\left(Y_p = 0\right) = \ \beta_{01} + \beta_{11}X_p + \beta_{21}Z_p$
  - Predict being a 0 using Link = Logit, Distribution = Bernoulli
  - Don't have to specify predictors for this part, can simply allow it to exist

- Submodel 2: $Log\left(E(Y_p) > 0\right) = \beta_{02} + \beta_{11}X_p + \beta_{21}Z_p$
  - Predict rest of positive counts using Link = Log, Distribution = ZAP or ZANB

- These models are not *readily* available in SAS, but NBH is in Mplus
  - Could be fit as a multivariate model in SAS GLIMMIX (I think)

Comparing Models for Count Data

- Whether or not a dispersion parameter is needed can be answered via a likelihood ratio test
  - For the most fair comparison, keep the model for the means the same
- Whether or not a zero-inflation model is needed should, in theory, also be answerable via a likelihood ratio test…
  - But people disagree about this
  - Problem? Zero-inflation probability can't be negative, so is bounded at 0
  - Other tests have been proposed (e.g., Vuong test—see SAS macro online)
  - Can always check AIC and BIC (smaller is better)
- In general, models with the same distribution and different links can be compared via AIC and BIC, but one cannot use AIC and BIC to compare across alternative distributions (e.g., normal or not?)
  - Log-Likelihoods are not on the same scale due to using different PDFs

And that's not all!

- There are many, many other possibilities for count outcomes, including different links and different distributions…
- In addition, count outcomes can also use models for continuous outcomes, up next after this example…

AN EVEN BRIEFER TOUR OF MODELS FOR CONTINUOUS OUTCOMES

Models for Continuous Outcomes > 0

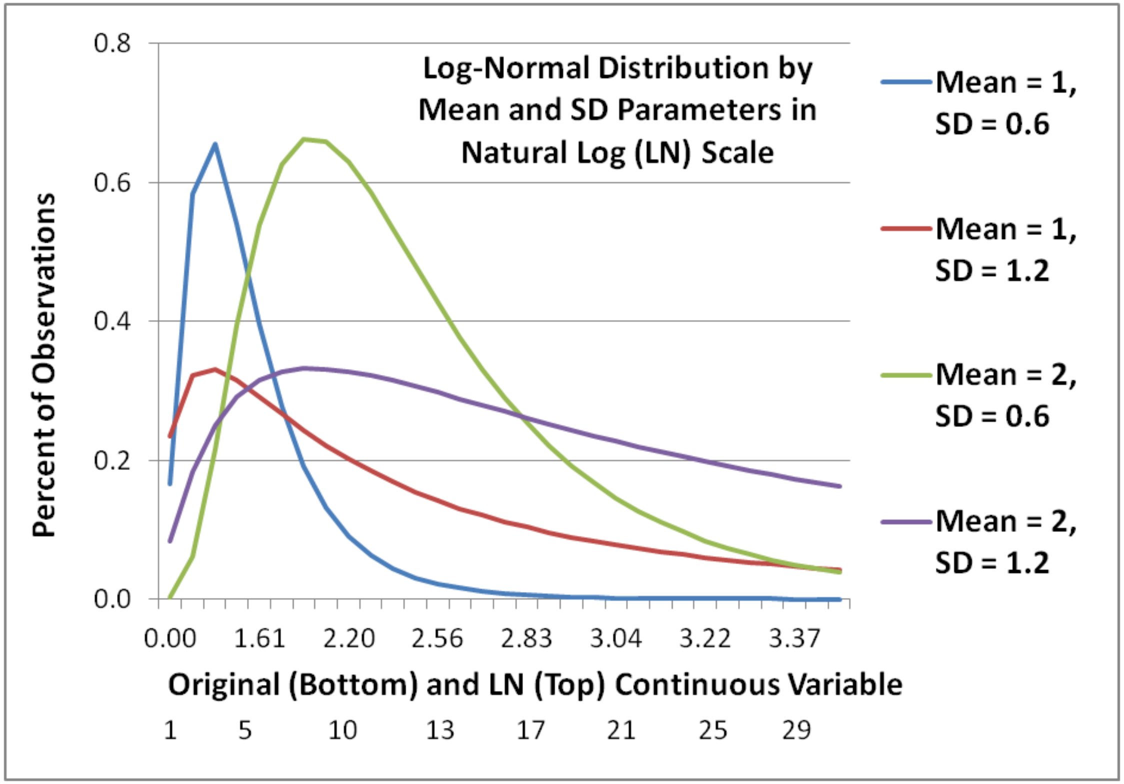
- There are many choices for modeling not-normal *continuous* outcomes (that include positive values only)
  - Most rely on either an identity or log link
  - Will find them in SAS PROC GENMOD and GLIMMIX (see also QLIM)

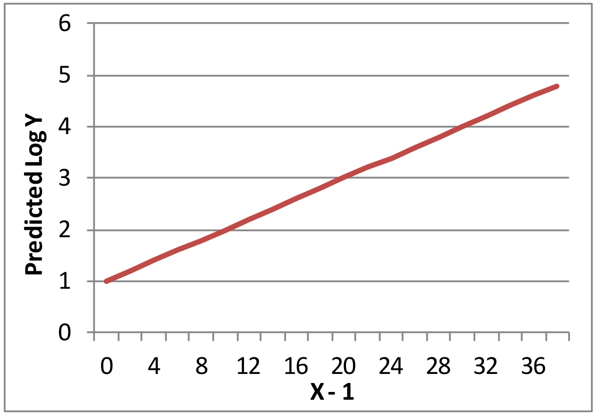
- GENMOD: DIST= (default link)
  - Gamma (Inverse), Geometric (Log), Inverse Gaussian (Inverse^2), Normal (Identity)

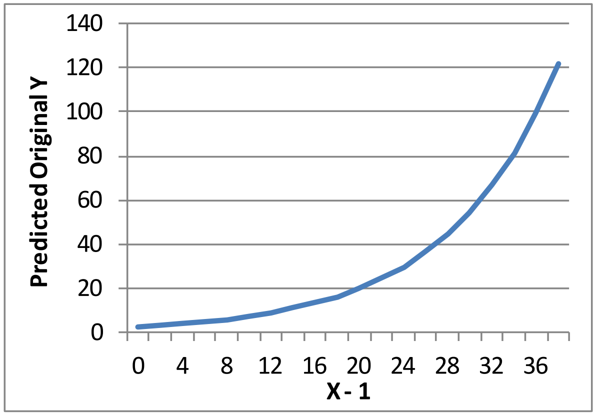
- GLIMMIX: DIST= (default link)
  - Beta (Logit), Exponential (Log), Gamma (Log), Geometric (Log), Inverse Gaussian (Inverse^2), Normal (Identity), LogNormal (Identity), TCentral (Identity), and BYOBS, which allows multivariate models by which you specify DV-specific models estimated simultaneously (e.g., two-part)

- Many others possible as well—here are just some examples…

Log-Normal Response Distribution (Link=Identity)

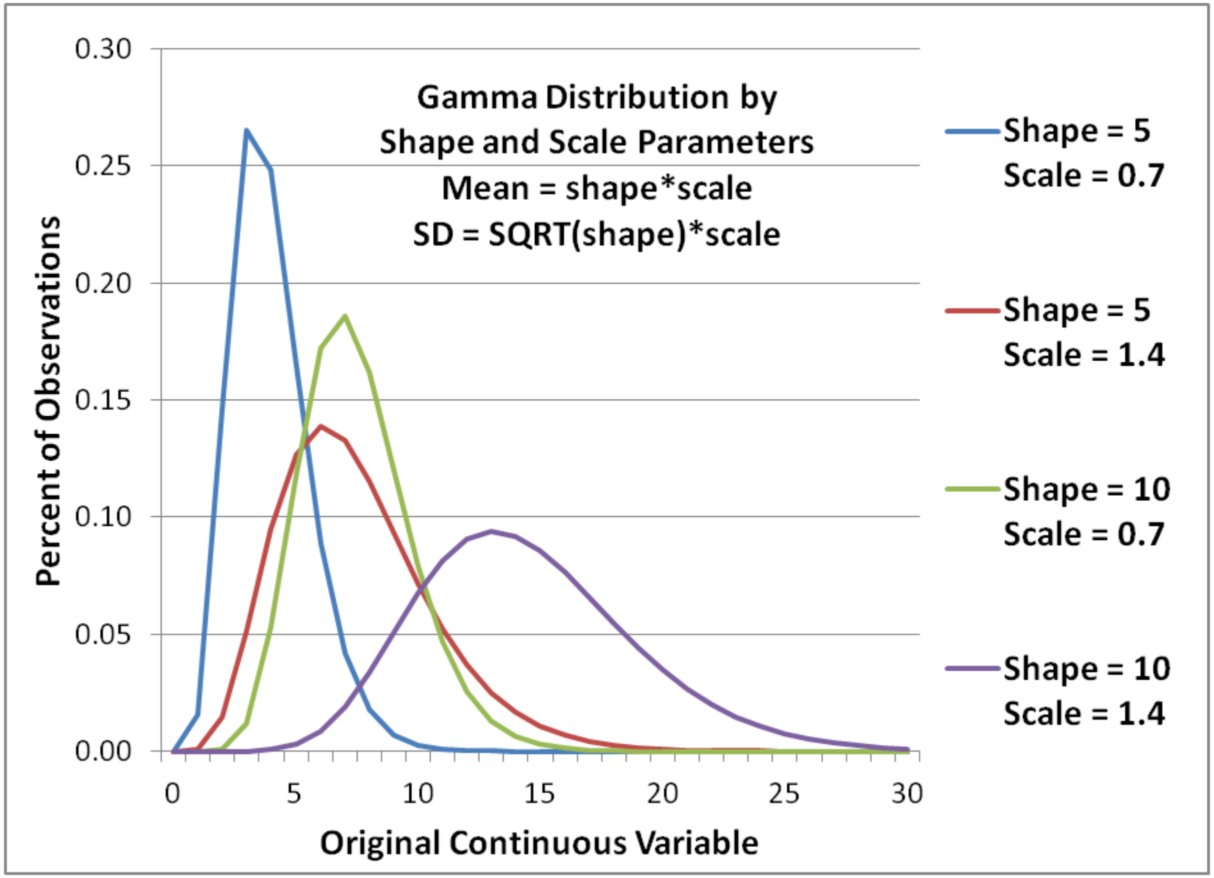


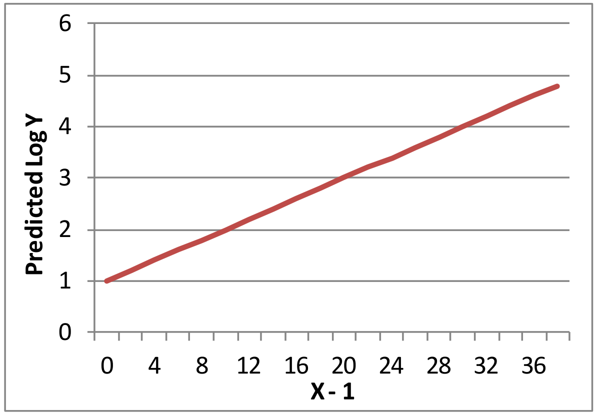


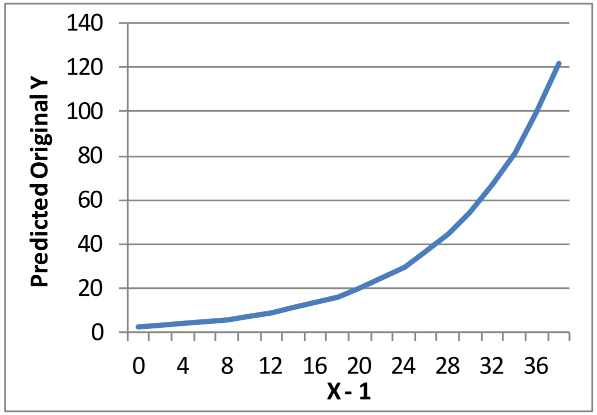


- Model: $Log\left(Y_p\right) = \beta_0 + \beta_1 X_p + \beta_2 Z_p + e_p$
  where $e_p \sim LogNormal(0, \sigma_e^2)$ → *log* of residuals is normal, not residuals
  - Happens to be the same as log-transforming your outcome in this case…
  - The LOG function keeps the predicted values positive, but results in an exponential, not linear prediction of original outcome from slopes
  - GLIMMIX provides “intercept” and “scale=SD” that need to be converted…

Gamma Response Distribution





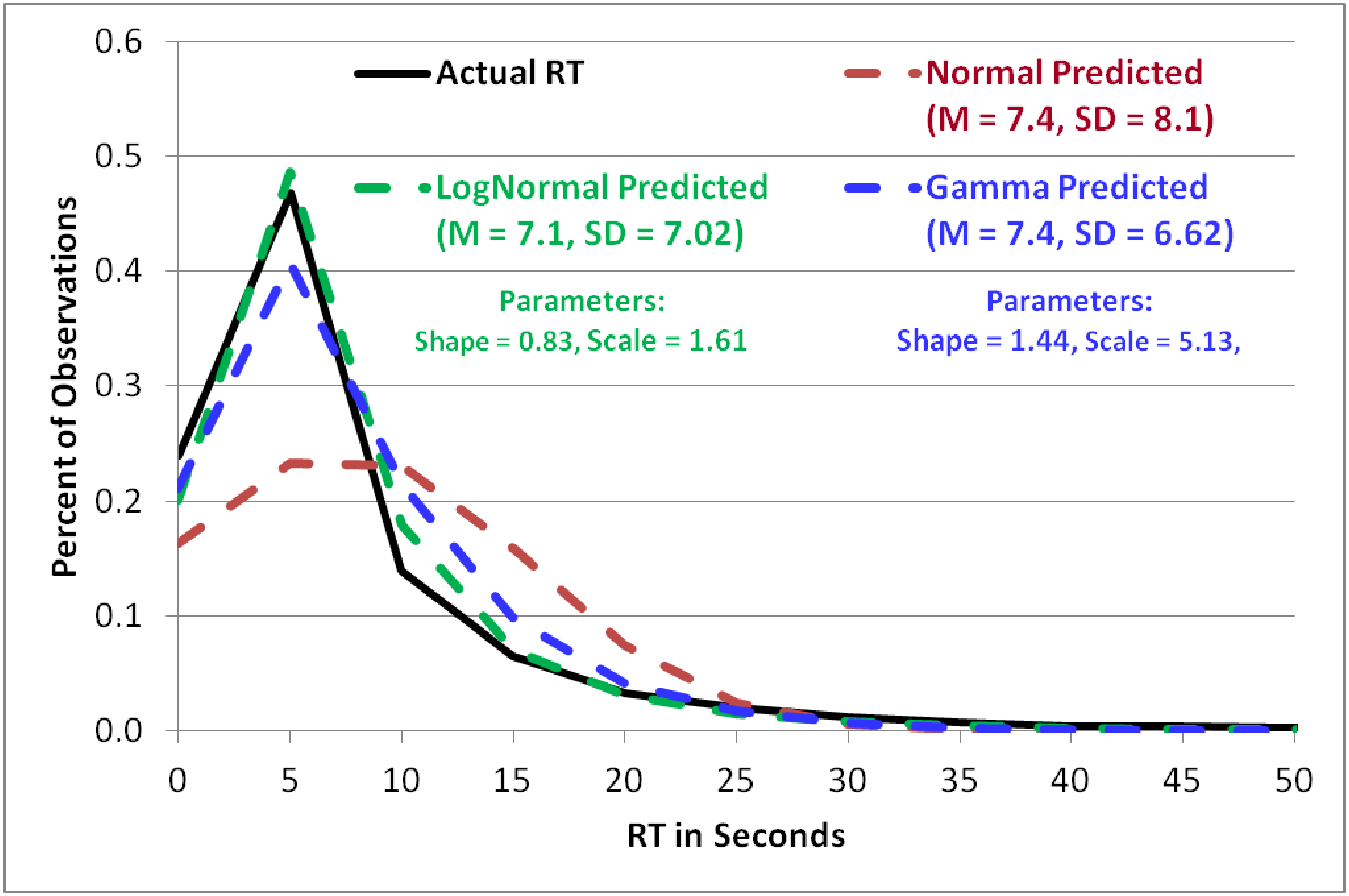


- Model: $Log\left(Y_p\right) = \beta_0 + \beta_1 X_p + \beta_2 Z_p + e_p$

  where $e_p \sim Gamma(0, \sigma_e^2)$ → *variance is based on shape and scale parameters*

  - Default Link is log in GLIMMIX, but inverse in GENMOD
  - Provides “intercept” and “scale=1/scale” that need to be converted…

Real Example using Response Time in Seconds



- Pretty much anything but normal (in red) looks ok!

Two-Part Models for Continuous Outcomes

- A two-part model is an analog to hurdle models for zero-inflated count outcomes (and could be used with count outcomes, too)
  - Explicitly **separates 0 from non-0 values** through a mixture of distributions (Bernoulli + Normal or LogNormal)
  - Creates two submodels to predict "if not 0" and "if not 0, how much"?
    - Easier to think about in terms of prediction (in my opinion)

- Submodel 1: $Logit\left(Y_p > 0\right) = \ \beta_{01} + \beta_{11}X_p + \beta_{21}Z_p$
  - Predict being a 0 using Link = Logit, Distribution = Bernoulli
  - Usually do specify predictors for this part

- Submodel 2: $\left(Y_p | Y_p > 0\right) = \beta_{02} + \beta_{11}X_p + \beta_{21}Z_p$
  - Predict rest of positive amount using Link = Identity, Distribution = Normal or Log-Normal (often rest of distribution is skewed, so log works better)

- Two-part is not *readily* available in SAS, but is in Mplus
  - Could be fit as a multivariate model in SAS GLIMMIX (I think)
  - Is related to "tobit" models for censored outcomes (for floor/ceiling effects)

Wrapping Up…

- Today we examined some generalized models for non-categorical but non-normal outcomes
  - Count data: log link, some kind of Poisson-based discrete distribution
  - Continuous data: log or identity link, some kind of not-normal distribution
  - There are many, many more to choose from

- Different programs/books will parameterize these models differently, so you'll need to read the documentation carefully

- The point? Never be stuck with "normal" again!