Generalized Linear Models

Goals

- To introduce the format and structure of generalized linear models.
- To show how the linear and logistic models fit into the generalized linear models framework.
- To introduce Poisson generalized linear models for count data.
- To describe diagnostics for generalized linear models.

Introduction

Generalized linear models (GLMs) extend the range of application of linear statistical models by accommodating response variables with non-normal conditional distributions.

Except for the error, the right-hand side of a generalized linear model is essentially the same as for a linear model.

$$g(E(Y_i|\mathbf{X_i})) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}$$

where $E(Y_i|\mathbf{X_i})$ is the expected value of the response variable Y_i given the explanatory variables. It is generally denoted as $\mu_i = E(Y_i|\mathbf{X_i})$. $g(.)$ is called the link function.

A GLM consists of three components:

1. A random component, specifying the conditional distribution of the response variable, Y_i, given the explanatory variables.

$$g(E(Y_i|\mathbf{X_i})) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}$$

- Traditionally, the random component is a member of an "exponential family" — the normal (Gaussian), binomial, Poisson, gamma, or inverse-Gaussian families of distributions — but GLMs have been extended beyond the exponential families.
- The Gaussian and binomial distributions are already familiar.
- Poisson distributions are often used in modeling count data. Poisson random variables take on non-negative integer values, $0, 1, 2, \ldots$.
- The gamma and inverse-Gaussian distributions are for positive continuous data

2. A linear function of the regressors, called the *linear predictor*,

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}$$

on which the expected value, μ_i, of Y_i depends.

The X's may include quantitative predictors, but they may also include transformations of predictors, polynomial terms, dummy variables, interaction regressors, etc.

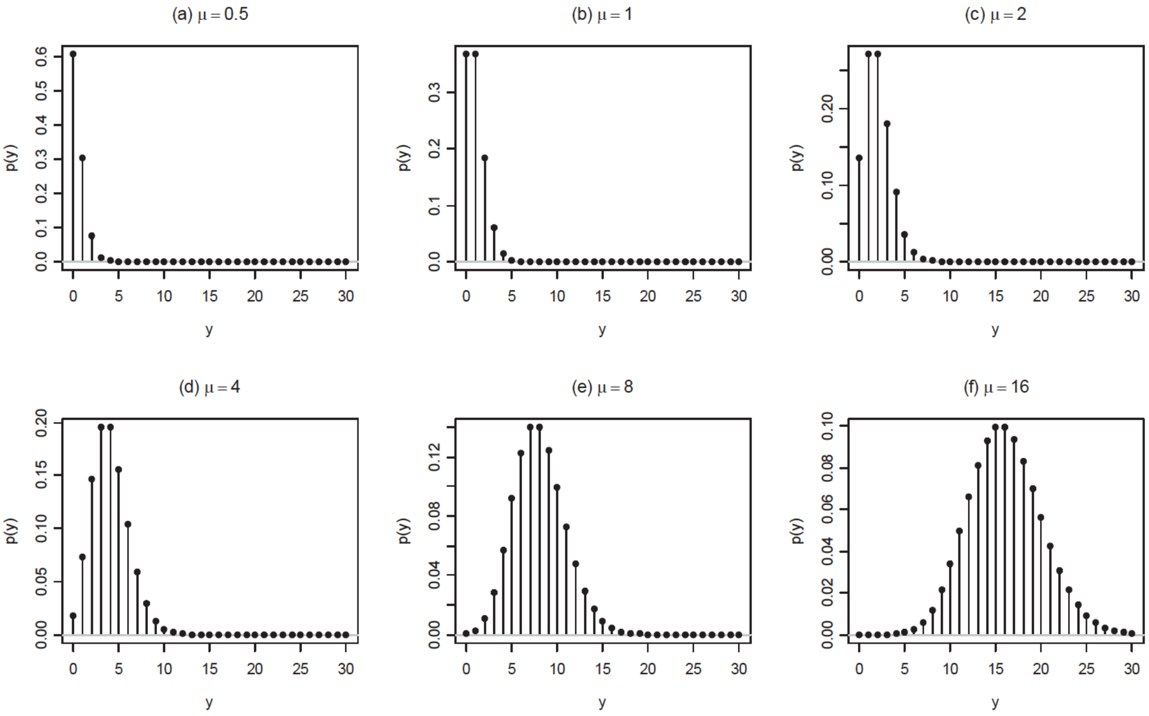


Figure 1: Poisson Distribution

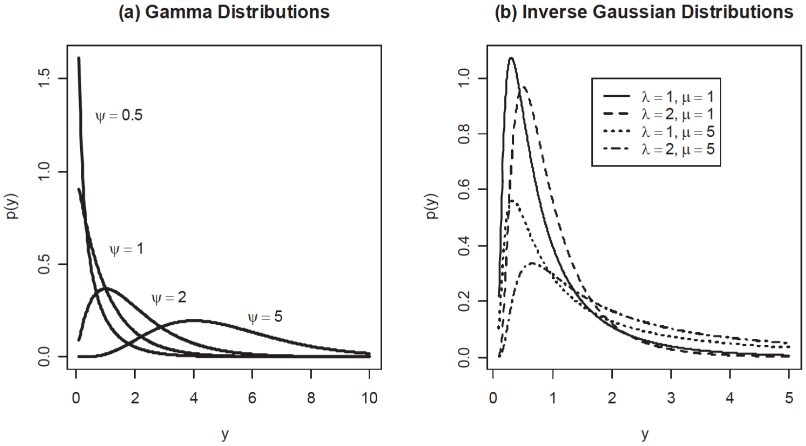
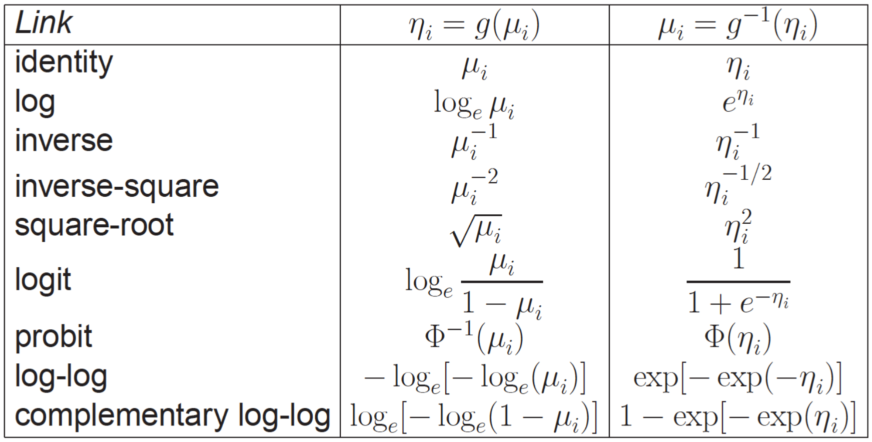


Figure 2: Gamma and Inverse-Gaussian Distributions

3. An invertible *link function*, $g(\mu_i) = \eta_i$, which transforms the expectation of the response to the linear predictor.

The inverse of the link function is sometimes called the mean function:

$$g^{-1}(\eta_i) = \mu_i$$



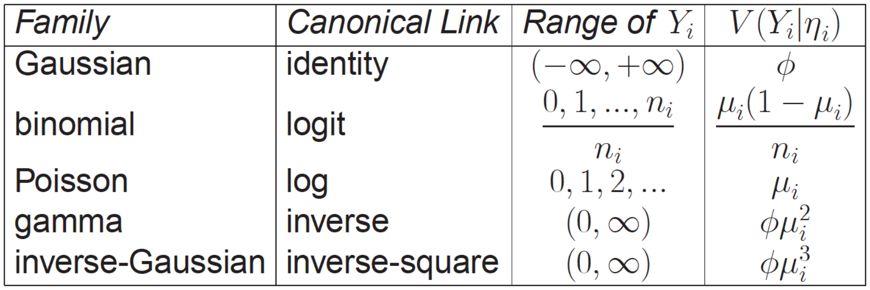
Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
identity	μ_i	η_i
log	$\log_e \mu_i$	e^{η_i}
inverse	μ_i^{-1}	η_i^{-1}
inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
square-root	$\sqrt{\mu_i}$	η_i^2
logit	$\log_e \dfrac{\mu_i}{1-\mu_i}$	$\dfrac{1}{1+e^{-\eta_i}}$
probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
log-log	$-\log_e[-\log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$
complementary log-log	$\log_e[-\log_e(1-\mu_i)]$	$1-\exp[-\exp(\eta_i)]$

Figure 3: Standard link functions and their inverses

The logit, probit, and complementary-log-log links are for binomial data, where Y_i represents the observed proportion and μ_i the expected proportion of “successes” in n_i binomial trials — that is, μ_i is the probability of a success. An special case is binary data, where there is $n = 1$ binomial trials (thus, Bernoulli), and therefore all of the observed proportions are either 0 or 1.

For the probit link, Φ, is the standard-normal cumulative distribution function, and Φ^{-1} is the standard-normal quantile function.

For distributions in the exponential families, the conditional variance of Y is a function of the mean, μ, together with a dispersion parameter, ϕ (as shown in the table below).



Family	*Canonical Link*	*Range of Y_i*	$V(Y_i\|\eta_i)$
Gaussian	identity	$(-\infty, +\infty)$	ϕ
binomial	logit	$\dfrac{0, 1, ..., n_i}{n_i}$	$\dfrac{\mu_i(1-\mu_i)}{n_i}$
Poisson	log	$0, 1, 2, \ldots$	μ_i
gamma	inverse	$(0, \infty)$	$\phi\mu_i^2$
inverse-Gaussian	inverse-square	$(0, \infty)$	$\phi\mu_i^3$

Figure 4: Canonical Links

For the binomial and Poisson distributions, the dispersion parameter is fixed to 1.

For the Gaussian distribution, the dispersion parameter is the usual error variance, which we previously symbolized by σ_ϵ^2 (and which does not depend on μ).

The *canonical link* for each familiy is not only the one most commonly used, but also arises naturally from the general formula for distributions in the exponential families.

Some examples:

- Combining the identity link with the Gaussian family produces the normal linear model. The maximum-likelihood estimates for this model are the ordinary least-squares estimates.
- Combining the logit link with the binomial family produces the logistic regression model.
- Combining the probit link with the binomial family produces the probit model.
- The log-log or complementary log-log link may be appropriate when the probability of the response as a function of the linear predictor approaches 0 and 1 asymmetrically.

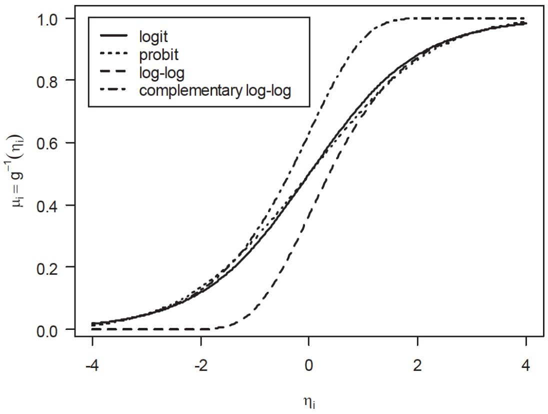


Figure 5: Comparison of logit, probit, and complementary log-log links. The probit link is rescaled to match the variance of the logistic distribution, $\pi^2/3$.

Other links may be more appropriate for the specific problem at hand.

One of the strengths of the GLM paradigm — in contrast, for example, to transformation of the response variable in a linear model — is the separation of the link function from the conditional distribution of the response.

GLMs are typically fit to data by the method of maximum likelihood.

Denote the maximum-likelihood estimates of the regression parameters as $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$.

These imply an estimate of the mean of the response

$$\widehat{\mu}_i = g^{-1}(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_k x_{ik})$$

The log-likelihood for the model, maximized over the regression coefficients, is

$$\log_e L_0 = \sum_{i=1}^{n} \log_e p(\widehat{\mu}_i, \phi; y_i)$$

where $p(\cdot)$ is the probability or probability density function corresponding to the family employed.

A “saturated” model, which dedicates one parameter to each observation, and hence fits the data perfectly, has log-likelihood

$$\log_e L_1 = \sum_{i=1}^{n} \log_e p(y_i, \phi; y_i)$$

Twice the difference between these log-likelihoods defines the *residual deviance* under the model, a generalization of the residual sum of squares for linear models:

$$G_0^2 = 2(\log_e L_1 - \log_e L_0)$$

Dividing the deviance by the estimated dispersion produces the scaled deviance: $G_0^2/\widehat{\phi}$.

Likelihood-ratio tests can be formulated by taking differences in the residual deviance for nested models.

Wald tests for individual coefficients are formulated using the estimated asymptotic standard errors of the coefficients.

GLM Diagnostics

Most regression diagnostics extend straightforwardly to generalized linear models.

Hat Values

Hat-values for a generalized linear model can be taken directly from the final iteration of the estimation procedure. They have the usual interpretation — except that the hat-values in a GLM depend on Y as well as on the configuration of the X's.

Residuals

Several kinds of residuals can be defined for generalized linear models:

Raw residuals are simply the differences between the observed response and its estimated expected value: $Y_i - \widehat{\mu}_i$

Pearson residuals are case-wise components of the Pearson goodness-of-fit statistic for the model:

$$e_i = \frac{\widehat{\phi}^{1/2}(Y_i - \widehat{\mu}_i)}{\sqrt{\widehat{V}(Y_i|\eta_i)}}$$

where ϕ is the dispersion parameter for the model and $V(Y_i|\eta_i)$ is the variance of the response given the linear predictor.

Standardized Pearson residuals correct for the conditional response variation and for the leverage of the observations:

$$r_i = \frac{e_i}{\sqrt{(1 - h_i)}}$$

Deviance residuals, d_i, are the square-roots of the case-wise components of the residual deviance, attaching the sign of $Y_i - \widehat{\mu}_i$.

Standardized deviance residuals are given as

$$s_i = \frac{d_i}{\sqrt{\widehat{\phi}(1 - h_i)}}$$

Several different approximations to studentized residuals have been suggested.

To calculate exact studentized residuals would require literally refitting the model deleting each observation in turn, and noting the decline in the deviance.

Here is an approximation

$$t_i = \sqrt{(1-h_i)s_i^2 + h_i r_i^2}$$

where, once again, the sign is taken from $Y_i - \widehat{\mu}_i$

A Bonferroni outlier test can also be used to identify outliers.

Influence

An approximation to Cook's distance influence measure is

$$D_i = \frac{r_i^2}{\widehat{\phi}(k+1)} \times \frac{h_i}{1-h_i}$$

Approximate values of dfbeta_{ij} and dfbetas_{ij} (influence and standardized influence on each coefficient) may also be obtained.

Poisson Regression

Poisson models arise in two common formally identical but substantively distinguishable contexts:

1. when the response variable in a regression model takes on non-negative integer values, such as a count;
2. to analyze associations among categorical variables in a contingency table of counts.

A random variable, Y, has a Poisson distribution with parameter μ if it takes integer values $0, 1, 2, \ldots$ with probability

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

for $\mu > 0$. The mean and the variance of the Poisson distribution is μ.

Since the mean is equal to the variance, any factor that affects one will also affect the other. Thus, the assumption of homoscedasticity would not be appropriate for Poisson data. The Poisson model captures very well the fact that, as is often the case with count data, the variance tends to increase with the mean.

The Poisson distribution also provides an approximation to the binomial distribution for the analysis of rare events, where π is small and n is large. A useful property of the Poisson distribution is that the sum of independent Poisson random variables is also Poisson.

The canonical link for the Poisson family is the log link. Suppose that we have a sample of n observations of $y_i, y_2, \ldots, y_n$ which are realizations of independent Poisson random variables, and we want to let the mean, μ, depend on a set of predictors. Using

$$\mu_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}$$

may result in predicted values that are negative. A straightforward solution to this problem is to model the logarithm of the mean using a linear model.

Thus, $\eta_i = \log(\mu_i)$ and the model can be written as

$$\log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}$$

Note that this is not the same as modeling the mean of log-transformed responses! That is,

$$\log(E[Y|x]) \neq E(\log Y|x)$$

The regression coefficient β_j represents the expected change in the log of the mean per unit change in the predictor, X_j. In other words increasing X_j by one unit is associated with an increase of β_j in the log of the mean.

The model for the mean itself is *multiplicative* and can be obtained by exponentiating.

$$\mu_i = e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_k X_{ik}}$$

The exponentiated regression coefficient, e^{β_j}, represents a multiplicative effect of the jth predictor on the mean. Increasing X_j by one unit multiplies the mean by a factor e^{β_j}.

It is important to note, however, that the model is additive in the log scale. In the original scale, the model is multiplicative, and postulates relative effects which translate into different absolute effects depending on the values of the other predictors.

A further advantage of using the log link stems from the empirical observation that with count data the effects of predictors are often multiplicative rather than additive. That is, one typically observes small effects for small counts, and large effects for large counts. If the effect is in fact proportional to the count, working on the log scale leads to a much simpler model.

The model is fit using maximum likelihood estimation. A measure of discrepancy between observed and fitted values is the deviance:

$$G^2 = 2\sum[y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) - (y_i - \widehat{\mu}_i)]$$

For large samples, the distribution of the deviance is approximately a χ^2 with $n-k$ degrees of freedom. Thus, the deviance can be used directly to test the goodness of fit of the model.

An alternative measure of goodness of fit is Pearson's chi-squared statistic:

$$\chi^2 = \sum \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}$$

In large samples, the distribution of Pearson's statistic is also approximately χ^2 with $n-k$ degrees of freedom.

An advantage of the deviance is that it can be used to compare nested models. The difference in deviances between two nested models is approximately χ^2 with degrees of freedom equal to the difference in the number of parameters between the models, under the assumption that the smaller model is correct.

Overdispersion

There is no such thing as overdispersion in ordinary linear regression because the variance of the normal distribution does not depend on the mean. But for discrete response variables, the commonly used distributions specify particular relationships between the mean and the variance and thus the variance is not estimated separately from the mean function as it is in ordinary linear regression.

Overdispersion occurs when there is greater variability in the data than would be expected based on the statistical model. In the context of count data, overdispersion would occur if $var(Y) > \mu$.

The binomial and Poisson GLMs fix the dispersion parameter, ϕ, to 1. But if we let $\phi > 1$ then we have overdispersion, and if $\phi < 1$ then we have underdispersion, although this is rare.

Now instead of assuming the variance is equal to the mean, consider the assumption that the variance is proportional to the mean, $var(Y) = \phi E(Y) = \phi\mu$.

It is possible to fit versions of the binomial and Poisson GLMs in which the dispersion is a free parameter, to be estimated along with the coefficients of the linear predictor. The resulting error distribution is not an exponential family. The models are fit by "quasi-likelihood."

The regression coefficients are unaffected by allowing dispersion different from 1, but because the estimated dispersion typically exceeds 1, this inflates the standard errors. The coefficient standard errors are multiplied by the square-root of $\widehat{\phi}$, where

$$\widehat{\phi} = \frac{\chi_k^2}{n-k}$$

Failing to account for "over-dispersion" produces misleadingly small standard errors.

The *over-dispersed* binomial and Poisson models arise in several different circumstances.

For example, in modeling proportions, it is possible that the probability of success μ varies for different individuals who share identical values of the predictors (this is called "unmodeled heterogeneity"); or the individual successes and failures for a "binomial" observation are not independent, as required by the binomial distribution.

In binomial models, overdispersion means that the true variance of y_i is greater than $n_i\pi_i(1-\pi_i)$.

If our model is correct, the Pearson residuals should behave like standardized residuals, i.e. like standard normal variates. But if their variance is substantially larger than one - and this extra variation is "spread across" the observational units, rather than concentrated in a small number of outliers - then we have evidence of overdispersion.

Rather than correcting the standard errors, an alternative approach to modeling over-dispersion in count data is to start from a Poisson regression model and add a multiplicative random effect, θ, to represent unobserved heterogeneity. This leads to the negative binomial regression model.

Negative Binomial Regression Model

A random variable, Y, has a Negative Binomial distribution with parameters p and r if it takes integer values $r, r+1, r+2, \ldots$ with probability

$$Pr(Y = y) = \left(\begin{array}{c} y-1 \\ r-1 \end{array} \right) p^r (1-p)^{y-r}$$

where the mean is $E(Y) = \mu = \frac{r}{p}$ and the variance is $Var(Y) = \frac{r(1-p)}{p^2} = \mu(\frac{1}{p} - 1)$.

We make an assumption regarding the distribution of θ and "integrate it out" of the likelihood, effectively computing the unconditional distribution of the outcome. It turns out to be mathematically convenient to assume that θ has a gamma distribution with parameters α and β. The gamma distribution has a mean of α/β and a variance of α/β^2. If we let $\alpha = \beta = 1/\sigma^2$, then the mean of the unobserved effect is one and its variance is σ^2. Then, the unconditional distribution of the outcome has a negative binomial distribution:

$$Pr(Y = y) = \frac{\Gamma(\alpha + y)}{y!\Gamma(\alpha)} \frac{\beta^\alpha \mu^y}{(\mu + \beta)^{\alpha + y}}$$

The negative binomial distribution with $\alpha = \beta = 1/\sigma^2$ has a mean $E(Y) = \mu$ and variance $var(Y) = \mu(1+\sigma^2\mu)$. If $\sigma^2 = 0$ then there is no unobserved heterogeneity and we obtain the Poisson variance. If $\sigma^2 > 0$, then the variance is larger than the mean.

Zero-Inflation

Another common problem with count data models, including both Poisson and negative binomial models, is that empirical data often show more zeroes than would be expected under either model.

The zero-inflated Poisson (zip) model postulates that there are two "latent classes" of people. For example, if we are studying number of alcoholic drinks over some period of time, there are some people who never drink and they would always be zero, and the rest who are not always zero, for whom the number of drinks has a Poisson distribution with $\mu > 0$. The model combines a logit model that predicts which of the two latent classes a person belongs to, with a Poisson model that predicts the outcome for those in the second latent class. In this model there are two kinds of zeroes: some are structural zeroes from the "always zero" class, and some are random zeroes from the other class.

Specifically, let π_i be the probability that the response Y_i for the ith individual is necessarily 0. The ZIP model turns out to be a piecewise probability function as follows:

$$P(Y = y_i) = \begin{cases} \pi_i + (1-\pi_i)e^{-\mu_i} & , y_i = 0 \\ (1-\pi_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} & , y_i > 0 \end{cases}$$

Accordingly, the two models are given by:

$$\log_e \frac{\pi_i}{1-\pi_i} = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \ldots + \gamma_p z_{ip}$$

where z_{ij} are the regressors for predicting membership in the first latent class, and

$$\log_e \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

where μ_i is the expected count for an individual in the second latent class, and the x_{ij} are regressors for the Poisson submodel. The two sets of regressors, z_{ij} and x_{ij} are often the same but they can be different.

The probability of observing a 0 count is

$$Pr(Y_i = 0) = \pi_i + (1-\pi_i)e^{-\mu_i}$$

The conditional expectation and variance of Y_i are:

$$E(Y_i) = (1-\pi_i)\mu_i Var(Y_i) = (1-\pi_i)\mu_i(1+\pi_i\mu_i)$$

The zip model and the Poisson model are nested because the Poisson model corresponds to the zip model where the probability of "always zero" is zero for everyone.

A zero-inflated negative binomial model combines a logit model for the latent classes with a negative binomial for the counts in the "not always zero" class.

Another approach to excess zeroes is to use a logit model to distinguish counts of zero from larger counts, effectively collapsing the count distribution into two categories, and then use a truncated Poisson model, namely a Poisson distribution where zero has been excluded, for the positive counts. This is called a hurdle model.

This approach differs from the zero-inflated models because the classes are observed rather than latent, one consists of observed zeroes and the other of observed positive counts.