EPSE 682: Multivariate Analysis

Ed Kroc

University of British Columbia

ed.kroc@ubc.ca

February 8, 2022

Some added difficulties of Binomial regression

- Interpretation of model coefficients can be more challenging (i.e., odds/log-odds ratios rather than raw rates of change in a continuous response).
- Validating a model (i.e., residual diagnostics more difficult).
- Building a best fitting valid model (i.e., hard to diagnose model misspecification like missing curvature, interactions, etc.).
- Statistical power to detect/estimate effects is *structurally, intrinsically decreased*.
- Complete or quasi-complete separation of the binary response over a predictor or set of predictors.
- Over/underdispersion of the conditional response process (i.e., a kind of violation of the i.i.d. Bernoulli assumption of the Binomial model).
- Note: **only correct for over/underdispersion after you have constructed a valid (or valid enough) model**.

Multinomial regression

- There is a natural extensions of the Binomial (and Quasibinomial) model to the situation where the response variable has more than two possible outcomes; i.e., when we can have $Y \in \{0, 1, \ldots, k\}$.
- In general, such a random variable is *categorical*, but just as we assumed i.i.d. Bernoulli trials (conditionally) for Binomial regression, one can assume i.i.d. $\{0, 1, \ldots, k\}$ trials (conditionally) here and thus define a *Multinomial model* for the conditional response process (there are other new assumptions one has to make too).
- Furthermore, the Multinomial model has been refined to (several different) *Ordinal Multinomial* models, appropriate when the response categories $\{0, 1, \ldots, k\}$ actually have a meaningful order to them; i.e., when they are not just placeholder labels to distinguish nominal categories.

Multinomial regression

- As a general rule, multinomial regression should be **avoided** at all costs, unless you are working with at least 500–1000's of observations per predictor, and assuming those observations are pretty nicely spread out across *all* your response categories.

- A multinomial model relies on suspicious assumptions, is notoriously difficult to perform diagnostics with, and suffers from extremely low power. All the difficulties of Binomial regression are severely amplified, and the assumed multinomial structure of the data is a far more restrictive assumption than the two-category binomial one.

- Moreover, it is far harder to fix model misspecification by moving to a quasilikelihood framework for responses with more than two categories.

- Avoid Ordinal Multinomial models at all costs. I have never seen a reasonable application in practice.

Multinomial regression

- The response data Y are now categorical on *more than two categories*, so take values in $\{0, 1, 2, \ldots, k\}$.
- Just as in regression with Binomial (0/1) data, if we can model the *probabilities* of Y falling into any of its $k + 1$ categories, then we know everything about the random phenomenon.
- Now though, there are k probabilities that we need to model:

$$\begin{aligned}
\pi_1 &:= \Pr(Y = 1 \mid \boldsymbol{X}) \\
\pi_2 &:= \Pr(Y = 2 \mid \boldsymbol{X}) \\
&\vdots \\
\pi_k &:= \Pr(Y = k \mid \boldsymbol{X})
\end{aligned}$$

- Note that if we knew these, then we would automatically know

$$\pi_0 := \Pr(Y = 0 \mid \boldsymbol{X}) = 1 - \pi_1 - \pi_2 - \cdots - \pi_k$$

Multinomial regression

- The most typical way you see this modelled is via *multinomial logistic regression*, another kind of GLM.
- Remember: 2 pieces to define a GLM:
  (1) a likelihood for the data, and
  (2) a link function relating the mean of the response (conditional on the predictors) to a linear combination of the predictors.
- Here, we assume a *Multinomial* model for the response data and a logit link.

Multinomial regression

- Explicitly, suppose our m_i sample observations at $\boldsymbol{X} = \boldsymbol{x}$ generate z_0 counts of $Y = 0$, z_1 counts of $Y = 1, \ldots,$ and z_k counts of $Y = k$. Then:

$$\ell(\pi_0, \pi_1, \ldots, \pi_k \mid \boldsymbol{x}, \boldsymbol{y}) = \prod_i \binom{m_i}{z_0, z_1, \ldots, z_k} \pi_0^{z_0} \pi_1^{z_1} \cdots \pi_k^{z_k},$$

  where $\binom{m_i}{z_0, z_1, \ldots, z_k}$ is a *multinomial coefficient* that counts the number of ways to simultaneously select z_0 objects of type 0, z_1 objects of type $1, \ldots,$ and z_k objects of type k from m_i objects. [Think: how many ways could you arrange 2 green balls, 3 red balls, and 1 blue ball in a straight line?]

- Notice: if you sub-in $k = 1$, this model reduces back to the Binomial model.

Multinomial logistic regression

- Multinomial regression now requires you to declare a *reference* category and then posits a *link* function between the mean number of outcomes in each of the k other categories.
- In Binomial regression, the choice was arbitrary (i.e. either "0" or "1" could be declared a Bernoulli "success"), but now we have options; the decision is made according to the research problem.
- We will declare $Y = 0$ as the reference category. Then multinomial *logistic* regression posits:

$$\log\left(\frac{\pi_i}{\pi_0}\right) = \sum_{j=0}^{p} \beta_{i,j} X_j, \quad 1 \leqslant i \leqslant k$$

  Notice that we are now modelling k different logits, each with their own set of coefficients that capture the linear associations between the predictor and logit response.

Multinomial logistic regression

- Everything now proceeds as with Binomial logistic regression, except we would no longer talk about the "odds of success" increasing or decreasing with values of $\boldsymbol{X}$ (estimated via the model coefficients).
- Instead, now we would talk about the "odds of being in category i relative to the reference category 0". E.g. odds of voting NDP vs. Conservative, Green vs. Conservative, and Liberal vs. Conservative.
- One major problem with this kind of model is that it assumes *independence of irrelevant alternatives* (IIA). That is, removing one category from consideration should *not* change the odds of responding in another category.
- This assumption is often unreasonable in practice. E.g. if there were no Green candidates, it is very likely that more people would vote NDP and/or Liberal vs. Conservative, thus, changing these odds.
- Some other multinomial regression models relax this assumption, but some version of it is always there.

Multinomial regression

- The IIA assumption is really a critically bad one, but usually necessary to make the estimation problem tractable.
- Relaxing the multinomial structure by moving to quaslikelihood does not work nearly as well as in the Binomial case.
- With categorical response data in practice, it's far more typical to see very imbalanced responses across the categories than balanced ones, and this is going to make modelling very difficult; in particular, your model will be driven almost entirely by the categories that contain most of the responses.
- In practice, if you are forced to model a categorical response process on more than 2 categories and do not have truly "big data," either:
  (1) Break the response process up into multiple Binomial/qBin responses and perform separate logistic regressions, or,
  (2) Collapse response categories down to just 2 and perform a single Binomial/qBin regression.

Multinomial regression

- For example, instead of modelling Green, NDP, Liberal, vs. Conservative as a single categorical response, model:
    - Green vs. not as a Binomial response
    - NDP vs. not as a Binomial response
    - Liberal vs. not as a Binomial response
    - And Conservative vs. not as a Binomial response
- Or maybe modelling Liberal vs. Conservative, Green vs. Conservative, NDP vs. Conservative will be more relevant for your research questions.

Some important and common GLMs

- **Ordinary linear regression** (Gaussian/Normal regression with identity link): continuous, well-behaved response data.
- **Logistic regression** (i.e. Binomial logistic regression), **quasi-Binomial** regression: binary response data.
- Multinomial regression: categorical response with more than two categories (usually AVOID at all costs).
- **Poisson** regression, **ZIP** regression, **quasi-Poisson** regression, negative binomial regression: count response data.
- Gamma regression, lognormal regression, inverse Gaussian regression: continuous, but skewed error structure.
- Beta regression, ZOIB regression: proportion/percentage (0-100%) response data.
- ...and many others. We will talk about most of these at least a little.

Count response processes

- Binomial response process: for every individual in the population, we can observe either a "success" or a "failure", i.e. 0/1 response.
- Count response process: for every individual in the population, we can observe a nonnegative integer: $0, 1, 2, 3, \ldots$.
- Examples:
    - Number of years of post-secondary school, say, modelled as a function of high school GPA, parents' number of years of post-secondary, and socioeconomic status.
    - Number of snow leopards observed at various camera traps over a one-year period, say, modelled as a function of elevation, snow cover, and viable prey count.
    - Number of online purchases made at a particular retail site, say, modelled as a function of income, credit line, and total debt.

Poisson random variables

- There are several common probability distributions that are used to model count response data: Poisson, Negative Binomial, Quasipoisson, Zero-inflated Poisson (ZIP), and Quasi-ZIP.
- The *Poisson random variable* is the simplest and we will start there.
- We say $Y \sim \mathit{Poisson}(\lambda)$, for some positive parameter $\lambda > 0$, if

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

  for each $y = 0, 1, 2, 3, \ldots$. Note that $y! = y \cdot (y-1) \cdots 2 \cdot 1$, and that $0! = 1$ (for very good mathematical reasons that are unfortunately too technical for us to get into).
- Note that count random variables, like the Poisson, are *discrete* but can take on (theoretically) any one of infinitely many values.

Poisson random variables

- We say $Y \sim \textit{Poisson}(\lambda)$ if

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

for each $y = 0, 1, 2, 3, \ldots$.

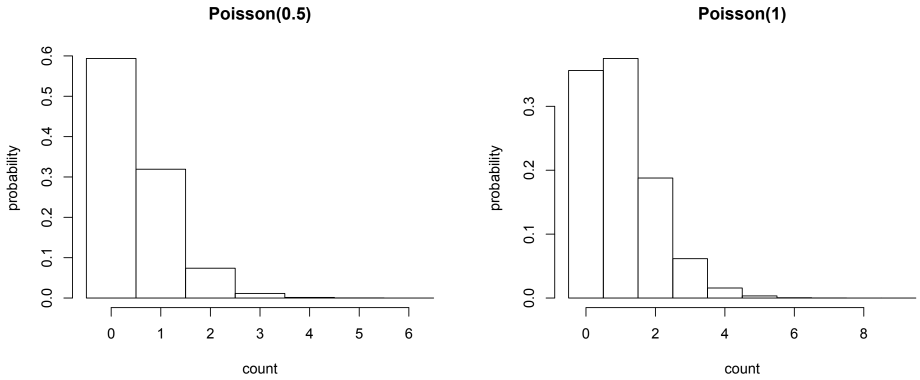
- One can show (with calculus) that:

$$\mathbb{E}(Y) = \lambda, \qquad \mathrm{Var}(Y) = \mathbb{E}(Y) = \lambda.$$

So, just as with Binomial phenomena, the *mean* of a Poisson response automatically determines its *variance*.

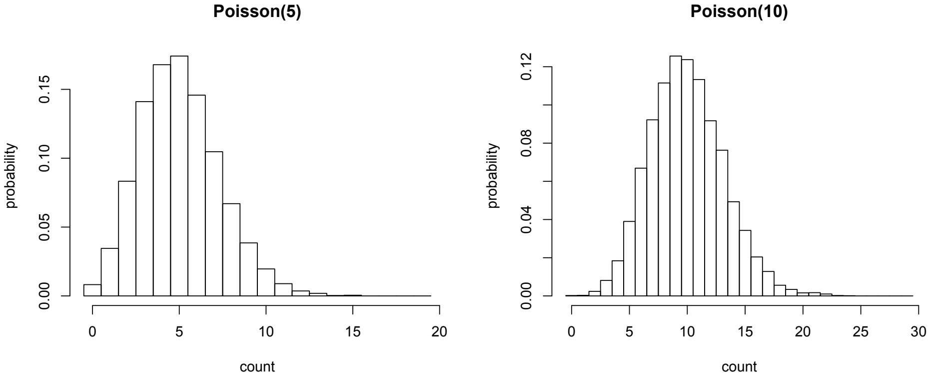
- Notice that this means all the implied problems from binomial regression (i.e., separation, over/underdispersion) are problems in Poisson regression.

Poisson random variables



- Explore the shape of the Poisson distribution for different values of its parameter here: `https://homepage.divms.uiowa.edu/~mbognar/applets/pois.html`

Poisson random variables



- Explore the shape of the Poisson distribution for different values of its parameter here: `https://homepage.divms.uiowa.edu/~mbognar/applets/pois.html`

Poisson regression

- We may choose to model count response processes via **Poisson regression with a log link** (to remove restrictions on the mean values of the response process):

  (1) Link: $g(\mathbb{E}(Y \mid \boldsymbol{X})) = \log(\mathbb{E}(Y \mid \boldsymbol{X})) = \sum_{j=0}^{p} \beta_j X_j$

  (2) Likelihood:

$$\ell(\boldsymbol{\beta} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{y}) = \prod_{i=1}^{n} \frac{\exp(-\exp(\sum_{j=0}^{p} \beta_j x_{ji})) \cdot \exp(y_i \sum_{j=0}^{p} \beta_j x_{ji})}{y_i!}$$

- This looks really ugly, but notice that we are just plugging into the expression for the probability mass function of a Poisson random variable with (conditional) mean parameter:

$$\lambda = \exp(\sum_{j=0}^{p} \beta_j X_j)$$

Poisson regression

- Now, just as with any GLM, we can estimate the unknown model coefficients $\beta_0, \ldots, \beta_p$ using maximum likelihood estimation (MLE).
- Realize: A Poisson model assumes a very strong (and often unreasonable) relationship between the (conditional) mean and the (conditional) variance of the response process:

$$\mathrm{Var}(Y \mid \boldsymbol{X}) = \mathbb{E}(Y \mid \boldsymbol{X})$$

  Or, put another way, *a Poisson model assumes that the (conditional) standard deviation of the response process is always the square root of the (conditional) mean response*.

- Over/underdispersion are often major concerns here.
- Also, a new phenomenon will be an arguably even bigger concern: **zero inflation**.

Quasipoisson regression

- We may choose to model over/underdispersed count response processes via **Quasipoisson regression with a log link function**:

  (1) Link: $g(\mathbb{E}(Y \mid \boldsymbol{X})) = \log(\mathbb{E}(Y \mid \boldsymbol{X})) = \sum_{j=0}^{p} \beta_j X_j$

  (2) Quasilikelihood:

$$\ell_{quasi}(\beta_0, \ldots, \beta_p \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{y}) = \phi \cdot \ell(\beta_0, \ldots, \beta_p \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{y})$$

- Now, Quasipoisson distribution for the (conditional) response process implies the weaker condition:

$$\mathrm{Var}(Y \mid \boldsymbol{X}) = \phi \cdot \mathbb{E}(Y \mid \boldsymbol{X})$$

- Note though that this still means the (conditional) variance is a *linear function* (constant multiple) of the (conditional) mean. **Negative Binomial** regression will weaken this structure.

Zero inflation

- With count data in the real world, there are often *far more observed zero counts* than what would be expected under a simple Poisson structure.
- This means that those histograms we saw for Poisson random variables would be putting far too little weight (probability) on the chance of observing a count of zero.
- Such data often arise when there are actually *two* response processes at work simultaneously: one binary $(0/1)$ that indicates whether or not there will be *any* events, and one that tells us how many events will occur (positive integer-valued: $1, 2, 3, \ldots$).
- Zero-inflation is likely to occur in any or all of the three count response process examples we originally considered:

Zero inflation

Average number of years of post-secondary school, say, modelled as a function of high school GPA, parents' number of years of post-secondary, and socioeconomic status.

- There are very likely other predictors here that are important to determine if somebody *actually even pursues post-secondary*.
- Such people would receive a count of '0' for their response.
- But these people are likely *fundamentally different* than those who start post-secondary schooling but don't make it through the first year, yet those people will also receive a count of '0' for their response.
- Thus, it's plausible that we could be trying to model *two response processes at once*: (1) to predict if someone will enrol in post-secondary, and (2) to predict how long they will stay there *assuming they already enrolled*.

Zero inflation

Average number of snow leopards observed at various camera traps over a one-year period, say, modelled as a function of elevation, snow cover, and viable prey count.

- There are very likely other predictors here that are important to determine if a snow leopard will *even ever be in the area of the camera trap*, e.g. presence/absence of a competing apex predator.
- Thus, some of the '0' responses will simply be recording the fact that snow leopards do not use the land because of predator/territorial competition, while others may reflect the fact that snow leopards just do not use the land because of lack of prey, snow cover, etc..
- Thus, it's plausible that we could be trying to model *two response processes at once*: (1) to predict if a snow leopard *could* be in the area, and (2) to predict how many times they are in the area *assuming they have access to it*.

Zero inflation

Average number of online purchases made at a particular retail site, say, modelled as a function of income, credit line, and total debt.

- There are very likely other predictors here that are important to determine if somebody is likely to *buy anything* at the site, like if they actually need any of the items for sale or not.
- Thus, some of the '0' responses will simply be there because the respondent didn't *need/want* anything, while others may be there because the respondent couldn't *afford/justify* a purchase.
- Thus, it's plausible that we could be trying to model *two response processes at once*: (1) to predict if people are interested in the retail items, and (2) to predict how much they will buy *assuming that they are already interested*.

Example: Modelling the number of snow leopards

- Response data: Number of snow leopards observed at various camera traps over a one-year period.
- Predictors: elevation (roughly 4000 to 6000 m), snow cover (0% to 100%), and a measure of prey count (0% to 100%).
- Natural to start with a Poisson GLM (log link):

Example: Modelling the number of snow leopards

Poisson GLM (log link):

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.5603845  0.7337032  -2.127   0.0334 *
elev        -0.0000751  0.0001355  -0.554   0.5795
snow         2.0207290  0.3616166   5.588 2.30e-08 ***
prey         3.0486225  0.4506277   6.765 1.33e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 683.07  on 249  degrees of freedom
Residual deviance: 609.76  on 246  degrees of freedom
AIC: 819.81
```

- Note: model estimates are on the log-scale (because of the link).

Example: Modelling the number of snow leopards

- So, based on our model, at 4000 m elevation, 50% prey availability, and 15% snow cover, we expect (on average) about 1 leopard sighting every year:

$$\widehat{\mathbb{E}}(Y \mid \boldsymbol{X}) = \exp(-1.56 - 0.000075 \cdot (4000) + 2.02 \cdot (0.15) + 3.05 \cdot (0.5)$$
$$= 0.967$$

- Whereas if the snow cover was 50% instead, we expect (on average) about 2 leopard sightings every year:

$$\widehat{\mathbb{E}}(Y \mid \boldsymbol{X}) = \exp(-1.56 - 0.000075 \cdot (4000) + 2.02 \cdot (0.5) + 3.05 \cdot (0.5))$$
$$= 1.962$$

- Interpreting coefficients: A one percent increase in snow cover corresponds to an increase of about $e^2 = 7.5$ of the *incidence rate ratio* (relative risk) of leopard sightings per year; i.e. strong positive relationship between snow cover and leopard sightings.

Example: Modelling the number of snow leopards

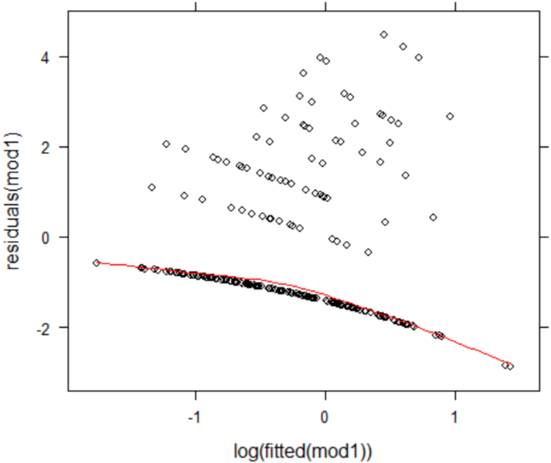
- Note: since we use the log link function, we have that:

$$
\begin{aligned}
\frac{\mathbb{E}(Y \mid x_1 + 1, x_2, \ldots, x_p)}{\mathbb{E}(Y \mid x_1, x_2, \ldots, x_p)} &= \frac{\exp\left(\beta_0 + \beta_1(x_1 + 1) + \sum_{j=2}^{p} \beta_j x_j\right)}{\exp\left(\beta_0 + \beta_1 x_1 + \sum_{j=2}^{p} \beta_j x_j\right)} \\
&= \frac{\exp(\beta_0) \cdot \exp(\beta_1(x_1 + 1)) \cdot \exp\left(\sum_{j=2}^{p} \beta_j x_j\right)}{\exp(\beta_0) \cdot \exp(\beta_1 x_1) \cdot \exp\left(\sum_{j=2}^{p} \beta_j x_j\right)} \\
&= \frac{\exp(\beta_1 x_1) \cdot \exp(\beta_1)}{\exp(\beta_1 x_1)} = \exp(\beta_1)
\end{aligned}
$$

- This ratio is often called the *incidence rate ratio* or the *relative risk*.
- Quantifies how much a (marginal) change in X_1 is expected to associate with a larger or smaller number of incidents. If $\exp(\beta_1) < 1$, then a marginal increase in X_1 is associated with fewer incidents; if $\exp(\beta_1) > 1$, then a marginal increase in X_1 is associated with more incidents.

Example: Modelling the number of snow leopards

Residuals vs. fitted values for our model:

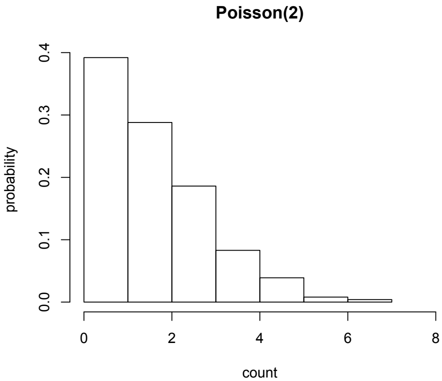


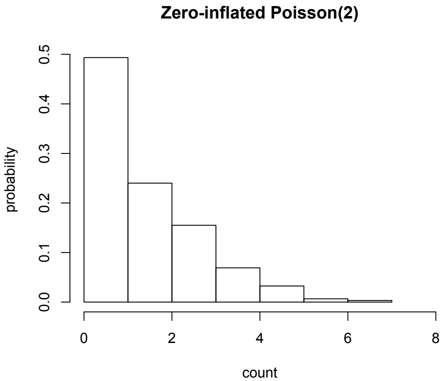
- Doesn't look great. The average residual is always below zero for any fitted/predicted value: *too many observed zeros?*

Example: Modelling the number of snow leopards

- Just like usual, could be missing important predictors, or not capturing interactions or marginal curvature.
- Under/overdispersion seems to be present based off the χ^2-test of residual deviance: 609 is in the $>$ 99th percentile of the $\chi^2(246)$ distribution: So **Quasipoisson** fit seems appropriate.
- Yet, quasilikelihood won't change the residuals at all (or the fitted values, or the coefficient estimates); it only adjusts the estimate of residual variance/deviance in order to fix the model standard errors.
- Seeing residuals consistently below zero on average suggests **zero inflation** is happening; i.e., there are more observed zeros than what should be there a Poisson response process. But both structural and conditional zeros may be present in the response process: model via a **Zero-inflated Poisson (ZIP)** model.

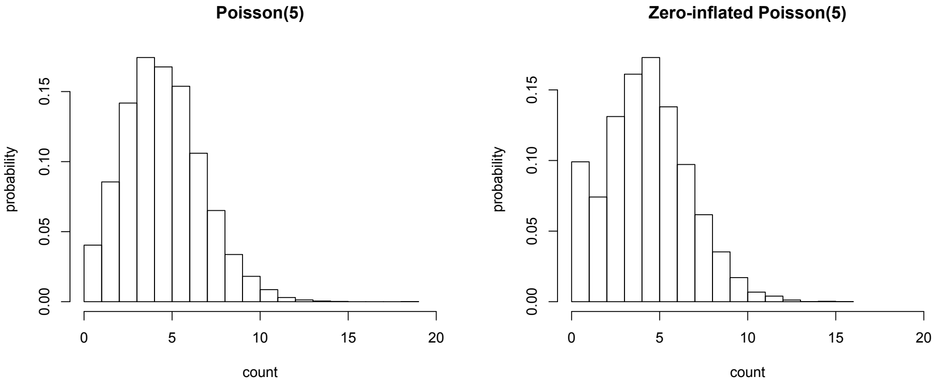
Zero inflation





- A zero-inflated Poisson (ZIP) distribution on the right; the chance of observing a zero count is about 25% higher than what it would be under an assumed Poisson (conditional) response distribution.

Zero inflation



- A zero-inflated Poisson (ZIP) distribution (right); the chance of observing a zero count is about twice as high as what it would be under an assumed Poisson (conditional) response distribution.

Zero-inflated Poisson (ZIP) regression

A ZIP model is an example of a *(conditional) mixture model*; i.e. it simultaneously models *two* different response processes.

- First, model the *structural zeros* with a *Binomial GLM*, logit link:

  (1) Link: $g_1(\mathbb{E}(Y_{0/1} \mid \boldsymbol{X})) = \text{logit}(\mathbb{E}(Y_{0/1} \mid \boldsymbol{X})) = \sum_{j=0}^{p_1} \gamma_j X_j$

  (2) Likelihood: Binomial

  (3) Response data: 0's and 1's (i.e. no chance of a count vs. chance of a count)

- Then, model the *incidence counts* for those observations that arose from *structural successes* with a *Poisson GLM*, log link:

  (1) Link: $g_2(\mathbb{E}(Y_{\geqslant 0} \mid \boldsymbol{X})) = \log(\mathbb{E}(Y_{\geqslant 0} \mid \boldsymbol{X})) = \sum_{j=0}^{p_2} \beta_j X_j$

  (2) Likelihood: Poisson

  (3) Response data: observed counts $\geqslant 0$.

Zero-inflated Poisson (ZIP) regression

Items of note about ZIP models:

- ZIP models (and mixture models in general) work by essentially creating *two different* response variables, then modelling each jointly.
- Hidden in the math though is the fact that the models are *dependent* on each other; i.e. the two models can be written with a *single likelihood function*, which is then (often) maximized to estimate model parameters. I have suppressed this likelihood function because it is very ugly.
- Consequently, assessing model fit via residual diagnostics operates *just as before*; i.e. diagnostics are performed on the residuals of the *composite* mixture model, not its individual pieces.
- The predictors do *not* have to be the same between the two parts of the ZIP model; they can in fact be entirely different.

Example: Modelling the number of snow leopards

Data: Number of snow leopards observed at various camera traps over a one-year period, say, modelled as a function of elevation, snow cover, and viable prey count.

- A series of famous studies in conservation ecology established a strong association between presence of snow leopards and snow cover (it appears causal, but no one knows why).
- Thus, modelling snow leopard abundance is likely to require some kind of mixture model, like a ZIP. That is, we should likely model *two response processes at once*:
  1. to predict if a snow leopard *could* be in the area (strongly dependent on snow cover); this will account for the *structural zeros*
  2. to predict how many times they are in the area *assuming they would consider accessing it* (dependent on additional variables).

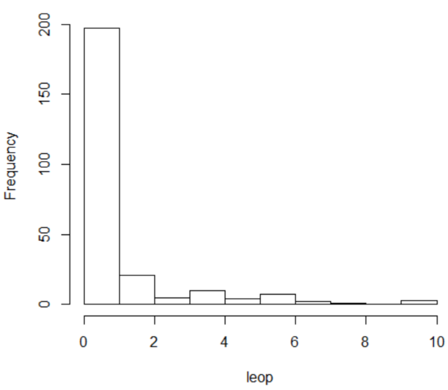
Example: Modelling the number of snow leopards

Recall our Poisson GLM fit (log link):

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.5603845  0.7337032  -2.127   0.0334 *
elev        -0.0000751  0.0001355  -0.554   0.5795
snow         2.0207290  0.3616166   5.588 2.30e-08 ***
prey         3.0486225  0.4506277   6.765 1.33e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 683.07  on 249  degrees of freedom
Residual deviance: 609.76  on 246  degrees of freedom
AIC: 819.81
```

- Note: model estimates are on the log-scale (because of the link).

Example: Modelling the number of snow leopards



- Looking at the raw counts (responses), we see an extremely high proportion of zeros, which *may* suggest zero-inflation.

- Note though that this diagnostic tool is *not* definitive. Because what really matters is if the response process *conditional on the predictors* is zero-inflated, not the raw response process itself (plotted).

Example: Modelling the number of snow leopards

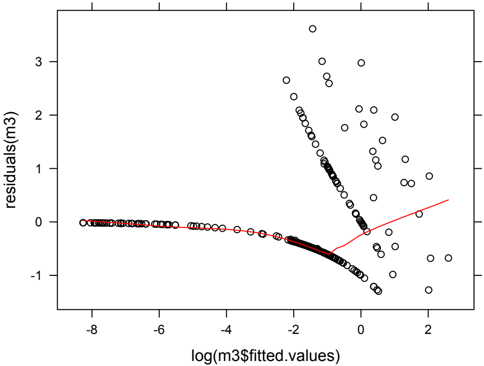
Fit a simple ZIP model to our data:

```
Call:
zeroinfl(formula = leop ~ elev + snow + prey | snow)

Pearson residuals:
     Min       1Q   Median       3Q      Max 
-1.29719 -0.50775 -0.35504 -0.01713  3.61594 

Count model coefficients (poisson with log link):
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.2077235  0.3018376  -0.688   0.4913    
elev        -0.0003956         NA      NA       NA    
snow         4.6955574  0.3228556  14.544   <2e-16 ***
prey         1.1252412  0.4524627   2.487   0.0129 *  

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)    6.013      1.609   3.737 0.000186 ***
snow           5.063      0.316  16.016  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Positive predictive effect of snow, and improved model fit: AIC $= 515$.
- But couldn't estimated std. error of 'elev' coefficient; common for numerical algorithms to break down in mixture models (bootstrapping can solve this).

Example: Modelling the number of snow leopards

Residuals vs. fitted values for our ZIP model:



- Looks way better! Notice also the improved predictive power, reflected in the larger scale/spread of fitted/predicted values.

ZIP regression

More notes about ZIP models:

- Interpretation of model coefficients is the same as for Poisson or Binomial GLMs, as appropriate. Just remember that you are partitioning the response process into *two different* response processes.

- It is quite common that ZIP models fit better than Poisson/Quasipoisson models.

- Estimation of model parameters and standard errors is more complex (less stable) for mixture models (like ZIP); thus, may run into problems like we saw when estimating standard errors. Can fix by simplifying model, or sometimes *by bootstrapping*.

- qZIP (Zero-inflated Quasipoisson) is theoretically possible, though I am not aware of any R packages that easily implement it ('pscl' implements regular ZIP models easily). Regardless, we can essentially fit a qZIP *by bootstrapping* an ordinary ZIP model.