

Ref: Zuur, Walker, Saveliev & Smith (2009) *Mixed Effects Models and Extensions in Ecology with R* give examples and analysis of zero-truncated models in Chapter 11. Poisson with no zeroes.

- If the mean of the Poisson process is large, zeros will be rare so no 0's is expected.
- Occasionally zeroes are missing values:

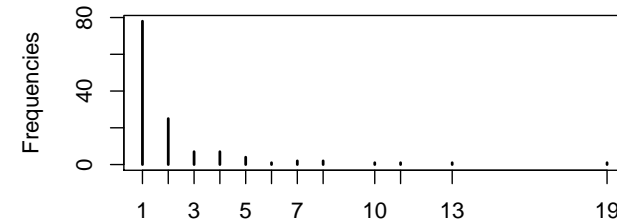
Ecologists in Portugal studied snakes killed on roadways and recorded the number of days a snake carcass lay on the road. A zero means the snake crossed the road, and was not part of the dataset. Even if the carcass was only hours old, a 1 was recorded.

	NB	Trunc.NB	SE NB	SE Trunc.NB
(Intercept)	0.37	-2.34	0.11	0.27
PDayRain	-0.00	0.10	0.19	0.46
Tot_Rain	0.12	0.27	0.02	0.07
Road_LocV	0.45	1.09	0.15	0.38
PDayRain:Tot_Rain	-0.11	-0.25	0.02	0.07

Truncated Negative Binomial has coefficients further from zero with greater SE.

With Zuur's AED library and VGAM library.

```
> plot(table(Snakes$N_days),ylab="Frequencies")
> snakefit1 <- MASS::glm.nb(N_days ~ PDayRain * Tot_Rain +
+                               Road_Loc, data=Snakes)
> snakefit2 <- vglm(N_days ~ PDayRain * Tot_Rain + Road_Loc,
+                   data=Snakes, family=posnegbinomial,
+                   control = vglm.control(maxit = 100))
```



Use negative binomial to account for overdispersion

Go to a site in the Greater Yellowstone Ecosystem and look for grizzly bears.

Why might we see zero bears?

- Habitat is not suitable.
- "Design error" It's winter and bears are all in dens.
- "Observer error" we missed signs which were there.
- "Bear error" good habitat, but bears haven't found it.
- Naughty naughts: bad zeroes from sampling outside the range (downtown Bozeman?) Remove these from the sample.

Hurdle models:

- ① Model zeroes versus non-zeroes as binomial with logistic regression.
- ② Given some were observed, use a truncated Poisson or negative binomial to model frequency.

Economists use censored regression models for variables like earnings or “labor supply”. Tobit models assume a latent variable with a threshold at which the response becomes positive (probit regression). Above the threshold use a linear model with the same predictors (same $\mathbf{X}\beta$) for the positive responses. Combined likelihood includes the binomial and linear model for those over threshold.

Mixture of two models

- ① $1 - p$ of the time we observe a “false” zero due to observer or bear error.
- ② p of the time we observe a Poisson or negative binomial, which could also give a zero (bear error or non-suitable habitat).

See Zuur et al. for examples.

- Survival data (log right tailed, censored) modeled with Gamma, Weibull, or proportional hazards models.
- Nonparametric models (highly parametric?)
 - gam generalized additive models
 - neural networks
 - support vector machines

for nonlinear trends

Decision theory assumes there are costs (minimize loss) and gains (maximize utility) for actions described by a *value* function.

- a_i is benefit of switching from unsafe to safe well (in \$?)
- $b_i + c_i x_i$ is the \$ cost of switching when well is $100x_i$ m away.

Logit or probit model: switch if $a_i > b_i + c_i x_i$

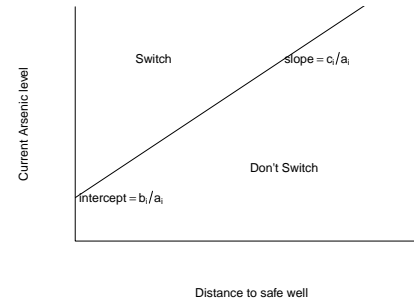
$$\Pr(y_i = 1) = \Pr\left(\frac{a_i - b_i}{c_i} > x_i\right)$$

a_i, b_i, c_i are not identifiable, but let $d_i = \frac{a_i - b_i}{c_i}$ and assume it has a logistic (normal) distribution with mean μ and spread σ .

$$\Pr(y_i = 1) = \Pr(d_i > x_i) = \Pr\left(\frac{d_i - \mu}{\sigma} > \frac{x_i - \mu}{\sigma}\right) = \text{logit}^{-1}\left(\frac{\mu - x}{\sigma}\right)$$

or in probit regression, $\Phi\left(\frac{\mu - x}{\sigma}\right)$. We need a slope and intercept for x .

Can estimate the population-average model, not individual values for each household.



Each individual has their own cost and value.

Switch if $a_i(A_s) > b_i + c_i x_i$.

Analysis shows people mistakenly used $\log(A_s)$ instead of A_s .