

Event count models

Silje Synnøve Lyder Hermansen

2020-11-19

The dependent variable

Count data

Count data is common in political science

- ▶ Discrete: consists only in integers (0, 1, 2, ... no digits)
- ▶ Bounded at zero, often long tail upwards.

Count models: What are they good for?

When do we use count models?

The data generating process allows us to

- ▶ observe and count a number of events and
- ▶ define a time frame or geographical space for the occurrence(s)

⇒ *e.g. number of meetings between decision makers, violent events, legislative proposals, etc.*

Why not a binomial logistic regression?

These are indeed binary outcomes but we don't have information on the event level

⇒ *Variables are on the exposure level; related to when (where) the events took place.*

Why not OLS?

The variable could be approximated to a continuous measure but

- ▶ it is bounded at zero, so predictions would be wrong → *same problems as logit*
- ▶ it is skewed. Some people add a constant and logtransform:
 $\log(y + 0.1)$ → *but heteroskedasticity and non normal errors remain*

⇒ *We replace the normal distribution with another probability distribution*

The generalized linear model strategy

There are many count models

- ▶ Poisson model: the base-line
- ▶ Other models: to address problems with the poisson

The Poisson model

Poisson process

The poisson distribution maps probabilities of events within a window to outcomes

- ▶ **Exposure ($t, t + h$):** A window of opportunity between two boundaries (geographical or spacial)
- ▶ **Probability of event (λ):** Simply the logtransformed mean of events within that window
 - ▶ Probability of event = $h\lambda$
 - ▶ Probability of no event = $1 - h\lambda$

Formula

The equation the model estimates:

$$E(y_i) \equiv h\lambda_i = h \times \exp(\alpha + \beta \times x_i) \quad (1)$$

Estimation of the exposure

What to do with the exposure parameter?

$$E(y_i) \equiv h\lambda_i = h \times \exp(\alpha + \beta \times x_i) \quad (2)$$

Two strategies :

- ▶ **Offset:** Move it into the equation but constrain parameter:
 $\exp(\alpha + \beta \times x_i + 1 \times \log(h_i)) \rightarrow$ *we don't see it in the BUTON*
- ▶ **Estimate a parameter:** $\exp(\alpha + \beta_1 \times x_i + \beta_2 \times \log(h_i))$

\Rightarrow *If the exposure is the same for all units, we set it to 1 and ignore it.*

Interpretation: back and forth

Interpretation is relatively easy with all count models

- ▶ Recoding (for estimation): we logtransform the mean of the y (within x -values)
- ▶ We back-transform (for interpretation): $\exp(\lambda)$ is simply an approximation (with digits) of our counts!

Interpretation: effects

Interpretation is relatively easy

- ▶ Recoding (for estimation): we logtransform the mean of the y (within x -values)
- ▶ We back-transform (for interpretation):
 - ▶ Predicted value: $\exp(\hat{\lambda})$ is simply an approximation (with digits) of our counts
 - ▶ Effect of β : $\exp(\beta)$ is multiplicative of predicted $\hat{\lambda} \rightarrow$ easy!

\Rightarrow *Make scenarios, predict, knock yourself out*

Dispersion

The main assumption of the Poisson model

The model assumes equidispersion: The spread equals the mean

- ▶ The y can be overdispersed, but not the $\hat{\lambda} \rightarrow$ as in OLS

\Rightarrow *The standard errors will be too small*

Identifying overdispersion

- ▶ Poissonness plot
- ▶ Rootograms
- ▶ Formal tests: Using residuals and significance tests.

Reasons for overdispersion

- ▶ Lack of exposure time
- ▶ Poor choice of variables (include more, also random intercepts)
- ▶ Too many zeros
- ▶ Events are related

Addressing overdispersion

The quasi-poisson model

- ▶ Adds an additional parameter, ϕ , to the variance estimation \rightarrow *similar to robust standard errors*

$\Rightarrow \beta$ *remains the same, standard errors are larger*

The negative binomial model

The event is in fact generated by two processes

- ▶ $\lambda_i = \exp(\beta \times x_i + 1 \times u_i)$
- ▶ $v = \exp(u_i)$ is in itself generated by a gamma distribution $v_i \sim f\Gamma(\alpha)$
- ▶ The latent variable is manipulated directly: the rate increases over y

Excess zeros

Substantially that two data generating processes are at work.

- ▶ One producing zeros
- ▶ One producing (at least some) positive counts

⇒ *We can model this in two parallel regressions with possibly different x or just an additional intercept.*

Hurdle models

Observations have a higher hurdle/threshold/distance to pass in order to obtain a positive count (from 0 to 1) than between positive counts (1 to 2, 2 to 3, etc)

- ▶ Hurdle part: A binomial logit where success is $y > 0$
- ▶ Count model: A zero-truncated poisson (or negative binomial) on all the positive counts.

⇒ *Can accomodate under-dispersion too.*

Zero-inflated models

There are two sources of zeros, but only one of positive counts.

- ▶ Zero-inflated part: A binomial logit where success is the “always zeros”.
- ▶ Count model: A poisson or negative binomial that is not truncated.

⇒ *functions as a switch that is turned on/off after a threshold. The observation is then passed to the count-model group.*

Recap on GLMs

What are the criteria for model selection?

You can think of model selection as a set of criteria that should be met

Try out the model selection decision tree to see my mental map!

https://siljehermansen.github.io/teaching/choose_glm/