Models for Count Data II

Sean Hellingman

Regression for Applied Data Science (ADSC2020)

shellingman@tru.ca

Winter 2024

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- *Continuing along under the assumption of a count response variable.*

- Provided that the models pass the assumptions, we can compare models to be used for inferences.

- May also compare models based on prediction accuracy.
    - Do not need to pass assumptions but violations *may* cause problems with prediction accuracy.

Model Selection

- We have covered three different models for count data.
- In the presence of overdispersion, the quasi-Poisson and the negative binomial regression models can be used.
  - The quasi-Poisson model is not estimated using MLE so comparing models is a little more difficult.
- Poisson and negative binomial models may be directly compared

Likelihood Ratio Test

- ***Likelihood ratio tests*** *can be used to assess the goodness of fit of two competing statistical models.*
- As Poisson and negative binomial models are estimated with MLE, they can be compared.
- In R:
  - `library("lmtest")`
  - `lrtest(Model1,Model2)`
- The null hypothesis is that Model1 is *as good as or better* than Model2.
  - *Model2 is usually a more complex model*

Example 1

- Estimate the following models assuming a Poisson distribution and a negative binomial distribution (4 total models and do not forget the offset):
  - $ln(\frac{G}{TOI}) = 1 + S + Age + Pos$
  - $ln(\frac{G}{TOI}) = 1 + S + Age + Pos + BLK + HIT$

- Use the likelihood ratio test to compare the models
  1. Compare the models with the same distributions
  2. Compare the models with differing distributions

- Which model appears to be best?

Information Criteria

- We can use AIC and BIC to compare **Poisson and negative binomial** models.
- Remember, lower values of AIC and BIC are considered to be better.
- BIC has a higher penalization for the number of parameters than AIC.
- In R:
  - `AIC()`
  - `BIC()`

Example 2

- Of the four models estimated in Example 1, which is the best with regards to the AIC and BIC.
- Does this change your conclusions from Example 1?

Comments on Model Selection

- The quasi-Poisson models are much more difficult to compare.
    - *Lowest estimated deviance value*
- You can also use stepwise selection on GLMs including Poisson and negative binomial regression models.
- *All three may be compared based on prediction accuracy*

Predictions

- All three of the models may be used to make predictions.
  - `predict(CountModel, type = "response")`
- The *caret* package does not support the `glm.nb` objects, but it can be used for Poisson and quasi-Poisson.

Cross-Validation

- You can code any cross-validation you wish to do on your own and choose which measure(s) of accuracy you wish to use.
- Using R functions (that work for all three):
  - `library(boot)`
  - `CV.Model <- cv.glm(data, CountModel, K = folds)`
  - `CV.Model$delta`
- `delta` is MSE so if you want your results in the units of the response `sqrt(CV.Model$delta)` (RMSE).

Example 3

- Estimate the following models assuming a Poisson distribution, a quasi-Poisson distribution, and a negative binomial distribution (6 total models and do not forget the offset):
  - $ln(G) = 1 + S + Age + Pos$
  - $ln(G) = 1 + S + Age + Pos + BLK + HIT$

- Which of the six models is the best based on 10-fold cross-validation?

Zero-Inflated Data

- Sometimes count data contains more zero observations than would be expected for a specific distribution.

Zero-Inflated Data

- Sometimes count data contains more zero observations than would be expected for a specific distribution.
- Assuming overdispersion may help with zero-inflated data but there are other solutions.
- Two-component models called hurdle models can help with zero-inflated data.

Hurdle Models

- **Hurdle models** are two-component models:
  1. A truncated count component
  2. A hurdle component models zero vs. larger counts.

- More formally, the hurdle model combines a count data model $f_{\text{count}}(y; x, \beta)$ and a zero hurdle model $f_{\text{zero}}(y; z, \gamma)$:

$$f_{hurdle}(y; x, z, \beta, \gamma) = \begin{cases} f_{\text{zero}}(y; z, \gamma) & \text{if } y = 0 \\ (1 - f_{\text{zero}}(0; z, \gamma)) \cdot f_{\text{count}}(y; x, \beta)/(1 - f_{\text{count}}(0; x, \beta)) & \text{if } y > 0 \end{cases}$$

- $f_{\text{count}}(y; x, \beta)$ is left truncated at $y = 1$
- $f_{\text{zero}}(y; z, \gamma)$ is right-censored at $y = 1$

- *The count model is only employed if the hurdle for modeling the occurrence of zeros is exceeded.*

Negative Binomial Hurdle Model

- Combine a negative binomial count model with a logistic hurdle:

$$f(x;\mu,\theta) = \frac{f(x;\mu,\theta)}{P_{\mu,\theta}(Y>0)}, \quad y = 1, 2, ..., .$$

- Where μ and θ are the parameters found in the untruncated negative binomial distribution.
- $P_{\mu,\theta}(Y>0)$ indicates probability that $Y>0$ calculated with respect to the untruncated distribution.

- The hurdle and count components of the model are estimated separately.

Negative Binomial Hurdle Model in R

- To estimate negative binomial hurdle models in R:
  - `library(pscl)`
  - `HurdleModel <- hurdle(count.response ~ Var.1 + Var.2 + ... + Var.j, data = data, dist = "negbin", offset = log(unit.size))`
  - `summary(HurdleModel)`
- Any predictions made from the `HurdleModel` object will predict the counts.

Example 4

- Suppose that we are now interested in the number of game-winning goals (GW) players are scoring.
- Examine the GW variable for the presence of zero-inflated data.
- Estimate two negative binomial hurdle models and comment on the results.
  - Sets of explanatory variables:
    - $1 + S + Age + Pos$
    - $1 + S + Age + Pos + BLK + HIT$

Comments on Hurdle Models

- The usual model assumptions hold for both components.
- The interpretation of the coefficients is the same as the logistic model and the negative binomial model respectively.
- *There are other solutions for unbalanced data.*

Exercise 1

- Take some time to estimate some regression models with count response variables to make predictions and test prediction accuracy.
- Do you have zero-inflated data?
    - Estimate hurdle models to contend with the zero-inflated data.

References & Resources

1. De Jong, P., & Heller, G. Z. (2008). *Generalized linear models for insurance data*. Cambridge University Press.

2. Geyer, C. J. (2007). *Lower-truncated Poisson and negative binomial distributions.* University of Minnesota, MN.

- `glm()`
- `family()`
- `lrtest()`
- `AIC()`
- `cv.glm()`
- `hurdle()`
- `glm.nb()`