Models for Count Data I

Sean Hellingman ©

Regression for Applied Data Science (ADSC2020)

shellingman@tru.ca

Winter 2025

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- Generalized linear models (GLMs) can be used in more situations than linear regression
  - More flexibility in the response variable.
  - Relaxation of some of the assumptions required.

- One such case is when there is a count response variable.

Count Response Variables

- **Count variables** are by nature are non-negative integers.
  - *Discrete random variables.*
- Count response variables may be found in many fields:
  - Insurance
  - Sports
  - Ecology
- The *Poisson distribution* is a member of the exponential family and is often used to model count outcomes.

Review: Poisson Distribution I

- The **Poisson distribution** is a discrete distribution used to model the number of occurrences in some unit of measure.
- Examples:
  - Number of customers within an hour.
  - Number of baskets per minute in a basketball game.
  - Number of errors per line of R code.

Review: Poisson Distribution II

- There is no limit on the number of occurrences (X can be any non-negative integer).

- The PMF of the Poisson distribution is:

$$p(x) = \begin{cases} \frac{e^{-\lambda}\lambda^{x}}{x!}, & \text{for } x = 0, 1, \ldots \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

- Expected value: λ

- Variance: λ

Poisson Regression Model

- When constructing Poisson regression, the mean μ, is modeled in terms of explanatory variables:

$$g(\mu) = \boldsymbol{X}\beta \tag{2}$$

- Using the identity link function:

$$\mu = \boldsymbol{X}\beta$$

- Using the logarithmic (log) link function:

$$ln(\mu) = \boldsymbol{X}\beta \quad \text{AND} \quad \mu = e^{\boldsymbol{X}\beta}$$

- *The log link function guarantees positive values.*

Poisson Regression in R

- Estimating a Poisson Regression model in R is very similar to a linear regression model:
  - `PoiModel <- glm(count.response ~ Var.1 + Var.2 + ... + Var.j, family = poisson(link="log"), data = data)`

- Estimated using MLE.
- `link = "log"` is included by default.

Example 1

- Import the *NHL.txt* dataset into R.
  - Sourced from *www.hockey-reference.com* March 07^{th}, 2024
- Take a moment to get to know the data.
- Estimate the following Poisson regression model:
  $ln(G) = 1 + S + Age + Pos$
- What do the coefficients actually tell us?

Interpreting Coefficients

- Using the `summary()` function, **the coefficients express how a one unit change in the explanatory variable changes the log of the expected count** (response variable).

- *Poisson regression models the log of the expected count as a function of the explanatory variables.*

- Positive values indicate an increase in the expected count and negative values indicate a decrease in the expected count.

- May also examine *incident rate ratios* (see references: *Poisson Regression in R*)

Offset in Count Models

- Data are often collected from units of different sizes (t).
  - *Number of occurrences in some unit of measure.*
- Need to include these differences in the model.
  - Sometimes called *exposure* in insurance.

$$ln(\frac{\mu}{n}) = \beta_0 + \beta_1 X_1 + ...$$

- To include the offset in the model:

$$ln(\mu) = ln(n) + \beta_0 + \beta_1 X_1 + ...$$

Offset in Count Models in R

- The coefficient is set to be 1.
- The expected value then becomes proportional to the unit size (t)
- In R:
  - `PoiModel <- glm(count.response ~ Var.1 + Var.2 + ... + Var.j, family = poisson(link="log"), data = data, offset = log(unit.size))`

Example 2

- Correct your regression model from Example 1 to account for the time the players have been on the ice (TOI).
- What do the coefficients actually tell us?

Assumptions

- We have relaxed the normality of the response and homoscedasticity assumptions.
- **The following assumptions still apply for Poisson regression:**
  1. Count response variable
  2. There is a linear relationship between the continuous predictor variables and the natural logarithm of the dependent variable.
  3. There is **no** multicollinearity of the explanatory variables.
  4. **The variability is equal to the mean**

Linearity IA

- We can use visualizations to verify this assumption.
- Directly plot the relationships of numeric variables (logit of the response vs explanatory variables) **after a model has been estimated**:
  - `counts <- predict(PoiModel, type = "response")`
  - `mydata <- data %>%`
  - `dplyr::select_if(is.numeric)`
  - `predictors <- colnames(mydata)`
  - `mydata <- mydata %>%`
  - `mutate(lncounts = log(counts)) %>%`
  - `gather(key = "predictors", value = "predictor.value", -lncounts)`

Linearity IB

- Create the scatterplots:
  - `ggplot(mydata, aes(lncounts, predictor.value))+`
  - `geom_point(size = 0.5, alpha = 0.5) +`
  - `geom_smooth(method = "loess") +`
  - `theme_bw() +`
  - `facet_wrap(~predictors, scales = "free_y")`
- If the individual plots show an approximately linear relationship, we can say the model passes the linearity assumption.

Linearity Solutions

- If the linearity assumption is violated:
  - Try to transform explanatory variables to create a linear relationship (polynomials).
  - May be able to use regression splines.

Multicollinearity

- The no multicolinearity assumption can be checked using the `vif()` function from the *car* package.

- Recall: *If the value is larger than 5 or 10 we should consider removing one or more of the variables.*
- Examine the `GVIF^(1/(2*Df))` when there are 2 or more degrees of freedom.
    - *Square this value.*

Example 3

- Check the linearity assumption for the model from Example 2.
- Check the model estimated in Example 2 for the presence of multicolinearity.

Outliers

- **Regression outliers** are those observations whose values (of the response and explanatory variables) deviate from the regression relationship which holds for the majority of observations.

- Cook's distance may be used to examine Poisson regression models for potential outliers (values over 0.5 and 1.0).

- In R:
  - Plots:
    `plot(PoiModel,3)` AND `plot(PoiModel,4)`
  - To get the numeric values:
    `cooks.distance(PoiModel)`

Example 4

- Check the model estimated in Example 2 for the presence of outliers.

Equidispersion

- Recall: *The mean and the variance of the Poisson distribution are assumed to be the same.*
  - $X \sim Po(\lambda) \Rightarrow E[X] = Var(X) = \lambda$
- **This assumption may not hold in many cases.**
- **Overdispersion** occurs when the variance is actually larger than the mean.

Exponential Family

- It is now assumed that the response follows a distribution from the *natural exponential family.*
  - Not the same as the exponential distribution.

- Density:

$$f_\theta(y) = exp[\{y\theta - b(\theta)\}/a(\phi) + c(y, \phi)] \quad (3)$$

  - ϕ: dispersion parameter
  - θ: canonical parameter (function of β)
  - a,b,c: functions

Exponential Family Variance Functions

Distribution	$\mathrm{E}(y)$	$V(\mu) = \frac{\mathrm{Var}(y)}{\phi}$
Binomial(n, π)	$n\pi$	$n\pi(1-\pi)$
Poisson(μ)	μ	μ
Normal(μ, σ^2)	μ	1
Gamma(μ, ν)	μ	μ^2
Inverse Gaussian(μ, σ^2)	μ	μ^3
Negative Binomial(μ, κ)	μ	$\mu(1+\kappa\mu)$

Overdispersion

- We can use R to check for overdispersion:
  - `library(AER)`

  - `dispersiontest(PoiModel,trafo=1) #linear specification`
  - `dispersiontest(PoiModel,trafo=2) #quadratic specification`
- trafo = transformation function

- trafo=1 $\Rightarrow$ Quasi-Poisson
- trafo=2 $\Rightarrow$ Negative binomial

Example 5

- Check the model estimated in Example 2 for the presence of overdispersion.

Quasi-Poisson in R

- To estimate the Poisson model with overdispersion, *quasi-likelihood* estimation methods are used.
- To estimate using R:
    - `QPoiModel <- glm(count.response ~ Var.1 + Var.2 + ... + Var.j, family = quasipoisson(), data = data, offset = log(unit.size))`

Negative Binomial Distribution

- Negative binomial distribution which may arise as a gamma mixture of Poisson distributions.
- One way of expressing the negative binomial probability mass function is:

$$f(y;\mu,\theta) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)\cdot y!}\cdot\frac{\mu^y\cdot\theta^\theta}{(\mu+\theta)^{y+\theta}} \tag{4}$$

- with mean μ and shape parameter θ.
- Estimated using the maximum likelihood estimation methodology.

Negative Binomial Distribution in R

- Estimate a negative binomial regression model in R:
  - `library(MASS)`
  - `NBModel <- glm.nb(count.response ~ Var.1 + Var.2 + ... + Var.j + offset(log(unit.size)), link = "log", data = data)`

Example 6

- Estimate a quasi-Poisson and a negative binomial regression model to improve the model from Example 2.

- What do these coefficients mean?

Comments on Count Models

- Usually begin with a Poisson regression model.
- Check the model for presence of overdispersion.
    - There are other methods you may use to check for overdispersion.
- If over dispersion exists, select an appropriate model.
    - Quasi-Poisson
    - Negative Binomial

Exercise 1

- Take some time to estimate some regression models with count response variables, run appropriate diagnostics, and use Quasi-Poisson or Negative Binomial as needed.

References & Resources

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

- `glm()`
- `family()`
- Poisson Regression in R
- Poisson Regression
- `dispersiontest()`
- `glm.nb()`