Models for unbounded count data

Bert van der Veen

Department of Mathematical Sciences, NTNU

Outline

- Models for count data
- Residual diagnostics in GLMs
- Other useful models

Questions about yesterday?

The binomial GLM

Response data: r the number of successes in N trials
Predictor variables: x_i albeit continuous and/or categorical
Parameters: probability of success p_i in trial i
Goal: estimate pi_i for each observation

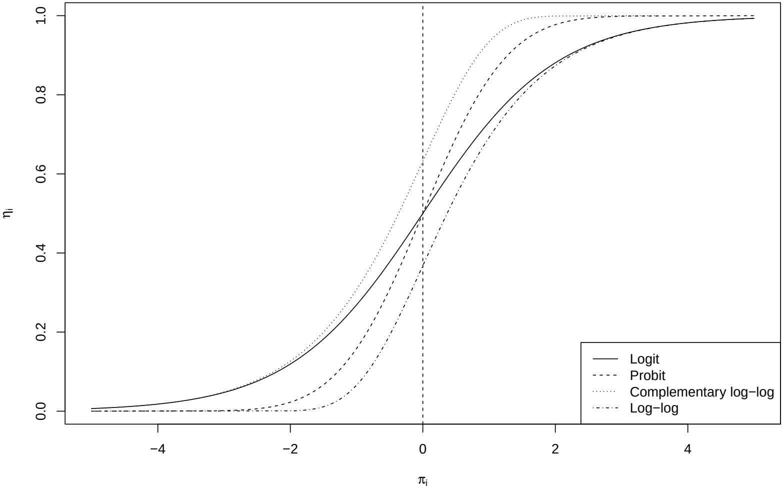
Binomial GLM use

- When a linear regression is not appropriate :)
- For binary data or counts of successes/failures

Common examples

- OME signal identification
- cancer rates
- Predicting species' distributions
- Number of germinated plant seeds
- Prevalence of disease in a population
- Probability of observing a behavior
- Proportion of orchids 🙃

Binomial link functions (2)



Typical count cases

- Number of caught fish
- Number of deaths due to lung cancer or other diseases
- Seizure counts
- Times a behavior is expressed
- Number of pidgeons in a city
- Number of Bigfoot reports
- Number of wrongful convictions
- Number of stars in the night sky

The Poisson GLM

Response data: k_i the count
Predictor variables: x_i albeit continuous and/or categorical
Parameters: mean λ
Goal: estimate λ_i for each observation

The Poisson distribution

$$\mathcal{L}(y_i;\Theta) = \mathsf{exp}\{y_i \log(\lambda_i) - \lambda_i - \log(y_i!)\} \quad (1)$$

The Poisson paramater λ is the mean of the counting process

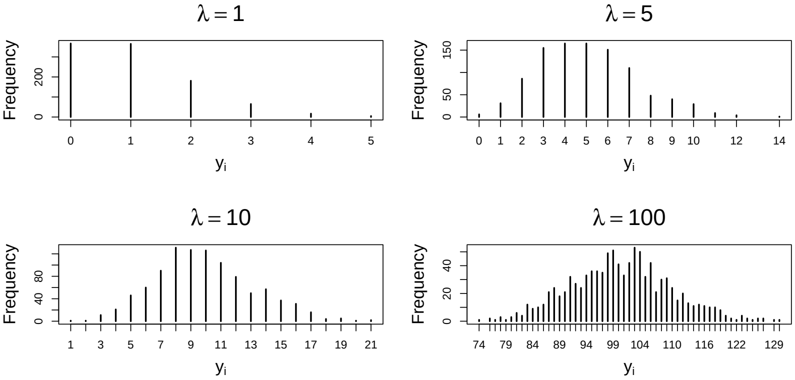
Is Poisson regression in the EF?

$$\mathcal{L}(y_i;\Theta) = \text{exp}\left\{\frac{y_i \log(\lambda) + \log(\lambda)}{1} + \log(y_i!)\right\} \tag{2}$$

All GLMs can be formulated as:

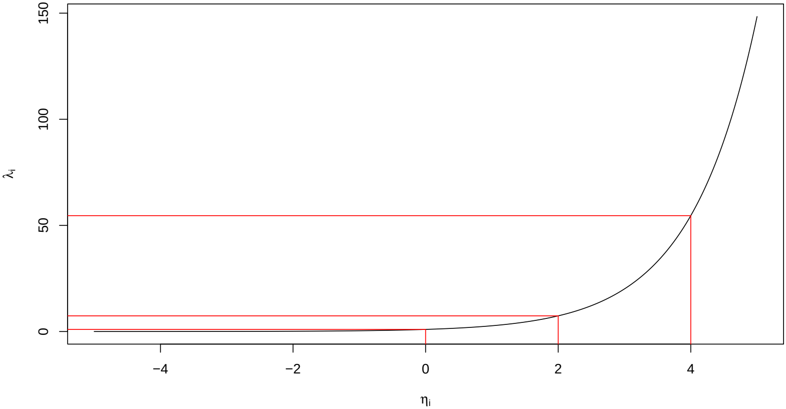
$$\mathcal{L}(y_i;\Theta) = \text{exp}\{\frac{y_i \eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi)\} \tag{3}$$

The Poisson distribution visually



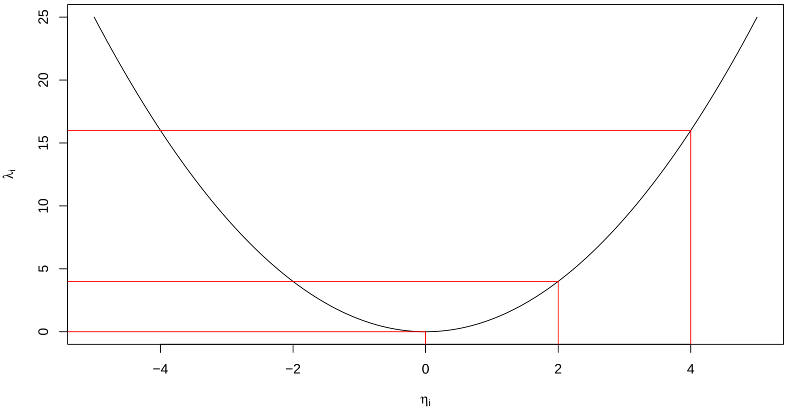
Log-link function

So log is the canonical link. This looks like:



square root-link function

An alternative is the square root-link function $\lambda_i = (\alpha + x_i\beta)^2$



Poisson assumptions

- An event can occur $0 \dots \infty$ times
- Events are independent
- Events cannot occur simultaneously
- Variance equals the mean
- The rate of events is constant

The rate of events

Counts are usually collected over time or space:

- The amount of fish we catch in an hour
- The number of deaths on a population of 10.000
- The number of seizures a patient has during the night
- The number of times a behavior is expressed during a treatment
- Number of pidgeons in a city the size of Los Angeles
- The number of bigfoot reports collected in a small forest, yesterday, by 3 people
- Number of wrongful convictions in Germany last year
- Number of starts in the night sky

The Poisson distribution: rates

Alternatively we can write:

$$\mathcal{L}(y_i; \Theta) = \text{exp}\{y_i \log(rt) - rt - \log(y_i!)\} \tag{4}$$

so, $\lambda = rt$

- r is the rate at which counts occur, per time period t
- we can instead write $\lambda = t\exp(\eta) = \exp\{\eta + \log(t)\}$
- $\log(t)$ is called an **offset**

Example: going out fishing

- On average we catch 1 fish in 20 minutes
  $\lambda = 1 = \exp\{-2.99 + \log(20)\}$
- If we go fishing for an hour we catch $\exp(-2.99) * 60 = 3$ fish
- If we go fishing for one minute we catch $\exp(-2.99) = \frac{1}{20}$ fish
- Here, $r = \exp(-2.99)$ and t is the time we want to spend fishing
- We can also find the amount of time we need to spend to catch 5 fish
  - $\exp\{-2.99 + \log(t)\} = 5$, so $t = \frac{5}{\exp(-2.99)} = 100$ minutes

Log-linear regression

Log-linear regression is a class of models that uses the log-link function:

$$\begin{aligned} \log\{\mathbb{E}(y_i|x_i)\} = \eta_i = \alpha + x_i\beta \\ \mathbb{E}(y_i|x_i) = \lambda_i = \exp(\alpha + x_i\beta) \end{aligned} \tag{5}$$

Log-linear regression is commonly used to analyse count data

Log-linear regression

Log-linear regression is a "multiplicative" model

$$\begin{aligned} \lambda_i &= \exp(\alpha + x_i\beta) \\ &= \exp(\alpha)\exp(x_i\beta) \end{aligned} \tag{6}$$

Log-linear regression

Log-linear regression is a “multiplicative” model

$$\begin{aligned} \lambda_i &= \exp(\alpha + x_i\beta) \\ &= \exp(\alpha)\exp(x_i\beta) \end{aligned} \tag{6}$$

A unit increase in x_i scales λ_i by $\exp(\beta)$

So, when $\exp(\beta) = \frac{1}{2}$, $\exp(\alpha)$ halves for every unit of x_i
So, when $\exp(\beta) = 2$, $\exp(\alpha)$ doubles for every unit of x_i

Log-linear regression

Log-linear regression is a “multiplicative” model

$$\begin{aligned} \lambda_i &= \exp(\alpha + x_i\beta) \\ &= \exp(\alpha)\exp(x_i\beta) \end{aligned} \tag{6}$$

A unit increase in x_i scales λ_i by $\exp(\beta)$

So, when $\exp(\beta) = \frac{1}{2}$, $\exp(\alpha)$ halves for every unit of x_i
So, when $\exp(\beta) = 2$, $\exp(\alpha)$ doubles for every unit of x_i

Of course, this is more involved with multiple predictors

Example of a multiplicative process

Say that we have the model:

$$\log(\lambda) = \alpha + x_i\beta \tag{7}$$

- with $\alpha = -2.99$ and $\beta = \log(2) \approx 0.693$
- x_i is either 0 or 1: either I was fishing or you were
- $\exp(-2.99) = 0.05$ the average number of fish I caught in the time I spent fishing
- $\exp(-2.99 + \log 2) = \exp(-2.99) * 2 = 0.1$ the average number of fish you caught
- So, you caught twice as many fish! I am not very good at fishing

Example: campus crime

Count of violent crimes for an academic year

Figure 1: freepik.com

Figure 2: campussecuritytoday.com

Campus crime: the data

Data via Legler and Roback (2021)

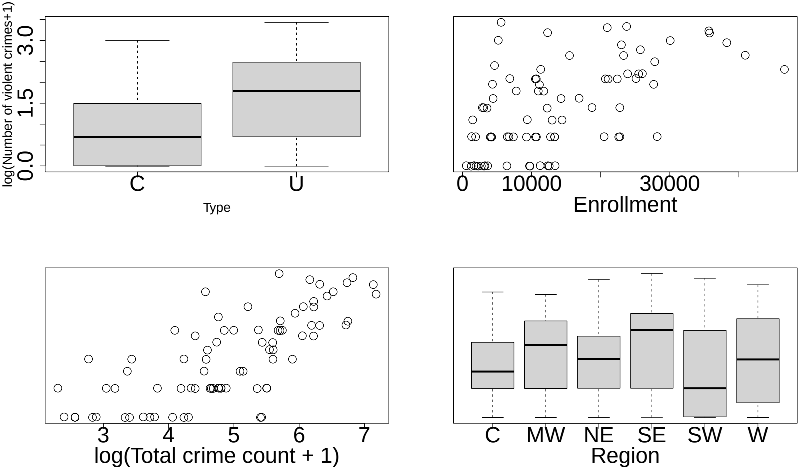
- 81 observations
- Number of violent crimes, Total number of crimes, Number of property crimes
- Student enrollment
- Type (University or College)
- Region of the country

num_viol	total_crime	num_prop	Enrollment	type	region
30	296	266	5590	U	SE
0	10	10	540	C	SE
23	1256	1233	35747	U	W
1	211	210	28176	C	W
1	117	116	10568	U	SW
0	29	29	3127	U	SW
7	291	284	20675	U	W

How could these data be analysed with a binomial regression?

num_viol	total_crime	num_prop	Enrollment	type	region
30	296	266	5590	U	SE
0	10	10	540	C	SE
23	1256	1233	35747	U	W
1	211	210	28176	C	W
1	117	116	10568	U	SW
0	29	29	3127	U	SW
7	291	284	20675	U	W

What is the relationship between violent crimes and school variables?

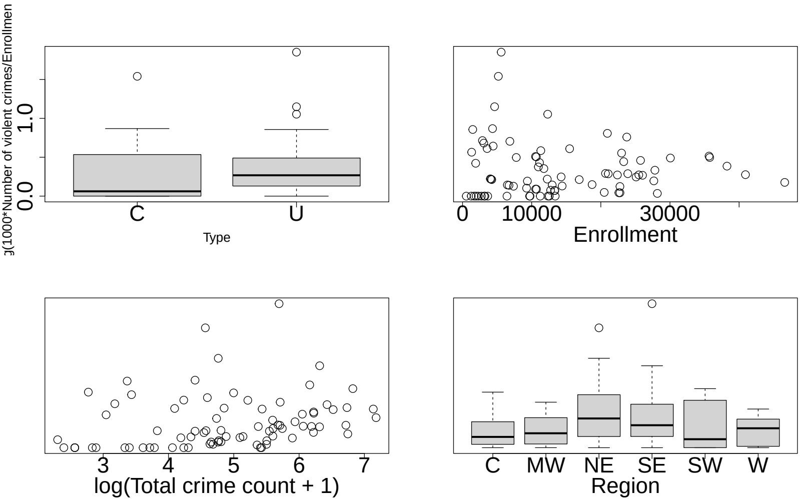


Campus crime: fit the model

```r
model <- glm(num_viol ~ region+type,
             family = "poisson", data = campus)
```

What issue can we identify for this model?

What is the relationship between violent crimes per 1000 enrolled and school variables?



Campus crime: fit the model

```r
modelo <- glm(num_viol ~ region + type
              + offset(log(Enrollment)),
            family = "poisson", data = campus)
```

```
##              No.offset Offset
## (Intercept)       0.34 -8.509
## regionMW          0.49  0.099
## regionNE          0.52  0.781
## regionSE          0.87  0.877
## regionSW          0.25  0.503
## regionW           0.74  0.273
## typeU             1.24  0.340
```

Campus crime: per 1000 enrolled

```
model1000 <- glm(num_viol ~ region + type
                  + offset(log(Enrollment/1000)),
                 family = "poisson", data = campus)
```

$$\exp(\alpha_2) * 1000 = \exp(\alpha_3) \; \exp(\alpha_2) = \exp(\alpha_3)/1000$$
$$\exp(\alpha_2) = \exp\{\alpha_3 - \log(1000)\}$$

Campus crime: per 1000 enrolled

```
##               No.offset Offset Per thousand
## (Intercept)        0.34 -8.509       -1.602
## regionMW           0.49  0.099        0.099
## regionNE           0.52  0.781        0.781
## regionSE           0.87  0.877        0.877
## regionSW           0.25  0.503        0.503
## regionW            0.74  0.273        0.273
## typeU              1.24  0.340        0.340
```

Do we have a good model?

More on this after the break

- Overdispersion or underdispersion
- Zero-inflation

Overdispersion

Our assumption: $\lambda = \text{var}(\mathbf{y})$
Reality: $\lambda \geq \text{var}(\mathbf{y})$

- Mean $=$ variance
- If there is more variation, this assumption fails
- Consequences: CIs underestimate, biased parameter estimates, inflation in model selection

For our example: many females have few satellites, but some females have very many.

Underdispersion

Our assumption: $\lambda = \text{var}(\mathbf{y})$
Reality: $\lambda \leq \text{var}(\mathbf{y})$

Considerably less common than overdispersion.

Detecting overdispersion

- Residual diagnostics
- $D(\mathbf{y}; \hat{\boldsymbol{\mu}})/(n-k)$: should be close to 1
- `performance::check_overdispersion` (relies on asymptotics)
- Simulation (later today)

Dealing with dispersion: options

- Correct for it (calculate dispersion)
- Fit a different model
    - Negative binomial (overdispersion, `MASS package`)
    - Conway-Maxwell Poisson (over- and underdispersion.)
    - Generalized Poisson(over- and underdispersion)
    - Quasi-likelihood models
    - Mixed models (not covered here)

Quasi-likelihood models

Introduced by Wedderburn (1974)

- No “real” likelihood is specified for the data
- Means no AIC, but deviance exists
- Largely defined by its variance function

For Poisson responses: does not correct the parameter estimates

Negative-binomial

$$\mathcal{L}(y_i; \Theta) = \frac{\Gamma(y_i + \phi)}{\Gamma(\phi) y_i!} \left(\frac{\phi}{\mu_i + \phi} \right)^{\phi} \left(\frac{\mu_i}{\mu_i + \phi} \right)^{y_i} \tag{8}$$

- $\text{var}(\mathbf{y}) = \boldsymbol{\mu} + \frac{\boldsymbol{\mu}^2}{\phi}$
- For large ϕ Poisson!
- Requires more data/information due to extra parameter

Is negative-binomial regression in the EF?

$$\mathcal{L}(y_i;\Theta) = \exp[\frac{y_i \log\{\frac{\mu_i}{\mu_i+\phi_i}\} - \phi \log\{\frac{\mu_i+\phi}{\phi}\}}{1} + \log\{\Gamma(y_i+\phi)\} - \log\{\Gamma(\phi)\} - \log(y_i!)] \tag{9}$$

All GLMs can be formulated as:

$$\mathcal{L}(y_i;\Theta) = \text{exp}\{\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i,\phi)\} \tag{10}$$

Example: hurricanes

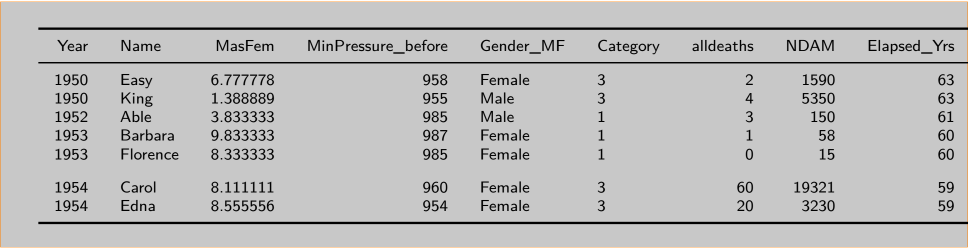
Deaths due to hurricanes

Figure 3: climate reality project

Hurricanes: the data

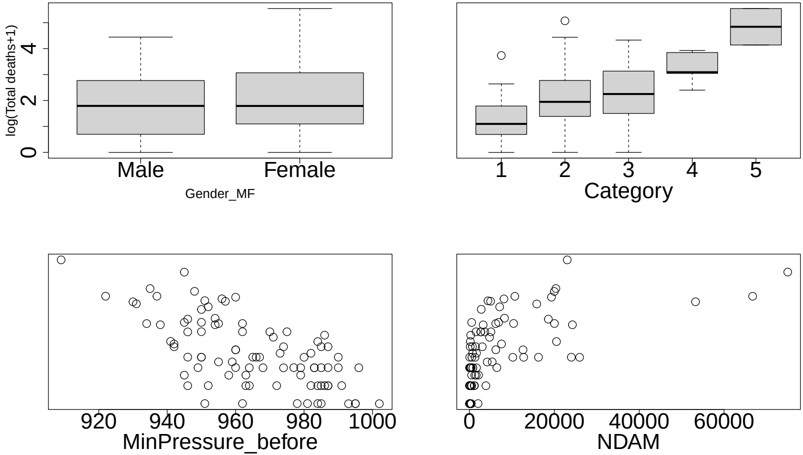
Data from Jung et al. (2914)

- 94 observations
- Year, name, Binary name categorization (male, female), Masulinity-Femininity score, strength of the hurricane, prior air pressure
- Excluded two outliers



Year	Name	MasFem	MinPressure_before	Gender_MF	Category	alldeaths	NDAM	Elapsed_Yrs
1950	Easy	6.777778	958	Female	3	2	1590	63
1950	King	1.388889	955	Male	3	4	5350	63
1952	Able	3.833333	985	Male	1	3	150	61
1953	Barbara	9.833333	987	Female	1	1	58	60
1953	Florence	8.333333	985	Female	1	0	15	60
1954	Carol	8.111111	960	Female	3	60	19321	59
1954	Edna	8.555556	954	Female	3	20	3230	59

Are female hurricanes more deadline than male hurricanes?



Hurricanes: fit the model

```
modelp <- glm(alldeaths ~ Gender_MF + MinPressure_before,
             family = "poisson", data = hurricanes)
```

Hurricanes: interpreting parameters

```
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)             55.000     1.2000    46.0  0.0e+00
## Gender_MFFemale          0.270     0.0570     4.8  1.7e-06
## MinPressure_before      -0.055     0.0013   -43.0  0.0e+00
```

How do we interpret the intercept?

Hurricanes: interpreting parameters

```
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)             2.100     0.0550    38.0  0.0e+00
## Gender_MFFemale         0.270     0.0570     4.8  1.7e-06
## MinPressure_beforeC    -0.055     0.0013   -43.0  0.0e+00
```

Average prior air pressure: 964.49 knots

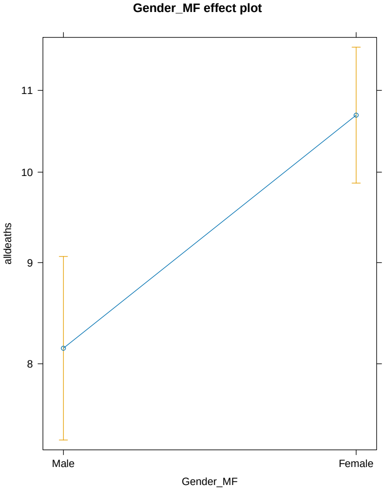
How do we interpret the intercept?
And its standard error?

Prior air pressure centered

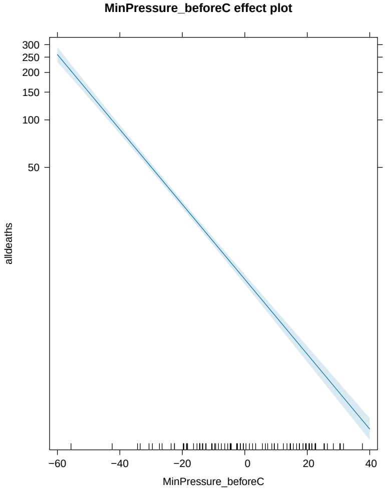
Hurricanes: interpreting parameters on the **response scale**

- (Intercept) = Male-named hurricanes, prior air pressure ≈ 965
- Female-named hurricanes: $\exp(0.27) = 1.3$, so 30% more deadly

Hurricanes: visual interpretation



Gender_MF effect plot



MinPressure_beforeC effect plot

Hurricanes: checking overdispersion

```
performance::check_overdispersion(modelp1)
```

```
## # Overdispersion test
##
##        dispersion ratio =   18.974
##   Pearson's Chi-Squared = 1650.771
##                 p-value =  < 0.001
```

```
## Overdispersion detected.
```

Hurricanes: Negative-binomial

```
modelnb <- MASS::glm.nb(alldeaths ~ Gender_MF + MinPressure_beforeC
```

and compare the models:

```
AIC(modelp1, modelnb)
```

```
##          df       AIC
## modelp1   3 1639.9581
## modelnb   4  612.1051
```

Hurricanes crabs: comparing estimates

```
##                      Poisson estimate NB estimate Poisson SE  NB SE
## (Intercept)                     2.100       2.200     0.0550 0.2000
## Gender_MFFemale                 0.270       0.047     0.0570 0.2500
## MinPressure_beforeC            -0.055      -0.056     0.0013 0.0061
```

- SEs have increased
- Coefficients have changed
- **Female and male-named hurricanes are equally deadly**
- Effect of pressure has remained

Count distributions

- Poisson
- Negative binomial (two types, with dispersion)
- Conway-Maxwell Poisson (with dispersion)
- Generalized Poisson (with dispersion)
- Skellam distribution (difference of counts)
- Binomial distribution (counts with a maximum)
- Truncated distributions (e.g., without zeros)
- Quasi-likelihood models

Summary

- Counts are analysed with log-linear models
- The collection effort of counts needs to be considered (offset)
- When the Poisson assumption is violated, we change to another count distribution