

Part II: Generalized Linear Models

Chapter II.6

Modeling Counts

▶ To be discussed...

- ▶ Poisson GLMs
- ▶ Modeling Rates
- ▶ Overdispersion for Poisson GLMs
- ▶ Negative Binomial GLMs
- ▶ GLMs for Zero Inflated Discrete Distributions


6. Modeling Counts

The Poisson distributions are a discrete family with probability mass function (pmf) indexed by the rate parameter $\mu > 0$, with sample space \mathbb{N}_0 (s. Remark II.1.3).

A r.v. $Y \sim \mathcal{P}(\mu)$ has pmf

$$p(y) = \mathbb{P}(Y = y) = \frac{e^{-\mu} \mu^y}{y!}, \quad y = 0, 1, 2, \dots$$

➤ II.6.1 Remark (Poisson distribution)

- It is useful for modeling 'count data', i.e. count of events that occur randomly over time or space at a particular rate, when outcomes in disjoint spaces are independent.
- The Poisson distribution is a member of the *exponential dispersion family* with natural parameter $\theta = \log(\mu)$, $b(\theta) = \exp(\theta)$, $\psi/\omega = 1$ (see Remarks II.2.5 and II.2.8).
 Thus $\mathbb{E}(Y) = b'(\theta) = \exp(\theta) = \mu$ and $\text{Var}(Y) = \frac{\psi}{\omega} b''(\theta) = \mu$.
- The Poisson distribution is unimodal and skewed, with skewness given by $\mathbb{E}(Y - \mu)^3 / \sigma^2 = 1/\sqrt{\mu}$.
- As μ increases, the Poisson distribution grows more symmetric and is eventually well approximated by a normal distribution.

▶ II.6.2 Remark (Variance Stabilization)

Consider n independent Poisson random variables Y_1, \dots, Y_n , with $E(Y_i) = \mu_i$ and $\text{Var}(Y_i) = \mu_i$. In order to model count data, the counts could be transformed in order to approximately stabilize the variance, so that ordinary least squares methods (for constant variance) can be applied (s. Example II.2.12).


By the delta method, we have


$$g(Y) - g(\mu) \approx (Y - \mu)g'(\mu), \text{ which leads to } \text{Var}[g(Y)] \approx [g'(\mu)]^2 \text{Var}(Y).$$

For $Y_i \sim \mathcal{P}(\mu_i)$ and $g(Y_i) = \sqrt{Y_i}$ it holds

$$\text{Var}(\sqrt{Y_i}) \approx \left(\frac{1}{2\sqrt{\mu_i}} \right)^2 \mu_i = \frac{1}{4}.$$

This approximation is better for larger μ_i ($\sqrt{Y_i}$ s is more closely linear in a neighborhood of μ_i).

 Hence $\sqrt{Y_1}, \dots, \sqrt{Y_n}$ could be modeled by linear models and ordinary least squares.

However we model $E(\sqrt{Y_i})$ and not $E(Y_i)$  GLM is a more adequate approach.

II.6.3 Poisson GLM

Consider a GLM for independent Poisson responses Y_i , $i = 1, \dots, n$ with *canonical link* $\eta_i = g(E(Y_i)) = g(\mu_i) = \log(\mu_i)$ and linear predictor $\eta_i = \sum_{k=1}^p \beta_k x_{ik}$. In matrix form, the Poisson GLM is defined as

$$\boldsymbol{\eta} = \log(\boldsymbol{\mu}) = \log[E(\mathbf{Y})] = \mathbf{X}\boldsymbol{\beta} ,$$

which in terms of the means (μ_i) is expressed by

$$\mu_i = \exp \left(\sum_{k=1}^p \beta_k x_{ik} \right) = (e^{\beta_1})^{x_{i1}} \dots (e^{\beta_p})^{x_{ip}} ,$$

i.e. for the i -th item and the j -th explanatory variable, the mean at $x_{ij} + 1$ equals e^{β_j} times the mean at x_{ij} , when all the other explanatory variables X_k , $k \neq j$, are kept fixed.

➤ II.6.4 Remark

The likelihood equations are (see Remark II.2.20):

$$\sum_{i=1}^n x_{ik} [y_i - E(Y_i)] = 0, \quad k = 1, \dots, p.$$

👉 Here we consider *Poisson regression* models (existence of continuous explanatory variables).

II.6.5 Remark (estimated covariance matrix)

The estimated covariance matrix of $\hat{\beta}$ is $\widehat{\text{Var}}(\hat{\beta}) = (\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}$, with \mathbf{W} a diagonal matrix with $w_{ii} = (\partial\mu_i/\partial\eta_i)^2/\text{Var}(Y_i) = \mu_i$ (s. Theorem II.2.22).


II.6.6 Remark (Goodness of Fit - Remark II.2.37)

- ① For Poisson GLMs, the deviance equals the *likelihood ratio statistic* (LRS) G^2 for testing model \mathcal{M} (H_0) against the saturated model (H_1):

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = G^2(\mathcal{M}) = 2 \sum_{i=1}^n y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right).$$

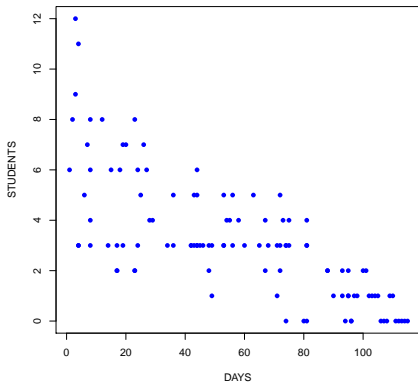
- ② For Poisson GLMs, the corresponding *score statistic* becomes the *Pearson's X^2* statistic
- $$X^2(\mathcal{M}) = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}, \text{ since } v(\mu_i) = \mu_i.$$

These statistics are asymptotically chi-squared distributed and can thus be used for goodness of fit testing, when the number n of Poisson observations is fixed and their means increase unboundedly. This holds for contingency tables with a fixed number of cells and large sample size (as we already know).

 It is more informative to test a model by comparing it to more complex models (e.g., with interaction terms or models not assuming $\text{Var}(Y_i) = \mu_i$) and investigate thus the lack of fit.

➤ II.6.7 Example (infectious disease)

The data set consists of counts of high school students diagnosed with an infectious disease within a period of days from an initial outbreak
(files : 'infectious_disease.dat', 'Poisson_glm_examples_R.pdf').



Example (continues)


```
Call:
glm(formula = Students ~ Days, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max 
-2.00482  -0.85719  -0.09331   0.63969   1.73696 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.990235    0.083935   23.71  <2e-16 ***
Days        -0.017463    0.001727  -10.11  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 215.36 on 108 degrees of freedom
Residual deviance: 101.17 on 107 degrees of freedom
AIC: 393.11
Number of Fisher Scoring iterations: 5
```

 The coefficient for Days is highly significant and negative, indicating that the mean number of diseased students decreases in days from outbreak: every day that passes, the mean number of diseased students decreases by $e^{-0.017} = 0.983$ students.

Modeling Rates

> II.6.8 Remark

The response count Y_i may correspond to number of occurrences of an event over a fixed amount t_i of time, spatial area or population size. For example modeling counts Y_i of patients suffering from a specific disease for various cities, t_i will be the population size of the i -th city. Then the random sample rate is Y_i/t_i , with expected value μ_i/t_i .

Given explanatory variables, a log-linear model for the expected rate has the form

$$\log\left(\frac{\mu_i}{t_i}\right) = \sum_{k=1}^p \beta_k x_{ik} ,$$

which can be written as

$$\log(\mu_i) = \log(t_i) + \sum_{k=1}^p \beta_k x_{ik} .$$

Thus, the fit corresponds to using $\log(t_i)$ as an explanatory variable in the GLM for $\log(\mu_i)$ and forcing its coefficient to equal 1. This adjustment term $\log(t_i)$ is called an **offset**.

In this case, the expected response count is given by

$$\mu_i = t_i \exp\left(\sum_{k=1}^p \beta_k x_{ik}\right) .$$

► II.6.9 Example (lung cancer survival)

Number of deaths for 539 males having lung cancer ^a
(files: 'lung_cancer.dat', 'Poisson_glm_examples_R.pdf').

The prognostic factors are histology (H) and stage (S) of disease, with observations grouped into 2-month intervals of follow-up after the diagnosis [time (T): 1 (0-2m), 2 (2-4m), 3 (4-6m), 4 (6-8m), 5 (8-10m), 6 (10-12m) and 7 ($\geq 12m$)]. For every combination of a particular length of follow-up, histology, and stage of disease, the number of deaths (count) and the number of months of observations of subjects still alive during that follow-up interval (risktime).

If μ_{ijk} denote the expected number of deaths and t_{ijk} the total time at risk for histology i and stage of disease j , in follow-up time interval k , the Poisson GLM for the death rate,

$$\log(\mu_{ijk}/t_{ijk}) = \beta_0 + \beta_i^H + \beta_j^S + \beta_k^T,$$

treats each explanatory variable as a qualitative factor.

The estimated stage-of-disease effects are $(\hat{\beta}_1^S, \hat{\beta}_2^S, \hat{\beta}_3^S) = (0, 0.47, 1.32)$. Thus, the estimated death rate at the third stage of disease is $\exp(1.32) = 3.76$ times that at the first stage, adjusting for follow-up time and histology.

The corresponding Wald 95% confidence interval $\exp(1.32 \pm 1.96 \cdot 0.152) = (2.79, 5.06)$.

^aAgresti (2015). *Foundations of Linear and Generalized Linear Models*, p. 233-235

Overdispersion for Poisson GLMs

It is often the case that the variability of count data is greater than that of a Poisson r.v., i.e. $\text{Var}(Y_i) > \text{E}(Y_i)$. This phenomenon is called **overdispersion**.

II.6.10 Remark

A distribution, often used as an alternative to the Poisson in the presence of *overdispersion*, is the Negative Binomial $\mathcal{NB}(k, \lambda)$.

If $Y \sim \mathcal{NB}(k, \lambda)$, with $k > 0$ and $\lambda > 0$, then

$$P(Y = y) = \frac{\Gamma(k + y)}{\Gamma(k)} \cdot \frac{\lambda^k (1 - \lambda)^y}{y!}, \quad y = 0, 1, 2, \dots$$

(equivalent expression to the pmf provided in Remark II.1.3).

- The expected value of Y is $\mu = \text{E}(Y) = \frac{k(1-\lambda)}{\lambda}$ while its variance $\sigma^2 = \text{Var}(Y) = \frac{k(1-\lambda)}{\lambda^2} = \mu + \delta\mu^2 > \mu$, where $\delta = 1/k > 0$.
- As $\delta \rightarrow 0$, $\sigma^2 \rightarrow \mu$ and the negative binomial converges to the Poisson distribution.

Negative Binomial GLMs

> II.6.11 Assumption

The dispersion parameter δ is assumed to be constant (unknown) for all n observations (like the variance in normal models).

> II.6.12 MLE for NB GLMs

It can be shown that for a NB GLM with link function g , the likelihood equations are (see II.2.16):

$$\sum_{i=1}^n \frac{x_{ik}(y_i - \mu_i)}{\text{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) = \sum_{i=1}^n \frac{x_{ik}(y_i - \mu_i)}{\mu_i + \delta \mu_i^2} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) = 0, \quad k = 1, \dots, p.$$

Furthermore, $\hat{\beta}$ and $\hat{\delta}$ are asymptotically independent (derive the Hessian matrix and verify that $E(\partial^2 L / \partial \beta_k \partial \delta) = 0$, for all k). Thus, the large-sample SE for $\hat{\beta}_k$ is the same for the cases of known and estimated δ .

For the covariance matrix of $\hat{\beta}$, the entries of the diagonal matrix \mathbf{W} are

$$w_i = \frac{(\partial \mu_i / \partial \eta_i)^2}{\text{Var}(Y_i)} = \mu_i / (1 + \delta \mu_i).$$

Negative Binomial GLMs


II.6.13 NB GLMs in R

Very often, the log link is used for NB GLMs (as in Poisson GLMs) rather than the canonical link. In R, NB GLMs with log-link are easily fitted by the `glm.nb()` function of the MASS package.


II.6.14 Example (horseshoe crab data)

We shall reconsider the crabs data set of Example II.3.25 and focus on the number of satellites (SAT) a female crab has.

The satellite counts are modeled as Poisson and negative binomial (NB) random variables.

 See: 'P_NB_ZI_(crabs)_R.pdf'

Fitting first the Poisson and NB GLMs having just the intercept, i.e assuming that the number of satellites per female crab are iid distributed (Poisson or NB), we see that the expected number of satellites is estimated to be $\hat{\mu} = \exp(\hat{\beta}_1) = \exp(1.07) = 2.9$ (under both models).

 Under the NB assumption we get $\hat{\delta} = 1/\hat{\theta} = 1.32$, indicating the presence of overdispersion ($\hat{\theta} = 0.758$ with std.error 0.126).

Example (continues)

We consider next the Poisson and NB GLMs (with log-link) modeling SAT in terms of the weight of the female crabs. The weight is expressed in the model in Kgr (`weight=WT/1000`).

- 👉 Implementing the associated R-code provided in 'P_NB_ZI_(crabs)_R.txt', we verify that the deviance for the Poisson and NB GLM is equal to 560.87 and 196.16, respectively, providing strong evidence that the NB GLM is of better fit.
- 👉 The parameter estimates, under each model, are $\hat{\beta}_{\mathcal{P}} = (-0.42841, 0.58930)$ and $\hat{\beta}_{\mathcal{NP}} = (-0.8647, 0.7603)$.
- 👉 Thus, under the considered NB GLM with log-link, it is estimated that a female crab being 1 Kgr heavier is expected to have about 2 ($e^{0.7603}$) satellites more than a crab of a 1 Kgr less.

➤ II.6.15 Remark

A common reason for overdispersion is heterogeneity (the response mean varies for different values of unobserved (explanatory) variables).

For the crabs example above, assume that SAT has a Poisson distribution at each fixed combination of the explanatory variables (weight, width, color, spine cond.). We considered the weight (in Kgr) as the only explanatory variable. Thus, the population of crabs having a certain weight is a mixture of several Poisson populations, corresponding to crabs of various levels for the other explanatory variables.

- A mixture model is a flexible way to account for overdispersion.
- It can be proved that a gamma mixture of Poisson distributions yields, marginally, the *negative binomial distribution*.


GLMs for Zero Inflated Discrete Distributions

Zero inflated data are count data for which the frequency of 0s is larger than expected under standard discrete models.

II.6.16 Zero-inflated Poisson Distribution

A discrete random variable Y is distributed according to a zero-inflated Poisson (ZIP) model (Lambert 1992), if

$$Y \sim \begin{cases} 0, & \text{with prob. } \pi = 1 - \phi \\ \mathcal{P}(\mu), & \text{with prob. } \phi \end{cases}$$

 Unconditional probability distribution of Y :

$$P(Y = 0) = (1 - \phi) + \phi e^{-\mu} = \pi + \phi e^{-\mu},$$

$$P(Y = j) = \phi \frac{e^{-\mu} \mu^j}{j!}, \quad j = 1, 2, \dots$$

II.6.17 Remark (Models for zero inflated (ZI) distributions)

- Zero inflation can be considered for other discrete distributions as well (e.g. ZINB).
- In a GLM set-up, consider independent random response variables Y_i , $i = 1, \dots, n$, with $Y_i \sim \text{ZIP}(\mu_i, \phi_i)$. The explanatory variables affecting the zero inflation parameters ϕ_i need not be the same as those affecting the Poisson parameters μ_i .

Thus, the parameters could be modeled by

$$\log(\mu_i) = \mathbf{x}_{1i}\boldsymbol{\beta}_1 \quad \text{and} \quad \text{logit}(\pi_i) = \mathbf{x}_{2i}\boldsymbol{\beta}_2,$$

for the count model and the zero inflation model, respectively, with $\phi_i = 1 - \pi_i$.

Hence ϕ_i are estimated by $\hat{\phi}_i = 1 - \hat{\pi}_i = (1 + \exp\{\mathbf{x}_{2i}\boldsymbol{\beta}_2\})^{-1}$.

👉 GLMs for ZINB responses are defined analogously.

👉 Zero inflated distributions can be fitted in the `pscl` R-package by the `zeroinfl()` function.

II.6.18 Example (horseshoe crab data: consideration of ZI)

The satellite counts are further modeled as zero inflated Poisson (ZIP) and ZINB distributions, applying the `zeroinfl()` function of the `pscl` package. 📖 See: 'P_NB_ZIP (crabs).pdf'

📖 $Y \sim \text{ZIP}$ (fitted by ZIP GLM with only intercept):

- The parameter ϕ is estimated to be $\hat{\phi} = 1 - \hat{\pi} = 1/(1 + e^{-0.6139}) = 0.649$, where $\hat{\beta}_{21} = -0.6139$ is the estimated intercept of the *zero-inflated model*.
- From the coefficients of the *count model*, we get $\hat{\mu} = e^{\hat{\beta}_{11}} = e^{1.50385} = 4.499$

Thus, under the ZIP model, the estimated number of female crabs with 0 satellites is

$$n(\hat{\pi} + \hat{\phi}e^{-\hat{\mu}}) = 173(0.3551 + 0.649e^{-4.499}) = 61.97$$

(compare to the estimated number of 0s under the Poisson model ($173e^{-2.99} = 9.3$, s. Example II.6.14) – the number of observed 0s in the sample is 62).

📖 $Y \sim \text{ZINB}$ (fitted by ZINB GLM with only intercept):

The estimated parameter values are given in 'P_NB_ZIP (crabs).pdf'.

Example (continues: ZINB GLM for crabs data)

```
> ZINB.mod <- zeroinfl(SAT ~ weight | weight + C, dist="negbin")
> summary(ZINB.mod)
```

```
Pearson residuals:
      Min       1Q   Median       3Q      Max
-1.3864 -0.7506 -0.2666  0.5298  3.7767

Count model coefficients (negbin with log link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.8961    0.3070   2.919  0.00351 **
weight       0.2169    0.1125   1.928  0.05383 .
Log(theta)   1.5802    0.3574   4.422  9.79e-06 ***

Zero-inflation model coefficients (binomial with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.2678    1.4078   0.901   0.368
weight      -1.7531    0.4429  -3.958  7.55e-05 ***
C             0.5985    0.2572   2.326   0.020 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 4.8558
Number of iterations in BFGS optimization: 12
Log-likelihood: -349.9 on 6 Df
```

$$\log \hat{\mu}_i = 0.896 + 0.217 \text{weight}_i$$

$$\text{dispersion: } \hat{\delta} = 1/\hat{\theta} = 1/4.8558 = 0.21$$

$$\text{probability mass at 0: } \hat{\pi}_i = 1 - \hat{\phi}, \text{ satisfying } \text{logit}(\hat{\pi}_i) = 1.866 - 1.753 \text{weight}_i + 0.598 C_i$$

* The color (C), from light to dark, is considered as ordinal (the assigned scores equal the levels of C).