Regression Analysis
Chapter 12: Binomial and Poisson Regression

Shaobo Jin

Department of Mathematics

Start With Linear Regression

We often express a linear regression model as

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i.$$

But if y_i is not continuous (e.g., only takes values 0 or 1, or only takes integer values), the model is not suitable.

Equivalently, our model is

$$\mathrm{E}\left(y_i \mid \boldsymbol{x}_i\right) = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

The key is that our model is build to model $\mu_i = \mathrm{E}\left(y_i \mid \boldsymbol{x}_i\right)$.

Components of Our Model

1. Random component: Response variable Y given $\boldsymbol{x}$ follows some distribution.
2. Linear predictor: The linear predictor for y_i is

$$\eta_i \quad = \quad \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

3. Link function $g()$: $g()$ transforms μ_i to the linear predictor

$$g(\mu_i) \quad = \quad \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

The link function must be monotonic and differentiable.

Binomial Data

Suppose that each individual has gone through n_i identical and independent trials with possible outcomes 0 and 1, and we record the number of 1's.

1. Random component: $y_i \mid \boldsymbol{x}_i \sim \text{Binomial}\,(n_i, \mu_i)$.
2. Linear predictor: $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$.
3. Link function $g\,()$: The mostly commonly used link function for binomial data is the logistic link:

$$g\,(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

The resulting model is called a logistic model.

Poisson Distribution For Counts

Suppose that we would like to model the number of cases (in principle unlimited). Then we have count data.

1. Random component: $y_i \mid \boldsymbol{x}_i \sim \text{Poisson}(\mu_i)$.
2. Linear predictor: $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$.
3. Link function $g()$: The mostly commonly used link function for Poisson data is the log link:

$$g(\mu_i) = \log(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

It is often called the log-linear model.

Interpretation of Logistic Model

Consider the logistic model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}.$$

Then,

$$\frac{P\left(Y_i = 1 | x_{i1} = a+1\right) / P\left(Y_i = 0 | x_{i1} = a+1\right)}{P\left(Y_i = 1 | x_{i1} = a\right) / P\left(Y_i = 0 | x_{i1} = a\right)}$$

$$= \frac{\exp\left\{\beta_0 + \beta_1 (a+1) + \beta_2 x_2\right\}}{\exp\left\{\beta_0 + \beta_1 a + \beta_2 x_2\right\}} = \exp\left\{\beta_1\right\}$$

is the (conditional) odds ratio (adjusting for other covariates).

Generally speaking, β_j is the expected change in the log odds for one unit increase in x_{ij}, provided that the sum of the other terms are fixed.

Common Mistakes

Some common mistakes people often commit:

1. The logistic model is

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i,$$

   where e_i is the error term.
2. Logistic regression does not have distributional assumptions.

Behind Link Function: Distribution Assumption!

- Suppose that, in an ideal world, we could observe continuous y_i^* and we could use the linear model

$$y_i^* = \boldsymbol{x}_i^T \boldsymbol{\beta} - \varepsilon_i.$$

- However, in reality, we only observe y_i such that

$$y_i = \begin{cases} 0, & \text{if } y_i^* < 0, \\ 1, & \text{if } y_i^* \geq 0. \end{cases}$$

- In such a case, we often assume that $Y_i \sim \text{Bernoulli}\,(\pi_i)$.
- Note that

$$\pi_i = P\left(Y_i = 1\right) = P\left(Y_i^* \geq 0\right) = P\left(\varepsilon_i \leq \boldsymbol{x}_i^T \boldsymbol{\beta}\right) = F_{\varepsilon_i}\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right).$$

- The link function corresponds to the distribution assumption that we put on ε_i.

Interpretation: Rate Ratio

Suppose that the Poisson model is

$$\log \mu_i = \beta_0 + \beta_1 x_{i1} + \sum_{p=2} \beta_p x_{ip}.$$

Then

$$\frac{\mathrm{E}\left(Y_i | x_{i1} = a+1\right)}{\mathrm{E}\left(Y_i | x_{i1} = a\right)} = \frac{\exp\left(\beta_0 + \beta_1 (a+1) + \sum_{p=2} \beta_p x_{ip}\right)}{\exp\left(\beta_0 + \beta_1 a + \sum_{p=2} \beta_p x_{ip}\right)} = \exp\left(\beta_1\right),$$

provided that $\sum_{p=2} \beta_p x_{ip}$ remain the same.

This ratio is called the rate ratio: one unit change in x_{i1} will result in a multiplicative effect of $\exp(\beta_1)$ on μ_i, provided that $\sum_{p=2} \beta_p x_{ip}$ remain the same.

Maximum Likelihood

The log-likelihood for the logistic model is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \log(\mu_i) + (n_i - y_i) \log(1 - \mu_i) + \log \binom{n_i}{y_i} \right],$$

where

$$\mu_i = \frac{\exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)}{1 + \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)}.$$

The maximum likelihood estimator does not have closed form expression and must be obtained numerically.

Maximum Likelihood

The log-likelihood for the Poisson model is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\mu_i \log(y_i) - \log(y_i!) - \mu_i\right],$$

where

$$\mu_i = \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right).$$

The maximum likelihood estimator does not have closed form expression and must be obtained numerically.

Confidence Interval

- Assume that n is large enough together with some other assumptions, the distribution of $\hat{\boldsymbol{\beta}}$ can be approximated by a normal distribution

$$N\left(\boldsymbol{\beta},\ \boldsymbol{\Omega}\right),$$

  for some $\boldsymbol{\Omega}$.
- The standard error of $\hat{\beta}_j$ can be approximated by $\sqrt{\hat{w}_j}$, where $\hat{w}_j$ is the (j,j)th element of $\hat{\boldsymbol{\Omega}}$.
- The Wald confidence interval is

$$\hat{\beta}_j - z_{1-\alpha/2}\sqrt{\hat{w}_j} \leq \beta_j \leq \hat{\beta}_j + z_{1-\alpha/2}\sqrt{\hat{w}_j},$$

  where $z_{1-\alpha/2}$ is the $1-\alpha/2$ percentile of $N\left(0,1\right)$.

Inference for $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\mu}}$

- The distribution of $\hat{\boldsymbol{\eta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is then approximately

$$N\left(\boldsymbol{\eta}, \boldsymbol{X}\boldsymbol{\Omega}\boldsymbol{X}^T\right).$$

- If $\hat{\kappa}_i$ is the (i,i)th diagonal element of $\boldsymbol{X}\boldsymbol{\Omega}\boldsymbol{X}^T$, the Wald confidence interval for η_i is

$$\hat{\eta}_i - z_{1-\alpha/2}\sqrt{\hat{\kappa}_i} \leq \eta_i \leq \hat{\eta}_i + z_{1-\alpha/2}\sqrt{\hat{\kappa}_i},$$

  where $z_{1-\alpha/2}$ is the $1-\alpha/2$ percentile of $N(0,1)$.
- The confidence interval for μ_i is

$$g^{-1}\left(\hat{\eta}_i - z_{1-\alpha/2}\sqrt{\hat{\kappa}_i}\right) < \mu_i < g^{-1}\left(\hat{\eta}_i + z_{1-\alpha/2}\sqrt{\hat{\kappa}_i}\right),$$

  where $g^{-1}()$ is the inverse function of $g()$.

Three Models

In GLM we can at least define three models

1. saturated model: fits the data "perfectly".
2. model of interest: the GLM that you fitted, $g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$ for each observation i.
3. null model: the GLM that only contains the intercept, $g(\mu_i) = \beta_0$ for any observation i.

Saturated Model and Null Model

The saturated model that fits the data "perfectly" uses y_i to estimate μ_i for all i, i.e., $\hat{\mu}_i = y_i$.

- This is not a good model because of overfitting.

The null model only includes the intercept in the model

$$g(\mu_i) = \beta_0.$$

- The fitted mean for individual i is $\hat{\mu}_i = \bar{y}$.

Example: Poisson Distribution

Assume Y_1, ..., Y_n are independent and $Y_i \sim \text{Poisson}(\mu_i)$. The log-likelihood function is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} [\mu_i \log(y_i) - \log(y_i!) - \mu_i].$$

1. For the saturated model, $\hat{\mu}_i = y_i$. The log-likelihood is denoted by

$$\ell(\boldsymbol{y}; \boldsymbol{y}) = \sum_{i=1}^{n} [\mu_i \log(y_i) - \log(y_i!) - y_i].$$

2. For the null model, $\hat{\mu}_i = \bar{y}$. The log-likelihood is denoted by

$$\ell(\bar{y}; \boldsymbol{y}) = \sum_{i=1}^{n} [\bar{y} \log(y_i) - \log(y_i!) - \bar{y}].$$

Model of Interest

- Our model of interest uses some covariates to model μ_i as $g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$.
- The fitted mean for individual i is $\hat{\mu}_i = g^{-1}\left(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}\right)$, where $g^{-1}()$ is the inverse function of $g()$.

Poisson Model

The log-likelihood of the Poisson model is denoted by

$$\ell(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) = \sum_{i=1}^{n} \left[\hat{\mu}_i \log(y_i) - \log(y_i!) - \hat{\mu}_i\right],$$

where $\hat{\mu}_i = g^{-1}\left(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}\right)$.

Scaled Deviance and (Residual) Deviance

H_0 : The model fits the data as good as the saturated model

H_1 : The model fits the data worse than the saturated model

The (residual) deviance is defined to be

$$D\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right) \quad = \quad -2\left[\ell\left(\hat{\boldsymbol{\mu}};\boldsymbol{y}\right)-\ell\left(\boldsymbol{y};\boldsymbol{y}\right)\right],$$

just the likelihood ratio test statistics. The greater the deviance, the poorer the fit.

If the model fits the data well, then $D\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right) \approx \chi^2\left(m-p\right)$, where *$m$ is the number of parameters in the saturated model* and p is the number of parameters in the model of interest. The model fits the data well if

$$D\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right) \quad \leq \quad \chi^2_{1-\alpha}\left(m-p\right).$$

Grouped Data and Ungrouped Data

```
####  Ungrouped data
Ungroup
```

```
##    y x1         x2
## 1  0  0  1.0304591
## 2  0  0  0.1288419
## 3  1  0 -0.9201021
## 4  0  0  0.1590724
## 5  1  0  0.4947146
## 6  1  1 -0.6870865
## 7  1  1  0.5701739
## 8  0  1  0.2244959
## 9  0  1 -1.6308122
## 10 1  1  0.6338810
## 11 0  1  2.5272024
## 12 0  1 -1.1834090
```

```
####  Grouped data
Group
```

```
##   fail success x1 x2
## 1    2       1  0  0
## 2    1       1  0  1
## 3    1       2  1  0
## 4    3       1  1  1
```

Grouped Data Expressed as Ungrouped

```
##     y x1 x2
## 1   0  0  0
## 2   0  0  0
## 3   1  0  0
## 4   0  0  1
## 5   1  0  1
## 6   1  1  0
## 7   1  1  0
## 8   0  1  0
## 9   0  1  1
## 10  1  1  1
## 11  0  1  1
## 12  0  1  1
```

```
##   fail success x1 x2
## 1    2       1  0  0
## 2    1       1  0  1
## 3    1       2  1  0
## 4    3       1  1  1
```

Response Residual

Similar to multiple linear regression, we can define various residuals.

1. The response residual is

$$y_i - \hat{\mu}_i.$$

2. The Pearson residual is

$$\frac{y_i - \hat{\mu}_i}{\sqrt{\text{Var}\left(\hat{\mu}_i\right)}}.$$

3. The deviance residual is

$$\text{sign}\left(y_i - \hat{\mu}_i\right) \cdot \sqrt{d_i},$$

whose squared values sum to the residual deviance $D = \sum_{i=1}^{n} d_i$.

Residual Plots in R

The residual plots can often be produced using

- plot(your GLM fitted object)
- car::residualPlots()

However, the residual plots are not always useful, depending the characteristic of y. One mistake that is most relevant to the course is that

Built in residual plots are always useful for diagnostics when you have binary data or count data.

We will focus on the randomized quantile residual: DHARMa::simulateResiduals() followed by DHARMa::plot()

How Does DHARMa Work?

Algorithm 1: Randomized quantile residuals

1 Fit your model.;
2 **while** *for each observation* i **do**
3 Simulate q response variables using the predicted μ_i from the model. ;
4 Compute the percentage that simulated response less than observed y_i and the percentage that simulated response less than or equal to observed y_i. ;
5 If two percentages are the same, the quantile residual is the percentage. If not the same, the randomized quantile residual is draw from a uniform distribution between two percentages. ;
6 **end**
7 If your model is correct, the cdf of y follows a uniform distribution. Hence, we expect the quantile residuals to be uniform and spread out everywhere.

Compare Two Models

- Suppose that we have two models (M_0 and M_1) and that M_0 nested in M_1 with different $\boldsymbol{x}$. The deviances for M_0 and M_1 are

$$M_0: \quad D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_0\right) \quad \text{and} \quad M_1: \quad D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_1\right).$$

- If $a(\phi) = \phi / w_i$ and $\phi = 1$, the difference in the deviance is

$$G^2\left(M_0 | M_1\right) \quad \stackrel{\text{def}}{=} \quad D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_0\right) - D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_1\right),$$

  which is the test statistic for H_0: M_0 versus H_1: M_1.
- We reject H_0 if

$$G^2\left(M_0 | M_1\right) \quad \geq \quad \chi^2_{1-\alpha}\left(p_1 - p_0 > 0\right),$$

  where M_0 has p_0 parameters and M_1 has p_1 parameters.

Be Careful with Ungrouped Data

- For grouped data, the number of groups be fixed but the sample size increases.
- For ungrouped data, the number of groups increases as the sample size increases.

	Grouped data	Ungrouped data
Res. Dev.	asymptotically chi-square, if the model fits the data well.	not always useful
Res. plots (L7)	useful	not always useful
Diff. in Res. Dev.	useful	useful

Classification Table

A classification table or a confusion matrix cross-classifies the binary response y and the predicted binary response.

	Observed	
Predicted	1 (TRUE)	2 (FALSE)
1 (TRUE)	m_{11} (TP)	m_{12} (FP)
2 (FALSE)	m_{21} (FN)	m_{22} (TN)

From the classification table, we can compute

$$\begin{aligned}
\text{sensitivity (true positive rate)} &= \frac{m_{11}}{m_{11}+m_{21}}, \\
\text{specificity (true negative rate)} &= \frac{m_{22}}{m_{12}+m_{22}}, \\
\text{false positive rate} &= \frac{m_{12}}{m_{12}+m_{22}}, \\
\text{false negative rate} &= \frac{m_{21}}{m_{11}+m_{21}}.
\end{aligned}$$

F-Score of Binary Classification

	Observed	
Predicted	1 (TRUE)	2 (FALSE)
1 (TRUE)	m_{11}	m_{12}
2 (FALSE)	m_{21}	m_{22}

In machine learning,

$$\begin{aligned} \text{Precision} &= \frac{m_{11}}{m_{11}+m_{12}}, \\ \text{Recall} &= \frac{m_{11}}{m_{11}+m_{21}}. \end{aligned}$$

The F-score is

$$\text{F-score} = 2\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2m_{11}}{2m_{11}+m_{12}+m_{21}} \in [0,1].$$

Rule of Thumb: >0.7 or >0.8 is considered to be OK. >0.9 as good.

Predictive Power: ROC curve

- An issue with the classification table is that you need to determine a cut-off probability in order to convert probabilities into classification.
- Receiver operation characteristic (ROC) curve plots the sensitivity (true positive) as a function of 1 – specificity (false positive) as the cut-off probability decreases from 1 to 0.
- The better predictive power, the higher the ROC curve and the greater the area under it.

Overdispersion

An super strong assumption of Poisson is that mean and variance are the same.

- Suppose that $Y|\lambda \sim \text{Poisson}(\lambda)$ and $\lambda|\mu \sim \text{Gamma}$ with mean μ and variance μ^2/θ for some shape parameter θ.
- The marginal distribution of Y is negative binomial distribution with

$$P(Y=y) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)\Gamma(y+1)}\left(\frac{\theta}{\mu+\theta}\right)^{\theta}\left(\frac{\mu}{\mu+\theta}\right)^{y}, \quad y=0,1,2,....$$

- The mean of Y is μ and the variance is $\mu+\mu^2/\theta$.
- The log link is commonly used for negative binomial regression.
- In the special case where $\theta \to \infty$, the negative binomial distribution reduces to the Poisson distribution.

Zero-Inflated model

- Sometimes, our data have an excess of zero counts, e.g., the number of times in the past week that individuals seek for medical care.
- The zero-inflated Poisson model assumes

$$Y_i \sim \begin{cases} 0, & \text{with probability } 1-\pi_i, \\ \text{Poisson}\left(\mu_i\right), & \text{with probability } \pi_i. \end{cases}$$

- We need a model for π_i and another model for μ_i. For example,

$$\begin{aligned} \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \boldsymbol{x}_{1i}^T\boldsymbol{\beta}_1, \\ \log\left(\mu_i\right) &= \boldsymbol{x}_{2i}^T\boldsymbol{\beta}_2. \end{aligned}$$

Positive Poisson Part

- In the zero-inflated Poisson model, Y_i is always 0 in population 1 and Y_i is Poisson in population 2 (Y_i can still be 0 in population 2).

Number of insurance claims: May not claim

- People who do not have insurance have 0 claims.
- People who have insurance and have a large excess can either claim or not claim, if an incident happens.

- Different case: Y_i is always 0 in population 1 and Y_i is a truncated Poisson in population 2 (Y_i cannot be 0 in population 2).

Number of insurance claims: Always claim

- People who do not have insurance have 0 claims.
- People who have insurance and whose excess is 0 always claim, if an incident happens.

Hurdle Model

- The hurdle model assumes

$$Y_i \sim \begin{cases} 0, & \text{with probability } 1-\pi_i, \\ \text{truncated-at-zero Poisson}\left(\mu_i\right), & \text{with probability } \pi_i, \end{cases}$$

  where μ_i is the mean for the untruncated Poisson distribution.
- We still need a model for π_i and another model for μ_i, e.g.,

$$\begin{aligned} \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \boldsymbol{x}_{1i}^T\boldsymbol{\beta}_1, \\ \log\left(\mu_i\right) &= \boldsymbol{x}_{2i}^T\boldsymbol{\beta}_2. \end{aligned}$$

Zero-Inflated Model and Hurdle Model With Negative Binomial

We can easily generalize the zero-inflated model and the hurdle model from Poisson distribution to negative binomial distribution as

$$Y_i \sim \begin{cases} 0, & \text{with probability } 1-\pi_i, \\ NB\left(\mu_i\right), & \text{with probability } \pi_i, \end{cases}$$

$$Y_i \sim \begin{cases} 0, & \text{with probability } 1-\pi_i, \\ \text{truncated-at-zero } NB\left(\mu_i\right), & \text{with probability } \pi_i. \end{cases}$$

In general, we have

	How to treat 0		
Distribution	Raw	Zero-inflated model	Hurdle model
Poisson	glm()	zeroinfl()	hurdle()
Negative binomial	glm.nb()	zeroinfl()	hurdle()

Poisson Model for Rates

- A usual Poisson regression model for counts is

$$\log \mu_i \quad = \quad \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

- μ_i is often proportional to an index t_i. Then the sample rate is y_i/t_i, with expectation μ_i/t_i.
  - The total number of questions raised in a lecture is count.
  - The total number of questions divided by the number of students is rate. A student can raise more than one question.
- The model for the expected rate is

$$\log\left(\frac{\mu_i}{t_i}\right) = \boldsymbol{x}_i^T \boldsymbol{\beta} \quad \Rightarrow \quad \log(\mu_i) = \log(t_i) + \boldsymbol{x}_i^T \boldsymbol{\beta},$$

which has an offset.