Generalized Linear Models
Poisson Regression and Multinomial (Logistic) Regression

University of California, San Diego
Instructor: Armin Schwartzman

Aircraft Damage dataset

- Consider the `Aircraft Damage` dataset taken from *Applied Linear Regression* (4th Edition) by Weisberg.

  This is a dataset on the result of strike missions during the Vietnam War with A-4 or A-6 aircrafts.

- The variables are:

  y: is the number of locations where the aircraft was damaged
  x_1: indicates the type of plane (0 for A-4; 1 for A-6)
  x_2: is the bomb load in tons
  x_3: is the total months of aircrew experience

Dealing with count data: standard model

- The response represents counts.

  Here the number of different values it takes is not large compared to the sample size. It could be considered numerical, but we have another option.

- The standard linear model

$$y|\mathbf{x} \sim \mathcal{N}(\mu(\mathbf{x}), \sigma^2), \qquad \mu(\mathbf{x}) = \mathbb{E}\,(y|\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x}$$

  is not appropriate because

  1. y is an integer
  2. $\mu(\mathbf{x})$ will be negative for some $\mathbf{x}$'s

  This model is relevant and may hold approximately if y takes a large number of values. This is because the Poisson distribution looks normal if its mean is large.

Dealing with count data: Poisson model

- A more appropriate is the Poisson regression model:

$$y|\mathbf{x} \sim \mathrm{Poisson}(\mu(\mathbf{x})), \qquad \log(\mu(\mathbf{x})) = \boldsymbol{\beta}^\top \mathbf{x}$$

- The logarithm could be replaced by any other (link) function $g : (0, \infty) \to (-\infty, \infty)$ monotone.

- Note that, by design, the variance is a function of the mean:

$$\sigma(\mathbf{x})^2 = \mathrm{Var}\,(y|\mathbf{x}) = \mu(\mathbf{x})$$

MLE for Poisson regression

A Poisson model is usually fitted by maximum likelihood. The log-likelihood is:

$$\begin{aligned}\ell(\mu_1, \ldots, \mu_n) &= \sum_{i=1}^n \left[y_i \log(\mu_i) - \mu_i - \log(y_i!)\right] \\ &= \sum_{i=1}^n \left[y_i \mathbf{b}^\top \mathbf{x}_i - \exp(\mathbf{b}^\top \mathbf{x}_i) - \log(y_i!)\right]\end{aligned}$$

since $\mu_i = \mu(\mathbf{x}_i) = \exp(\mathbf{b}^\top \mathbf{x}_i)$.

We want to maximize this function of $\mathbf{b}$. No closed form expression exists in general, but the problem is convex (maximize a concave function).

Deviance

The deviance is defined as twice the log-likelihood ratio:

$$\mathrm{DEV} = 2 \log \frac{\mathcal{L}(y_1, \ldots, y_n)}{\mathcal{L}(\hat{\mu}_1, \ldots, \hat{\mu}_n)} = 2\left[\ell(y_1, \ldots, y_n) - \ell(\hat{\mu}_1, \ldots, \hat{\mu}_n)\right]$$

For linear regression:

$$\ell(\mu_1, \ldots, \mu_n) = -\log(\sqrt{2\pi}\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu_i)^2 \qquad \Rightarrow \qquad \mathrm{DEV} = \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \hat{\mu}_i)^2$$

For Poisson regression:

$$\begin{aligned}\mathrm{DEV} &= 2\left(\sum_{i=1}^n \left[y_i \log(y_i) - y_i - \log(y_i!)\right] - \sum_{i=1}^n \left[y_i \log(\hat{\mu}_i) - \hat{\mu}_i - \log(y_i!)\right]\right) \\ &= 2\sum_{i=1}^n \left[y_i \log(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i\right]\end{aligned}$$

The deviance plays the role of the residual sum of squares.

Education by Age dataset

- Consider the `Education by Age` data taken from
  http://lib.stat.cmu.edu/DASL/Datafiles/Educationbyage.html.

  There are two categorical variables (factors): age group and highest degree.

- The main question is whether the two factors are independent.

- In general, suppose we have two paired categorical variables $\{(U_i, V_i) : i = 1, \ldots, n\}$, with

$$U_i \in \{u_a : a = 1, \ldots, A\}, \qquad V_i \in \{v_b : b = 1, \ldots, B\}$$

  If the observations are independent, then the cell counts

$$y_{ab} = \#\{i : (U_i, V_i) = (u_a, v_b)\}$$

  are sufficient statistics.

  These counts are organized in a (two-way) contingency table with A rows and B columns, which is the analog of a two-way table for numerical data.

- Note that $y = (y_{ab} : a = 1, \ldots, A; b = 1, \ldots, B)$ is multinomial with sample size n and probabilities $p_{ab} = \mathbb{P}(U = u_a, V = v_b)$.

Pearson's χ^2 test

- Testing for independence means testing $H_0 : p_{ab} = p_{a.}p_{.b}$, where

$$p_{a.} = \mathbb{P}(U = u_a), \qquad p_{.b} = \mathbb{P}(V = v_b)$$

- The most popular method is the chi-square test of independence. It rejects for large values of

$$\mathbb{X} = \sum_{a=1}^{A} \sum_{b=1}^{B} \frac{(y_{ab} - \widehat{y}_{ab})^2}{\widehat{y}_{ab}} \quad \text{where} \quad \widehat{y}_{ab} = \frac{y_{a.}\, y_{.b}}{y_{..}}$$

  - y_{ab} is the observed count for cell (a, b), and

$$y_{a.} = \sum_b y_{ab}, \qquad y_{.b} = \sum_a y_{ab} \qquad y_{..} = \sum_a \sum_b y_{ab} = n$$

    are the sum for row a, the sum for column b, and the total sum (equal to the sample size).
  - $\widehat{y}_{ab}$ is the predicted count for cell (a, b) under independence.
  - Under the null, as $n \to \infty$, $\mathbb{X}$ has the limiting distribution $\chi^2_{AB-A-B+1} = \chi^2_{(A-1)(B-1)}$.

Poisson model for contingency tables

- Count data is, strictly speaking, multinomial data (assuming the observations were independently sampled from a homogeneous population).
- As an approximation, we model the count data as Poisson distributed:

$$y_{ab} \sim \text{Poisson}(\mu_{ab}), \qquad \mu_{ab} = np_{ab}$$

  This approximation is accurate if the sample is large enough.
- Then testing for independence of the two factors is formalized as testing

$$H_0 : \mu_{ab} = \frac{\mu_{a.}\ \mu_{.b}}{n} \quad \forall a, b$$

- From a Poisson regression point of view, testing for H_0 corresponds to testing for the restricted model with no interaction term.

Cleveland Clinic Foundation heart disease study

- Consider the `cleveland` dataset taken from
  https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci

  8 variables are categorical, and 6 variables are numerical.

  We first focus on predicting `cond` based on the other (14) characteristics.
- The response `cond` is categorical (binary), therefore this is a classification task.
- A standard linear model is not that relevant here.

Logistic regression

- Assume the response y is binary and "coded" as $y \in \{0, 1\}$.
- We want to fit the following model:

$$y|\mathbf{x} \sim \text{Bernoulli}(\mu(\mathbf{x})), \qquad \mu(\mathbf{x}) = \mathbb{P}\left(y = 1|\mathbf{x}\right) = \mathbb{E}\left(y|\mathbf{x}\right)$$

  with

$$\mu(\mathbf{x}) = \frac{e^{\boldsymbol{\beta}^\top \mathbf{x}}}{1 + e^{\boldsymbol{\beta}^\top \mathbf{x}}}.$$

  This relationship is defined through the logit link function, yielding the *log odds*

$$\text{logit}(\mu(\mathbf{x})) = \log\left(\frac{\mu(\mathbf{x})}{1 - \mu(\mathbf{x})}\right) = \boldsymbol{\beta}^\top \mathbf{x}$$

- Note that, by design, the variance is a function of the mean:

$$\sigma(\mathbf{x})^2 = \text{Var}\left(y|\mathbf{x}\right) = \mu(\mathbf{x})(1 - \mu(\mathbf{x}))$$

Coefficient interpretation

- Let $\mathbf{e}_j = (0, \ldots, 1, \ldots, 0)^\top$ be a vector of zeros with a 1 in the j-th position. Suppose we increase variable x_j by 1 unit. Then the log odds ratio is:

$$\log\left(\frac{\mu(\mathbf{x}+\mathbf{e}_j)}{1-\mu(\mathbf{x}+\mathbf{e}_j)}\right) - \log\left(\frac{\mu(\mathbf{x})}{1-\mu(\mathbf{x})}\right) = \boldsymbol{\beta}^\top(\mathbf{x}+\mathbf{e}_j) - \boldsymbol{\beta}^\top\mathbf{x} = \beta_j$$

- The coefficient β_j is the log odds ratio when increasing x_j by unit while keeping all the other variables constant.

Classification boundary

- This model predicts (classifies) $y = 1$ at a new observation $\mathbf{x}$ if $\mu(\mathbf{x}) > 1/2$, meaning that it predicts the class that is the most likely at $\mathbf{x}$.

  As a consequence, the boundary b/w the two classes is the hyperplane:

$$\boldsymbol{\beta}^\top\mathbf{x} = 0$$

  (If the first entry of $\mathbf{x}$ is equal to 1 to represent the intercept, then this is an affine hyperplane.)

MLE and Deviance

- We again fit the model by maximum likelihood.

  Let $g = \mathrm{logit}$. The log-likelihood is:

$$\begin{aligned}\ell(\mu_1, \ldots, \mu_n) &= \sum_{i=1}^n \left[y_i \log(\mu_i) + (1-y_i)\log(1-\mu_i)\right] \\ &= \sum_{i=1}^n \left[y_i \log(g^{-1}(\mathbf{b}^\top\mathbf{x}_i)) + (1-y_i)\log(1-g^{-1}(\mathbf{b}^\top\mathbf{x}_i))\right]\end{aligned}$$

  Maximizing this concave function (of $\mathbf{b}$) is a convex optimization problem.

- The deviance has the following expression here:

$$\mathrm{DEV} = -2\sum_{i=1}^n \left[y_i \log(\hat{\mu}_i) + (1-y_i)\log(1-\hat{\mu}_i)\right]$$

  where $\hat{\mu}_i = g^{-1}(\widehat{\boldsymbol{\beta}}^\top\mathbf{x}_i)$.

Multinomial regression

- We turn to predicting `attplus` based on the individual characteristics. This is a categorical variable taking 5 distinct values.
- Assume the response y is categorical with K levels, e.g., $y \in \{1, \dots, K\}$.
- Let $\mu_k(\mathbf{x}) = \mathbb{P}(y = k|\mathbf{x})$. For $k = 1, \dots, K-1$, we model these as

$$\log\left(\frac{\mu_k(\mathbf{x})}{\mu_K(\mathbf{x})}\right) = \boldsymbol{\beta}_k^\top \mathbf{x}$$

  same as

$$\mu_k(\mathbf{x}) = \frac{e^{\boldsymbol{\beta}_k^\top \mathbf{x}}}{1 + \sum_{\ell=1}^{K-1} e^{\boldsymbol{\beta}_\ell^\top \mathbf{x}}}, \quad k = 1, \dots, K-1$$

$$\mu_K(\mathbf{x}) = \frac{1}{1 + \sum_{\ell=1}^{K-1} e^{\boldsymbol{\beta}_\ell^\top \mathbf{x}}}$$

- In this model, the boundary b/w the classes k and ℓ is the hyperplane:

$$(\boldsymbol{\beta}_k - \boldsymbol{\beta}_\ell)^\top \mathbf{x} = 0$$

Overdispersion

- Assuming a one-parameter family as in the Poisson or logistic models implicitly ties the variance to the mean, in that $\sigma^2 = V(\mu)$. This may be found to be incongruent with the data.
- Introduce the dispersion parameter $\phi = \sigma^2 / V(\mu)$. The one-parameter model is correct when $\phi = 1$. When $\phi > 1$, we have overdispersion.
- The function `glm` in `R` allows for an overdispersion parameter.