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:

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

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:

$$\mathrm{DEV} = 2\left[\ell(y_1, \dots, y_n) - \ell(\hat{\mu}_1, \dots, \hat{\mu}_n)\right]$$

For linear regression:

$$\ell(\mu_1, \dots, \mu_n) = -\sum_{i=1}^n (y_i - \mu_i)^2 \qquad \Rightarrow \qquad \mathrm{DEV} = \sum_{i=1}^n (y_i - \hat{\mu}_i)^2$$

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 $\chi^2_{AB-A-B+1}$.

Poisson model for contingency tables

- 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

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

  which means that the matrix of expected counts (μ_{ab}) has rank 1.

  We are testing against the full model where the μ_{ab}'s are unrestricted, except for the condition $\sum_{a,b} \mu_{ab} = n$.

- The corresponding estimates are:

$$\begin{aligned} (H_0) : \widehat{\mu}_{ab} &= \widehat{y}_{ab} = \frac{y_{a.}\ y_{.b}}{y_{..}} \\ (H_1) : \widehat{\mu}_{ab} &= y_{ab} \end{aligned}$$

Testing Poisson models

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

$$\begin{aligned} (H_0) : \log(\mu_{ab}) &= \eta + \alpha_a + \beta_b \\ (H_1) : \log(\mu_{ab}) &= \eta + \alpha_a + \beta_b + (\alpha\beta)_{ab} \end{aligned}$$

- The corresponding estimates are:

$$\begin{aligned} (H_0) : \widehat{\mu}_{ab} &= \widehat{y}_{ab} = \frac{\overline{y}_{a.}\overline{y}_{.b}}{\overline{y}_{..}} \\ (H_1) : \widehat{\mu}_{ab} &= y_{ab} \end{aligned}$$

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:

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

  with

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

  same as

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

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

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

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

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

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

- The deviance has the following expression here:

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

  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:

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

Generalized linear models

- Poisson, logistic and multinomial regression, as well as the standard linear regression when σ^2 is known, all assume that $y | \mathbf{x} \sim f_{\theta(\mathbf{x})}$, where f_θ is a one-parameter exponential family.

  In its *canonical form*:

$$f_\theta(y) = \exp\left(\theta\, y - c(\theta)\right) h(y)$$

- Let

$$\mu(\mathbf{x}) = \mathbb{E}(y | \mathbf{x}) = c'(\theta(\mathbf{x})), \quad \sigma(\mathbf{x})^2 = \mathrm{Var}(y | \mathbf{x}) = c''(\theta(\mathbf{x}))$$

- The link function relates $\mu(\mathbf{x})$ to a linear combination in $\mathbf{x}$:

$$g(\mu(\mathbf{x})) = \boldsymbol{\beta}^\top \mathbf{x}$$

  The canonical link function is such that $g(\mu) = \theta$, meaning, $g = (c')^{-1}$.

  We take the link function to be this in what follows, so that $\theta(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x}$.

- Note that the variance is a function of the mean:

$$\sigma^2 = c''(\theta) = c'' \circ (c')^{-1}(\mu) \stackrel{\text{def}}{=} V(\mu)$$

  V is called the variance function.

- Specifying a GLM amounts to setting the linear model, the link function and the variance function.

Examples

- For Poisson regression: $g = \log$ and $V = \mathrm{id}$.
- For a logistic model: $g = \mathrm{logit}$ and $V(\mu) = \mu(1-\mu)$.
- For linear regression (with known variance): $g = \mathrm{id}$ and $V = 1$.

Maximum likelihood estimnation

- Generalized linear models are fitted by maximum likelihood.

  There is no closed form expression in general.

- We want to maximize the log likelihood. Assuming g is strictly increasing and differentiable, we may look at critical points where the log likelihood has zero gradient:

$$\nabla \ell(\mathbf{b}) = 0, \qquad \ell(\mathbf{b}) = \sum_{i=1}^{n} \log f_{\theta_i}(y_i), \quad \theta_i = \mathbf{b}^\top \mathbf{x}_i$$

  The gradient $\nabla \ell(\mathbf{b})$ is often called the score vector.

- This is usually solved by the Newton-Raphson method:

$$\mathbf{b} \leftarrow \mathbf{b} + \mathcal{J}(\mathbf{b})^{-1} \nabla \ell(\mathbf{b})$$

  where $\mathcal{J}(\mathbf{b}) = (\mathcal{J}_{jk}(\mathbf{b}))$ with $\mathcal{J}_{jk}(\mathbf{b}) = -\partial_{jk} \ell(\mathbf{b})$ is the observed information matrix.

- In practice, we replace $\mathcal{J}(\mathbf{b})$ with $\mathcal{I}(\mathbf{b}) = \mathbb{E}\left[\mathcal{J}(\mathbf{b})\right] = \mathbb{E}\left[\nabla \ell(\mathbf{b}) \nabla \ell(\mathbf{b})^\top\right]$, leading to Fisher scoring, aka iteratively reweighted least squares.

- Simple calculations lead to:

$$\partial_j \ell = \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{V(\mu_i) g'(\mu_i)}$$

  and

$$\mathcal{I}_{jk} = -\partial_{jk} \ell = \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{V(\mu_i)(g'(\mu_i))^2}$$

  where $\mu_i = g^{-1}(\mathbf{b}^\top \mathbf{x}_i)$.

Deviance

- The deviance is defined as

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

  where $\hat{\theta}_i = \widehat{\boldsymbol{\beta}}^\top \mathbf{x}_i$. This is $-2\times$ the difference of the unscaled log-likelihoods of the model evaluated at the MLEs and the completely unconstrained model.

- The deviance is the analog of the error sum of squares due to regression in a standard linear model.

Residuals

- The Pearson residuals are defined as

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

  where $\hat{\mu}_i$ is the estimated mean of y at $\mathbf{x} = \mathbf{x}_i$.

- The deviance residuals are defined as

$$r_i = \mathrm{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$$

  where

$$d_i = 2\big[y_i g(y_i) - c(g(y_i)) - y_i g(\hat{\theta}_i) + c(\hat{\theta}_i)\big]$$

  is the contribution to the deviance of observation i.

- The diagnostic plots and inference are based on either of these residuals.

Asymptotic distributions

- If the model is correct then, in some asymptotic sense (when the sample size is large), the MLE is approximately normal

$$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}(\boldsymbol{\beta}, (\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1})$$

  where $\boldsymbol{\beta}$ is the true parameter and $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_n)$ is a weight matrix with $w_i^{-1} = V(\mu_i)g'(\mu_i)^2$ and $\mu_i = g^{-1}(\boldsymbol{\beta}^\top\mathbf{x}_i)$.

Analysis of deviance

- This is the analog of ANOVA for generalized linear models. It is also the likelihood ratio test for comparing two (nested) models.

- For testing a submodel (H_0) against a larger model (H_1), the inference is based on the deviances. We reject for large values of

$$D = \mathrm{DEV}_0 - \mathrm{DEV}_1$$

  Under the null, D has (asymptotically) the χ^2-distribution with $\mathrm{df}_1 - \mathrm{df}_0$ degrees of freedom.

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.

- One way to test for $\phi = 1$ versus $\phi > 1$ is to reject for large values of DEV of the full model (assuming it can be fitted), which under the null is approximately χ^2_{n-p}, where p is the size of the full model.

- If there is evidence of overdispersion, then one may want to fit a two-parameter exponential family model like

$$f_{\theta,\phi}(y) = \exp\left(\frac{\theta(\mathbf{x})\, y - c(\theta(\mathbf{x}))}{\phi}\right) h(y, \phi)$$

  The normal (standard linear regression) model is already of this form.