Analysis of Categorical Data
Chapter 4: Introduction to Generalized Linear Models

Shaobo Jin

Department of Mathematics

Intended Learning Outcome

Through this chapter, you should be able to

1. verify exponential dispersion family,
2. describe the components of GLM,
3. fit GLMs,
4. perform model comparison,
5. perform residual analysis.

Exponential Dispersion Family

A random variable Y_i belongs to the exponential dispersion family if the pmf/pdf is of the form

$$f\left(y_i;\theta_i,\phi_i\right)=\exp\left\{\frac{y_i\theta_i-b\left(\theta_i\right)}{\phi_i}+c\left(y_i,\phi_i\right)\right\}.$$

- θ_i is the natural parameter.
- $\phi_i>0$ is the dispersion parameter, which can be either known or unknown. We often have $\phi_i=\phi$ or $\phi_i=\phi/w_i$ with a known w_i.
- No y_i can be included in $b\left(\theta_i\right)$.
- No θ_i can be included in $c\left(y_i,\phi_i\right)$.

Example: Poisson Distribution

- The pmf of a Poisson distribution $\text{Poisson}\left(\mu_i\right)$ is

$$P\left(Y_i = y_i\right) = \frac{\mu_i^{y_i}}{y_i!} \exp\left\{-\mu_i\right\} = \exp\left\{y_i \log\left(\mu_i\right) - \mu_i - \log\left(y_i!\right)\right\},$$

  which does not directly fit into the exponential dispersion family

$$f\left(y_i; \theta_i, \phi\right) = \exp\left\{\frac{y_i \theta_i - b\left(\theta_i\right)}{\phi_i} + c\left(y_i, \phi_i\right)\right\}.$$

- However, if we define $\theta_i = \log\left(\mu_i\right)$, then

$$P\left(Y_i = y_i\right) = \exp\left\{\frac{y_i \theta_i - \exp\left(\theta_i\right)}{1} - \log\left(y_i!\right)\right\}.$$

  Here $\phi_i = 1$, which is a constant.

Example: Binomial Distribution

- The pmf of a binomial distribution $\text{Bin}\left(n_i, \pi_i\right)$ with n_i being the total number of trials and π_i being the success probability is

$$
\begin{aligned}
P\left(Z_i = z_i\right) &= \binom{n_i}{z_i} \pi_i^{z_i}\left(1-\pi_i\right)^{n_i - z_i} \\
&= \exp\left\{ z_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log\left(1-\pi_i\right) + \log\binom{n_i}{z_i} \right\},
\end{aligned}
$$

  whose expectation depends on n_i.

- Define $\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$ and consider $Y_i = Z_i / n_i$, then

$$
P\left(Y_i = y_i\right) = \exp\left\{ \frac{y_i\theta_i - \log\left[1+\exp\left(\theta_i\right)\right]}{1/n_i} + \log\binom{n_i}{n_i y_i} \right\}.
$$

  Here $\phi_i = \phi / w_i$ with $\phi = 1$ and $w_i = n_i$.

Moments of Exponential Family

For the exponential dispersion family,

$$\begin{aligned} \mathbb{E}\left(Y_i\right) &= b'\left(\theta_i\right), \\ \operatorname{var}\left(Y_i\right) &= \phi_i b''\left(\theta_i\right), \end{aligned}$$

where $V\left(\theta_i\right) = b''\left(\theta_i\right)$ is called the variance function.

Components of Generalized Linear Model

1. Random component: Response variable Y_i and its probability distribution from exponential dispersion family.
2. Linear predictor $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$: Model matrix $\boldsymbol{X}$ of size $n \times p$ and parameter vector $\boldsymbol{\beta}$ of size $p \times 1$. The linear predictor for y_i is

$$\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} = \sum_{j=1}^{p} x_{ij}\beta_j,$$

   where $\boldsymbol{x}_i^T$ is the ith row of $\boldsymbol{X}$.
3. Link function $g\left(\right)$: $g\left(\right)$ transforms $\mu_i = \mathbb{E}\left(Y_i\right)$ to the linear predictor

$$g\left(\mu_i\right) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

   The link function must be monotonic and differentiable.

Examples of Link Functions

Suppose that Y_i follows a Bernoulli distribution $(n_i = 1)$ or a binomial distribution $(n_i \neq 1)$. The most common link function is the logit link (logistic model or logit model):

$$g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right).$$

Suppose that Y_i follows a Poisson distribution. The link function is often the log-link $g(\mu) = \log \mu$.

Everything is Connected

A GLM transforms μ_i through the link function $g\left(\mu_i\right) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$.

θ_i, μ_i, η_i, $\boldsymbol{\beta}$ are all connected through $b\left(\theta_i\right)$ and $g\left(\mu_i\right)$.

$$\begin{tikzcd}
\theta_i \arrow[r, Leftrightarrow, "\mu_i = b'\left(\theta_i\right)"] & \mu_i \arrow[r, Leftrightarrow, "\eta_i = g\left(\mu_i\right)"] & \eta_i \arrow[r, Leftarrow, "\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}"'] & \boldsymbol{\beta}
\end{tikzcd}$$

Suppose that $b\left(\theta_i\right) = \exp\left(\theta_i\right)$ and $g\left(\mu_i\right) = \mu_i^3$. Then,

$$\begin{tikzcd}
\theta_i \arrow[r, Leftrightarrow, "\mu_i = \exp\left(\theta_i\right)", "\theta_i = \log\left(\mu_i\right)"'] & \mu_i \arrow[r, Leftrightarrow, "\eta_i = \mu_i^3", "\mu_i = \eta_i^{1/3}"'] & \eta_i \arrow[r, Leftarrow, "\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}"'] & \boldsymbol{\beta}
\end{tikzcd}$$

Canonical Link

- The link function of a GLM transforms the mean of the random component to the linear predictor $\eta_i = g\left(\mu_i\right)$.
- The link function that transforms the mean μ_i to the natural parameter θ_i is called the canonical link.

$$\begin{aligned} \theta_i &= g\left(\mu_i\right) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}, && \text{canonical link,} \\ \theta_i &\neq g\left(\mu_i\right) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}, && \text{otherwise.} \end{aligned}$$

- For a Poisson distribution, the canonical link is the log link.
- For a binomial distribution, the canonical link is the logit link.

Likelihood in Exponential Family

- For n independent observations, the likelihood is the product of densities or mass functions:

$$\prod_{i=1}^{n} f\left(y_i;\theta_i,\phi_i\right) = \prod_{i=1}^{n} \exp\left\{\frac{y_i\theta_i - b\left(\theta_i\right)}{\phi_i} + c\left(y_i,\phi_i\right)\right\}.$$

- The log-likelihood is

$$\sum_{i=1}^{n}\left\{\frac{y_i\theta_i - b\left(\theta_i\right)}{\phi_i} + c\left(y_i,\phi_i\right)\right\}.$$

The log-likelihood will be denoted by $\ell\left(\boldsymbol{\mu};\boldsymbol{y}\right)$, where the ith entry of $\boldsymbol{\mu}$ is $\mu_i = \mathbb{E}\left(Y_i\right)$ and the ith entry of $\boldsymbol{y}$ is y_i.

Maximum Likelihood Estimator

Since $g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$ and $\mu_i = \mathbb{E}(Y_i) = b'(\theta_i)$, θ_i is a function of $\boldsymbol{\beta}$. We can maximize the log-likelihood

$$\ell = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i) \right\}$$

to obtain the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}$.

The gradient be expressed as

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T \boldsymbol{D} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}),$$

where $\boldsymbol{X}_{n \times p}$ is the model matrix, $\boldsymbol{D}_{n \times n}$ is the diagonal matrix with (i,i)th element $\partial \mu_i / \partial \eta_i$, and $\boldsymbol{V}_{n \times n}$ is a diagonal matrix with (i,i)th element $\mathrm{var}(Y_i)$.

Example: Find Score Function

Gradient of Poisson regression

Consider the Poisson regression model, where $Y_i \sim \text{Poisson}\left(\mu_i\right)$ and $\log\left(\mu_i\right) = \eta_i = \beta_1 + \beta_2 x_i$. Show that

$$\begin{aligned} \boldsymbol{D} &= \text{diag}\left\{\exp\left(\beta_1 + \beta_2 x_i\right)\right\}, \\ \boldsymbol{V} &= \text{diag}\left\{\exp\left(\beta_1 + \beta_2 x_i\right)\right\}. \end{aligned}$$

General Problem

- Consider a general problem that, for a scalar-valued function $h(\boldsymbol{\beta})$, we need to find the solution of

$$\mathbf{0} = \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}.$$

- The solution is approximately the solution of

$$\mathbf{0} = \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \approx \frac{\partial h\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta}} + \frac{\partial^2 h\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}\right)$$

  for some known $\boldsymbol{\beta}^{(t)}$, which yields

$$\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)} - \left(\frac{\partial^2 h\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right)^{-1} \frac{\partial h\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta}},$$

  if the Hessian matrix is invertible.

Newton-Raphson Method or Newton's Method

We can name a first guess of $\boldsymbol{\beta}$, $\boldsymbol{\beta}^{(0)}$, and update parameter estimates using

$$\begin{aligned}
\boldsymbol{\beta}^{(1)} &\approx \boldsymbol{\beta}^{(0)} - \left(\frac{\partial^2 h\left(\boldsymbol{\beta}^{(0)}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)^{-1} \frac{\partial h\left(\boldsymbol{\beta}^{(0)}\right)}{\partial\boldsymbol{\beta}}, \\
\boldsymbol{\beta}^{(2)} &\approx \boldsymbol{\beta}^{(1)} - \left(\frac{\partial^2 h\left(\boldsymbol{\beta}^{(1)}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)^{-1} \frac{\partial h\left(\boldsymbol{\beta}^{(1)}\right)}{\partial\boldsymbol{\beta}}, \\
&\vdots
\end{aligned}$$

until $\frac{\partial h\left(\boldsymbol{\beta}^{(t+1)}\right)}{\partial\boldsymbol{\beta}}$ is sufficiently close to 0 or $\boldsymbol{\beta}^{(t+1)}$ and $\boldsymbol{\beta}^{(t)}$ are sufficiently close.

Newton-Raphson in GLM

In GLM, we need to find the solution of

$$\mathbf{0} = \frac{\partial \ell\left(\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T \boldsymbol{D} \boldsymbol{V}^{-1} \left(\boldsymbol{y} - \boldsymbol{\mu}\right).$$

The Newton-Raphson in GLM updates the parameter estimator as

$$\begin{aligned} \boldsymbol{\beta}^{(t+1)} &= \boldsymbol{\beta}^{(t)} - \left(\frac{\partial^2 \ell\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right)^{-1} \frac{\partial \ell\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta}} \\ &= \boldsymbol{\beta}^{(t)} + \left(-\frac{\partial^2 \ell\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right)^{-1} \frac{\partial \ell\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta}} \end{aligned}$$

until convergence. Here, we are taking the inverse of the observed information matrix.

Newton-Raphson to Fisher Scoring

- The Newton-Raphson method updates the parameter estimator as

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left(-\frac{\partial^2 \ell\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right)^{-1} \frac{\partial \ell\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta}}.$$

- The Fisher scoring updates the parameter estimator as

$$\begin{aligned} \boldsymbol{\beta}^{(t+1)} &= \boldsymbol{\beta}^{(t)} + \left[E\left(-\frac{\partial^2 \ell\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right)\right]^{-1} \frac{\partial \ell\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta}} \\ &= \boldsymbol{\beta}^{(t)} + \left[\boldsymbol{\mathcal{I}}\left(\boldsymbol{\beta}^{(t)}\right)\right]^{-1} \frac{\partial \ell\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta}}, \end{aligned}$$

where $\boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}$ for GLM with $\boldsymbol{W} = \boldsymbol{D} \boldsymbol{V}^{-1} \boldsymbol{D}$.

Iterative Reweighted Least Squares

- Plugging in the expression of information matrix and score function, Fisher scoring becomes

$$\begin{aligned}\boldsymbol{\beta}^{(t+1)} &= \left(\boldsymbol{X}^T\boldsymbol{W}^{(t)}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(t)}\left[\boldsymbol{X}\boldsymbol{\beta}^{(t)} + \left(\boldsymbol{D}^{(t)}\right)^{-1}\left(\boldsymbol{y}-\boldsymbol{\mu}^{(t)}\right)\right]\\ &= \left(\boldsymbol{X}^T\boldsymbol{W}^{(t)}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(t)}\boldsymbol{z}^{(t)}.\end{aligned}$$

- This means that, at each step, $\boldsymbol{\beta}$ is updated using weighted least squares with closed forms using the adjusted response variable $\boldsymbol{z}^{(t)}$.
- In other words, for GLM, estimators are obtained by an iterative reweighted least squares (IRLS) procedure.

Biproduct: Standard Error

- The IRLS procedure updates the parameter estimates by

$$\boldsymbol{\beta}^{(t+1)} = \left(\boldsymbol{X}^T \boldsymbol{W}^{(t)} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}^{(t)} \left[\boldsymbol{X}\boldsymbol{\beta}^{(t)} + \left(\boldsymbol{D}^{(t)}\right)^{-1} \left(\boldsymbol{y} - \boldsymbol{\mu}^{(t)}\right)\right].$$

- If n is large enough and all assumptions are correct, the distribution of $\hat{\boldsymbol{\beta}}$ can be approximated by

$$N\left(\boldsymbol{\beta},\ \left(\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X}\right)^{-1}\right).$$

  where $\hat{\boldsymbol{W}}$ is the latest $\boldsymbol{W}$ from IRLS.

- The standard error of $\hat{\beta}_j$ can be approximated by $\sqrt{\hat{\tau}_j}$, where $\hat{\tau}_j$ is the (j,j)th element of $\left(\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X}\right)^{-1}$.

Prediction

Once we have obtained $\hat{\boldsymbol{\beta}}$, we can predict η by $\hat{\eta} = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}$ and μ by $\hat{\mu} = g^{-1}\left(\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}\right)$, where $\boldsymbol{x}_0$ is the vector of regressors/features, and $g^{-1}()$ is the inverse function of $g()$.

- The distribution of $\hat{\eta} = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}$ is then approximately

$$N\left(\boldsymbol{\eta},\ \boldsymbol{x}_0^T \left(\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X}\right)^{-1} \boldsymbol{x}_0\right).$$

- A $1-\alpha$ confidence interval for η is

$$\hat{\eta} \pm z_{1-\alpha/2} \sqrt{\boldsymbol{x}_0^T \left(\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X}\right)^{-1} \boldsymbol{x}_0},$$

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

$$g^{-1}\left(\hat{\eta} \pm z_{1-\alpha/2} \sqrt{\boldsymbol{x}_0^T \left(\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X}\right)^{-1} \boldsymbol{x}_0}\right).$$

Maximum log-Likelihood of Our Model

- Given $\hat{\boldsymbol{\beta}}$, the fitted μ_i is $\hat{\mu}_i = g^{-1}\left(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}\right)$, where $g^{-1}()$ is the inverse function of $g()$.
- The fitted θ_i, denoted by $\hat{\theta}_i$, is the solution of $\hat{\mu}_i = b'\left(\hat{\theta}_i\right)$.
- The likelihood of our model becomes

$$L(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) \equiv \prod_{i=1}^{n} \exp\left\{\frac{y_i\hat{\theta}_i - b\left(\hat{\theta}_i\right)}{\phi_i} + c(y_i, \phi_i)\right\}.$$

Saturated Model

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

- Since $\mu_i = b'(\theta_i)$, the fitted θ_i is the solution of $\hat{\mu}_i = b'\left(\hat{\theta}_i\right)$.
- NOTE: there is no $\boldsymbol{\beta}$ directly involved here.

The likelihood of the saturated model is

$$L(\boldsymbol{y};\boldsymbol{y}) \equiv \prod_{i=1}^{n} \exp\left\{\frac{y_i\hat{\theta}_i^{(s)} - b\left(\hat{\theta}_i^{(s)}\right)}{\phi_i} + c(y_i,\phi_i)\right\},$$

where the superscript denotes that it is the saturated model.

(Residual) Deviance

Consider testing

$$\begin{aligned} H_0 &: \text{ The model fits the data as good as the saturated model} \\ H_1 &: \text{ The model fits the data worse than the saturated model} \end{aligned}$$

The likelihood ratio test statistic is $-2\log\left(\frac{L(\hat{\boldsymbol{\mu}};\boldsymbol{y})}{L(\boldsymbol{y};\boldsymbol{y})}\right)$.

In Poisson GLM or binomial GLM, the (residual) deviance is

$$D\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right) = -2\log\left(\frac{L\left(\hat{\boldsymbol{\mu}};\boldsymbol{y}\right)}{L\left(\boldsymbol{y};\boldsymbol{y}\right)}\right),$$

where ϕ_i is known in both models. If the model fits the data well, $D\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right) \approx \chi^2\left(m-p\right)$, where *m is the number of parameters in the saturated model*, p is the number of parameters in the model of interest, and m should not increases as n increases.

Example: Deviance for Binomial model

In a binomial model,

$$P\left(Y_i = y_i\right) = \left(\begin{array}{c} n_i \\ y_i \end{array}\right) \pi_i^{y_i} \left(1 - \pi_i\right)^{n_i - y_i}.$$

Our model yields predicted probability $\hat{\pi}_i$. Hence, the deviance is

$$D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}\right) = -2 \log \left(\frac{\prod_{i=1}^{n} \left(\begin{array}{c} n_i \\ y_i \end{array}\right) \hat{\pi}_i^{y_i} \left(1 - \hat{\pi}_i\right)^{n_i - y_i}}{\prod_{i=1}^{n} \left(\begin{array}{c} n_i \\ y_i \end{array}\right) y_i^{y_i} \left(1 - y_i\right)^{n_i - y_i}} \right).$$

Grouped Data and Ungrouped Data

```
#### Ungrouped data
Ungroup

##     y x1         x2
## 1  0  0  0.8458632
## 2  0  0  0.6726630
## 3  1  0 -0.4372080
## 4  0  0 -1.4194868
## 5  1  0  0.8742662
## 6  1  1 -0.7330018
## 7  1  1 -0.8285645
## 8  0  1 -0.2341681
## 9  0  1  0.5203699
## 10 1  1  0.1571108
## 11 0  1  0.2665822
## 12 0  1  0.2124662
```

```
#### 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
```

Grouped Data Expressed as Ungrouped

```
##
## Call:
## glm(formula = y ~ x1 + x2, family = binomial(), data = DF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2310  -0.9793  -0.8850   1.1513   1.5585
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.1251     1.0238  -0.122    0.903
## x1            0.2502     1.2310   0.203    0.839
## x2           -0.7372     1.2141  -0.607    0.544
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 16.301  on 11  degrees of freedom
## Residual deviance: 15.914  on  9  degrees of freedom
## AIC: 21.914
##
```

Grouped Data Expressed as Ungrouped

```
##
## Call:
## glm(formula = cbind(success, fail) ~ x1 + x2, family = binomial(),
##     data = NewDF)
##
## Deviance Residuals:
##       1        2        3        4
## -0.4758   0.6007   0.4758  -0.4373
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.1251     1.0238  -0.122    0.903
## x1            0.2502     1.2310   0.203    0.839
## x2           -0.7372     1.2141  -0.607    0.544
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.3912  on 3  degrees of freedom
## Residual deviance: 1.0049  on 1  degrees of freedom
## AIC: 12.261
```

Null Model and Null Deviance

- Consider a special model where only the intercept is included

$$g\left(\mu_i\right) = \beta_0,$$

  with $p = 1$.
- The fitted mean for individual i is $\hat{\mu}_i = g^{-1}\left(\beta_0\right)$, which is the same for all i.
- The estimator of θ_i is obtained from $\hat{\mu}_i = b'\left(\theta_i\right)$, still the same for all i.
- This is called a null model and its residual deviance is called the null deviance.
  - The null model represents the worst model that we can build.
  - The null deviance compares the null model with the saturated model.

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

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

- In binomial GLM or Poisson GLM, the difference in the deviance is

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

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.

AIC: Minimizing Distance of the Fit from the Truth

- The Akaike information criterion (AIC) is a nearly “unbiased” estimator of the “distance” between the assumed model and the unknown truth.
- It is a penalized log-likelihood

$$\mathrm{AIC} = -2\ell\left(\hat{\boldsymbol{\beta}}_M\right) + 2 \cdot \text{number of parameters in model } M.$$

- AIC is NOT

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i - \mu\right)^2 + 2 \cdot \text{number of parameters in model } M.$$

- We prefer to model with the smallest AIC or a parsimonious model that has AIC near the minimum.
- In practice, AIC tends to be conservative, in the sense that it tends to select more explanatory variables.

BIC: Consistent Model Selection

- Bayesian information criterion penalizes a complex model much more than AIC.

$$\mathrm{BIC} = -2\ell\left(\hat{\boldsymbol{\beta}}_M\right) + \log\left(n\right) \cdot \text{number of parameters in model } M.$$

- We prefer to model with the smallest BIC or a parsimonious model that has BIC near the minimum.
- BIC is consistent in model selection in the sense that

$$P\left(\text{Choose the true model if it is a candidate}\right) \to 1, \text{ as } n \to \infty.$$

- In contrast, AIC is not consistent.

Pearson Residual and Deviance Residual

- The Pearson residual is

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

- The deviance residual is

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

  whose squared values sum to the deviance.

It is often suggested to perform residual checking to investigate whether any patterns can be observed for grouped data. If the model fits data well, we should not observe any trends.

Unfortunately...

Unfortunately, the residual plots for models fitted by glm() may not be useful, when we have ungrouped data.

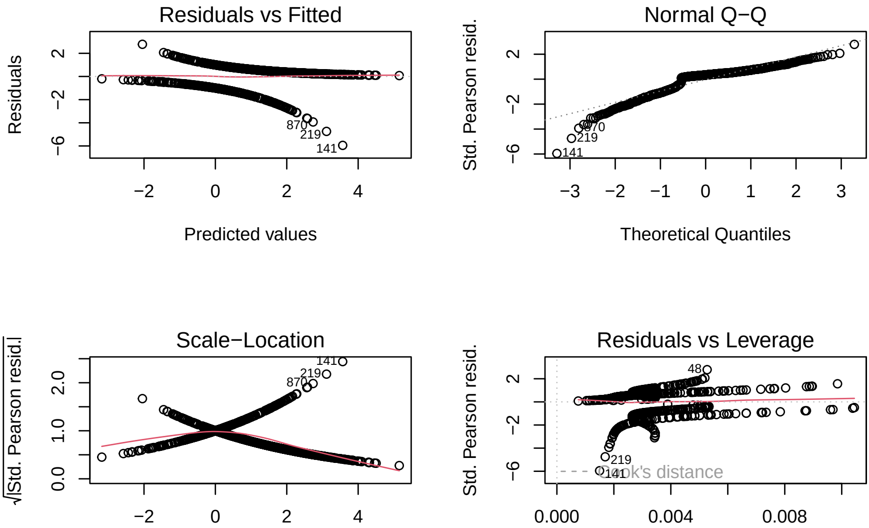
- We generate binary data from a logistic model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \beta_x x_2,$$

  where x_1 takes values 0 or 1, and x_2 is continuous.
- We fit the model using

```
logit <- glm(y ~ x1 + x2, data = Data, family = binomial())
```

Residual Plots

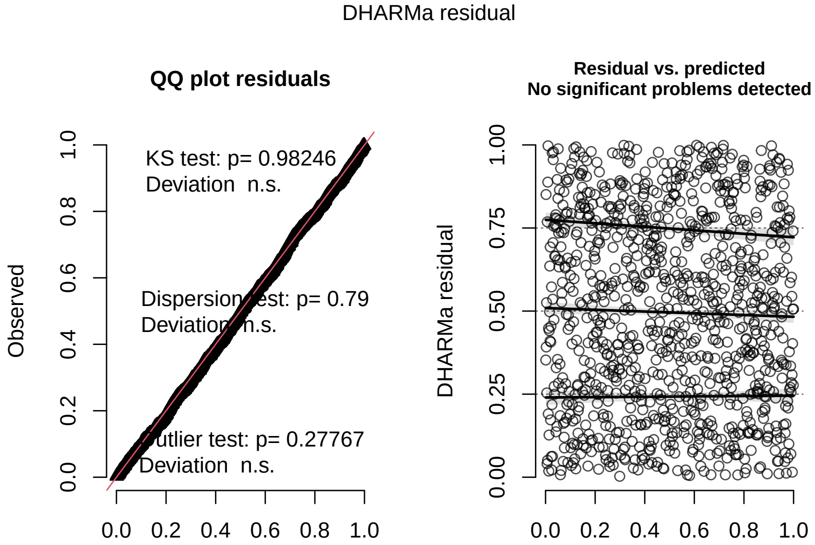


Alternative: Randomized Quantile Residuals

1. Fit your model using glm() or other functions
2. Simulate (randomized) quantile residuals using simulateResiduals()
    1. First, for each observation i, simulate q response variables using the predicted μ_i.
    2. Second, for each observation i, compute the percentage that simulated response less than y_i and the percentage that simulated response less than or equal to y_i.
    3. Third, 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.
3. Plot the (randomized) quantile residuals using plot().

If your model is correct, the cdf of y_i follows a uniform distribution. Hence, we expect the quantile residuals to be uniform and spread out everywhere.

Randomized Quantile Residual Plots


DHARMa residual
QQ plot residuals
KS test: p= 0.98246
Deviation n.s.
Dispersion test: p= 0.79
Deviation n.s.
Outlier test: p= 0.27767
Deviation n.s.
Observed
Residual vs. predicted
No significant problems detected
DHARMa residual