STAT347: Generalized Linear Models

Lecture 4

Winter, 2024

Jingshu Wang

Today's topics:

- Deviance analysis
- Model checking with the residuals
- Example: Building a GLM
- Reading: Agresti Chapters 4.4, 4.7, Faraway Chapters 8.3-8.4

Deviance analysis in GLM

```
## Call:
## glm(formula = y ~ weight + factor(color), family = poisson(),
##     data = Crabs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9833  -1.9272  -0.5553   0.8646   4.8270
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.04978    0.23315  -0.214   0.8309
## weight          0.54618    0.06811   8.019 1.07e-15 ***
## factor(color)2 -0.20511    0.15371  -1.334   0.1821
## factor(color)3 -0.44980    0.17574  -2.560   0.0105 *
## factor(color)4 -0.45205    0.20844  -2.169   0.0301 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 551.80  on 168  degrees of freedom
## AIC: 917.1
##
## Number of Fisher Scoring iterations: 6
```

$\sum_i (y_i - \bar{y})^2$

$\sum_i (y_i - \hat{\mu}_i)^2$

- In linear regression, we use

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{\mu}_i)^2}{\sum_i (y_i - \bar{y})^2} = \frac{\sum_i (\hat{\mu}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

To evaluate how well the model fits the data. We have an analogy in GLM, which is the deviance analysis.

Definition of deviance

Consider density function $f(y;\theta) = e^{\frac{y\theta - b(\theta)}{a(\phi)}} f_0(y;\phi)$ at two values θ_1 and θ_2. Measure the "distance" between two distributions:

$$D(\theta_1, \theta_2) = 2\mathbb{E}_{\theta_1}\left\{\log \frac{f(y;\theta_1)}{f(y;\theta_2)}\right\} = 2\mathbb{E}_{\theta_1}\left\{y(\theta_1 - \theta_2) - b(\theta_1) + b(\theta_2)\right\}/a(\phi)$$
$$= 2\left[\mu_1(\theta_1 - \theta_2) - b(\theta_1) + b(\theta_2)\right]/a(\phi)$$

Remember the 1-to-1 mapping between μ and θ, we also write $D(\mu_1, \mu_2) = D(\theta_{\mu_1}, \theta_{\mu_2})$

- $D(\mu_1, \mu_2) \geq 0$ and the equality holds only when $\mu_1 = \mu_2$
- Generally, $D(\mu_1, \mu_2) \neq D(\mu_2, \mu_1)$
- KL divergence: $D(\mu_1, \mu_2)/2$
- If f is the normal density, then $D(\mu_1, \mu_2) = (\mu_1 - \mu_2)^2/\sigma^2$

Residual deviance

- Saturated model: imagine the case that we collect an infinite number of covariates, then we can perfectly fit the data and obtain $\hat{\mu}_i = y_i$ for all samples.

- For a particular sample i, Deviance between the saturated model $\hat{\mu}_i = y_i$ and another model with μ_i (corresponding canonical parameter θ_i)

$$D(\theta_1, \theta_2) = 2\mathbb{E}_{\theta_1}\left\{\log \frac{f(y;\theta_1)}{f(y;\theta_2)}\right\} = 2\mathbb{E}_{\theta_1}\left\{y(\theta_1 - \theta_2) - b(\theta_1) + b(\theta_2)\right\}/a(\phi)$$

$$= 2\left[\mu_1(\theta_1 - \theta_2) - b(\theta_1) + b(\theta_2)\right]/a(\phi)$$

$$D(y_i, \mu_i) = \frac{2\left[y_i\left(\theta_{y_i} - \theta_i\right) - b\left(\theta_{y_i}\right) + b(\theta_i)\right]}{a(\phi)}$$

$$= -2\log[f(y_i, \theta_i)/f(y_i, \theta_{y_i})]$$

- $\theta_{y_i} = (b')^{-1}(y_i)$ [As $\mu_i = b'(\theta_i)$]

Residual deviance

- Residual deviance (total deviance):
  deviance between the fitted saturated model and the proposed model

$$D_+(y,\hat{\mu}) = \sum_i D(y_i,\hat{\mu}_i)$$

$$= -2\sum_i \log\left[f(y_i,\hat{\theta}_i)/f(y_i,\theta_{y_i})\right]$$

  - $\theta_{y_i} = (b')^{-1}(y_i)$
- Example: for Gaussian linear model $D_+(y,\hat{\mu}) = \sum_i (y_i - \hat{\mu}_i)^2/\sigma^2$

Null deviance

- Null model: the linear model that only includes intercept. Thus,

$$\mu_i \equiv \mu$$

- MLE estimate of μ from the null model will be $\hat{\mu} = \bar{y} = \sum_i y_i / n$

- Null deviance: deviance between the fitted saturated model and the null model

$$\sum_i D(y_i, \bar{y})$$

- "R^2" in GLM:

$$1 - \frac{D_+(y, \hat{\mu})}{\sum_i D(y_i, \bar{y})}$$

Deviance analysis for nested models

Let $\beta = \begin{pmatrix} \beta^{(1)} \\ \beta^{(2)} \end{pmatrix}$ where $\beta^{(1)} \in \mathbb{R}^{p_1}$ and $X = \begin{pmatrix} X^{(1)} & X^{(2)} \end{pmatrix}$.

We call $\mathcal{M}^{(1)}$ with

$$g(\mu_i) = X^{(1)}\beta^{(1)}$$

a nested model of the full model $\mathcal{M}$ where

$$g(\mu_i) = X\beta.$$

- Test for whether the nested model is already enough:

$$H_0 \colon \beta^{(2)} = 0$$

Deviance analysis for nested models

Let $\beta = \begin{pmatrix} \beta^{(1)} \\ \beta^{(2)} \end{pmatrix}$ where $\beta^{(1)} \in \mathbb{R}^{p_1}$ and $X = \begin{pmatrix} X^{(1)} & X^{(2)} \end{pmatrix}$.

We call $\mathcal{M}^{(1)}$ with

$$g(\mu_i) = X^{(1)}\beta^{(1)}$$

a nested model of the full model $\mathcal{M}$ where

$$g(\mu_i) = X\beta.$$

Let $\hat{\beta}^{(1)}$ be the MLE solution of the model $\mathcal{M}^{(1)}$ and $\hat{\mu}^{(1)}$ be the corresponding estimated expectations of y in the fitted model.

Then,

$$D_+(\hat{\mu}, \hat{\mu}^{(1)}) = D_+(y, \hat{\mu}^{(1)}) - D_+(y, \hat{\mu}) = -2\left[L(\hat{\beta}^{(1)}) - L(\hat{\beta})\right]$$

Deviance analysis for nested models

Let $\hat{\beta}^{(1)}$ be the MLE solution of the model $\mathcal{M}^{(1)}$ and $\hat{\mu}^{(1)}$ be the corresponding estimated expectations of y in the fitted model.

$$D_+(\hat{\mu}, \hat{\mu}^{(1)}) = D_+(y, \hat{\mu}^{(1)}) - D_+(y, \hat{\mu}) = -2\left[L(\hat{\beta}^{(1)}) - L(\hat{\beta})\right]$$

- Deviance additivity theorem (Efron, Annals of Statistics 1978)
- This is the likelihood ratio between the full and nested models

- Likelihood ratio test:
  If both p and p_1 are fixed, then asymptotically under $H_0\colon \beta^{(2)} = 0$

$$D_+(y, \hat{\mu}^{(1)}) - D_+(y, \hat{\mu}) \to \mathcal{X}^2_{p-p_1}$$

Deviance analysis table for model comparisons

Say we partition our covariates as

$$X = (1, X_{(1)}, X_{(2)}, \cdots, X_{(J)})$$

and $X_{(j)} \in \mathbb{R}^{d_j}$. We can sequentially add each partition of covariates into the model (in some pre-determined order) and understand each partition's "relative contribution" with a deviance analysis table.

- $\hat{\beta}^{(j)}$ is the MLE solution of the GLM model with covariates $X^{(j)} = (1, X_{(1)}, X_{(2)}, \cdots, X_{(j)})$
- $\hat{\mu}^{(j)}$ is the corresponding vector of expectations of $y = (y_1, \cdots, y_n)$ in the fitted model.

Deviance analysis table in R

Model	twice log-likelihood	residual deviance	difference	df	Compare with
$\hat{\beta}^{(0)}$ (null)	$2L(\hat{\beta}^{(0)})$	$D_+(y, \hat{\mu}^{(0)}) = \sum_i D(y_i, \bar{y})$			
$\hat{\beta}^{(1)}$	$2L(\hat{\beta}^{(1)})$	$D_+(y, \hat{\mu}^{(1)})$	$D_+(y, \hat{\mu}^{(0)}) - D_+(y, \hat{\mu}^{(1)})$	d_1	$\chi^2_{d_1}$
$\hat{\beta}^{(2)}$	$2L(\hat{\beta}^{(2)})$	$D_+(y, \hat{\mu}^{(2)})$	$D_+(y, \hat{\mu}^{(1)}) - D_+(y, \hat{\mu}^{(2)})$	d_2	$\chi^2_{d_2}$
$\vdots$					
$\hat{\beta}^{(J)}$	$2L(\hat{\beta}^{(J)})$	$D_+(y, \hat{\mu}^{(J)})$	$D_+(y, \hat{\mu}^{(J-1)}) - D_+(y, \hat{\mu}^{(J)})$	d_J	$\chi^2_{d_J}$

- Add variables sequentially to check if larger models are necessary
- Similar to the analysis of variable table in linear regression
- Typically the full model can not be the saturated model as df in a saturated model is too large

Deviance analysis table

- R output for the election counts example in Lecture 1

```
> result.glm <- glm(cbind(undercountNumber, votes) ~ pergore + factor(rural) + factor(econ) +
  factor(atlanta) + factor(equip), data = gavote, family = "binomial")
> anova(result.glm, test = "LRT")
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(undercountNumber, votes)

Terms added sequentially (first to last)


                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                              158      36829
pergore          1   5031.0       157      31798 < 2.2e-16 ***
factor(rural)    1   4197.2       156      27601 < 2.2e-16 ***
factor(econ)     2   7248.1       154      20353 < 2.2e-16 ***
factor(atlanta)  1    534.6       153      19818 < 2.2e-16 ***
factor(equip)    4   4150.5       149      15668 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

equip: the voting method, takes five values "LEVER", "OS-CC" (optimal scan, central count), "OS-PC" (optimal scan, precinct count), "Paper","PUNCH" (punch card)
econ: the economic level of the county, takes three values "middle", "poor" and "rich"
perAA: the percentage of African Americans
rural: whether the county is rural or urban
atlanta: whether the county is part of the Atlanta metropolitan area
gore: number of votes for Al Gore
bush: number of votes for George Bush
other: number of votes for other candidates
votes: total vote counts
ballots: number of ballots issued

This analysis is reliable only when model assumptions for each corresponding null hold

Model checking with the residuals

- As in the linear models, we can examine the residuals to help us check whether a model fits poor or not, and whether there are any outliers in the observations.

- Three types of residuals
  - Pearson residual

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)}} \qquad v(\hat{\mu}_i) = \widehat{\mathrm{Var}}(y_i)$$

  - Standardized residual (similar as in linear regression)

$$r_i = \frac{e_i}{\sqrt{1 - \hat{h}_{ii}}}$$

where h_{ii} is the ith diagonal element of the H_W defined equation (4.19) of the Agresti chapter 4.4.5.

Model checking with the residuals

- Three types of residuals
  - Pearson residual

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)}} \qquad v(\hat{\mu}_i) = \widehat{\mathrm{Var}}(y_i)$$

  - Standardized residual (similar as in linear regression)

$$r_i = \frac{e_i}{\sqrt{1 - \hat{h}_{ii}}}$$

where h_{ii} is the ith diagonal element of the H_W defined equation (4.19) of the Agresti chapter 4.4.5.

  - Deviance residual

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

Residuals examples

- For Gaussian linear model
  - Pearson residual

$$e_i = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}}$$

  - Deviance residual

$$d_i = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}} = e_i$$

Some intuition related to deviance residuals

- Deviance residuals are considered more "normal" than Pearson residuals

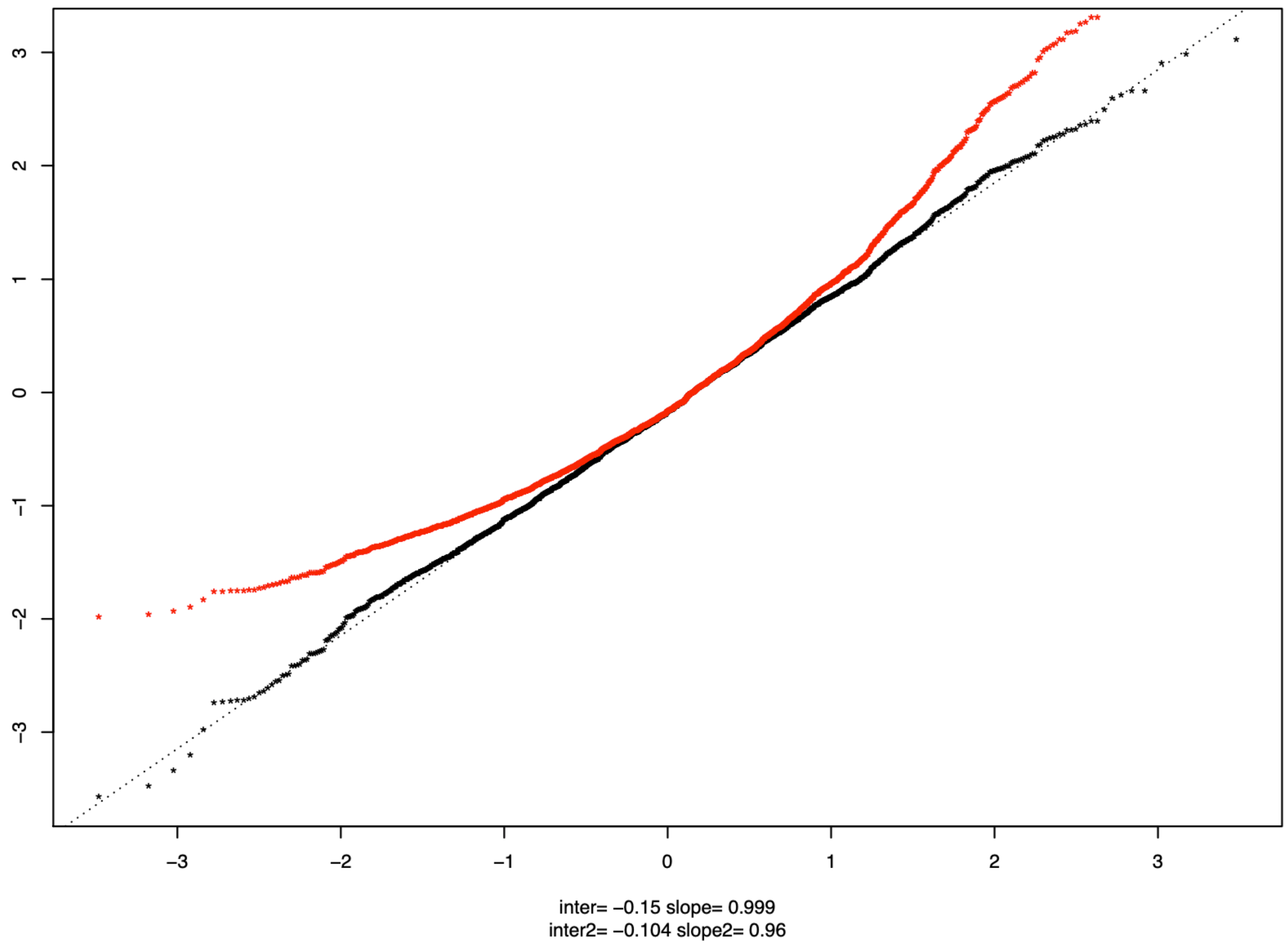
- Consider deviance residual of i.i.d samples

$$R = \text{sign}(\bar{y} - \mu)\sqrt{D(\bar{y}, \mu)}.$$

- It has been shown that R converges to $N(0,1)$ when sample size $n \to \infty$, and has better third order accuracy than corresponding Pearson residuals

- You can check Appendix C of McCullagh and Nelder, *Generalized Linear Models* for more math details

Some intuition related to deviance residuals



qq comparison of deviance residuals (black) with Pearson residuals (red); Gamma distribution $k = 1, \theta = 1, n = 5$; B = 2000 simulations.

Example: Building a GLM

- Check Example2 R notebook