STAT 528 - Advanced Regression Analysis II

Generalized Linear Models Diagnostics

Daniel J. Eck (with credit to Lu Yang)
Department of Statistics
University of Illinois

Learning Objectives Today

- GLM diagnostics using Lu Yang's Quasi-Empirical Residual Distribution Function method (paper in same directory as these slides).

Residuals

Recall Pearson and deviance residuals:

Pearson residuals

$$\hat{e}_{Pi} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

where $X^2 = \sum_i \hat{e}_{Pi}$.

Let deviance $D = \sum_i d_i$, deviance residuals

$$\hat{e}_{Di} = \mathrm{sign}\left(y_i - \hat{\mu}_i\right)\sqrt{d_i}$$

where $D = \sum_{i=1}^{n} \hat{e}_{Di}^2$.

Potential remedies addressed by residual diagnostics

Is there any nonlinear relationship between the predicted values and the residuals?
- A change link function - A change in the choice of predictors or transformations on these predictors

The assumptions of the GLM would require constant variance in the plot - A change in the variance function, quasi-likelihood GLM

Potential problems

In some cases, plots of the residuals are not particularly helpful.

For a binary response, the residual can only take two possible values for given predicted response. This is the most extreme situation, but similar discreteness can occur for binomial responses with small group sizes and Poisson responses that are small.

Small mean discrete outcomes: LGPIF data

We present an application of the proposed quasi-empirical residual distribution function to insurance claim frequency data.

Frequency, the number of reported claims from each policyholder, is an important component of insurance claim data and largely reveals the riskiness of a policyholder.

Here, we use a dataset from the Local Government Property Insurance Fund (LGPIF) in the state of Wisconsin, USA.

The LGPIF was established by Wisconsin government to provide property insurance for local government entities.

We focus on building and contents (BC) insurance, which is the major coverage offered by the LGPIF. The dataset contains 5660 observations from year 2006 to 2010

The frequency of claims is depicted below:

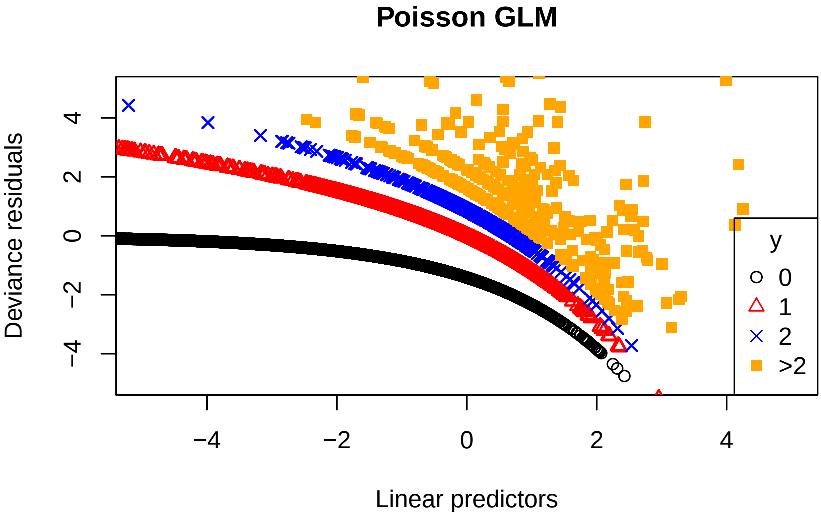
```r
freqinBC <- readRDS("freqinBC.rds")
table(freqinBC$FreqBC)
```

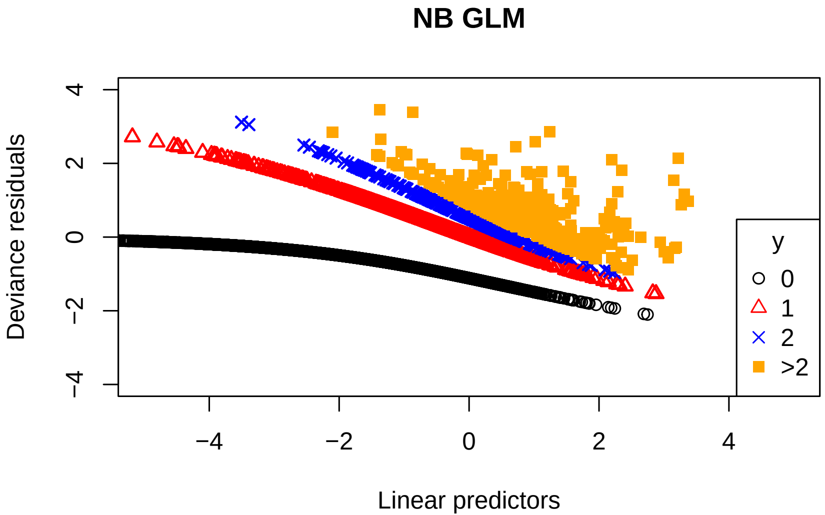
```
##
##    0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 3976  997  333  136   76   31   19   19   16    5    7    2    4    5    5    2
##   16   17   18   20   22   23   24   33   34   53   55   65   78   86   89   97
##    4    3    1    1    1    1    1    1    1    1    1    1    1    1    1    1
##   98  108  133  154  201  231
##    1    1    1    1    1    1
```

Poisson and negative binomial regression models are fit to this data

```r
freqmodelBC <- glm(FreqBC ~ lnCoverageBC + lnDeductBC + NoClaimCreditBC +
                     TypeCity + TypeCounty + TypeMisc + TypeSchool + TypeTown,
                   data = freqinBC, family = "poisson")
library(MASS)
freqmodelBCnb <- glm.nb(FreqBC ~ lnCoverageBC + lnDeductBC + NoClaimCreditBC +
                     TypeCity + TypeCounty + TypeMisc + TypeSchool + TypeTown,
                   data = freqinBC)
```

- Plots of residuals in these small mean cases tend to show curved lines of points corresponding to the limited number of observed responses. Such artifacts can obscure the main purpose of the plot.




NB GLM
Deviance residuals
4
2
0
−2
−4
Linear predictors
−4
−2
0
2
4
y
○ 0
△ 1
× 2
■ >2

Residuals in GLM

Two desirable properties of an informative diagnostic tool

1. Proximity to null patterns under true models

2. Discrepancy with null patterns under misspecified models

Residuals for linear regression models

- Residuals $r_i = Y_i - x_i'\hat{\beta}$, normally distributed under correctly specified models

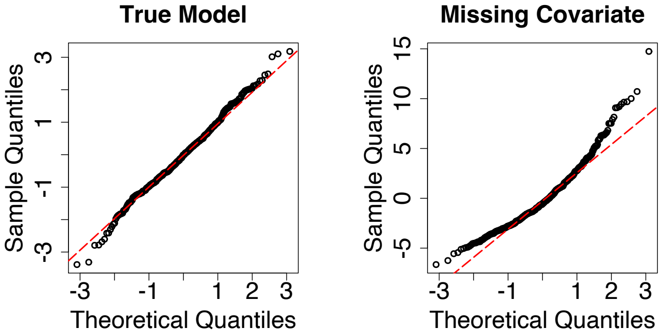


Figure 1: QQ plots for linear regression model residuals.

Residuals for linear regression models

- Features of residuals in linear regression models
  - Follow a known distribution under the correctly specified models
  - Nearly identically distributed
- Graphical diagnostics: QQ plot, PP plot, residuals versus predictor plot
  - Check normality assumption
  - Identify other important factors
  - etc.
- Construct overall goodness-of-fit tests using residuals

Beyond Normality: Cox and Snell (1968)

Linear Regression		Generalization
$e_i = Y_i - X_i'\beta$		$e_i = h(Y_i, X_i'\beta)$
$e_i \sim N(0, \sigma^2)$ i.i.d	$\Longrightarrow$	e_i i.i.d$\sim$known distribution
$r_i = Y_i - X_i'\hat{\beta}$		$R_i = h(Y_i, X_i'\hat{\beta})$
r_i are normally distributed under the true model		R_i follow a hypothesized pattern under the true model

Residuals for Continuous Outcomes

- For continuous variables Y_i, probability integral transform $F(Y_i|X_i, \beta) \sim \mathrm{Uniform}(0, 1)$
  - Gamma, inverse normal, lognormal distributions
- Cox-Snell residuals $F(Y_i|X_i, \hat{\beta}), i = 1, \ldots, n$ should be uniform with correctly specified models

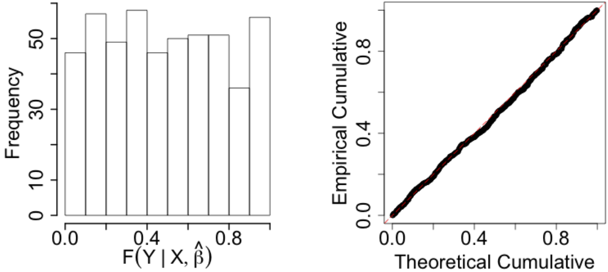


Figure 2: Histogram and PP plot of Cox-Snell residuals for a gamma example.

Commonly Used Residuals for Discrete Outcomes

- Discrete Y_i cannot be expressed as transformations of $X_i'\beta$ and i.i.d. errors so Cox-Snell residuals are not applicable

- Pearson and deviance residuals are approximately normal under a correctly specified model

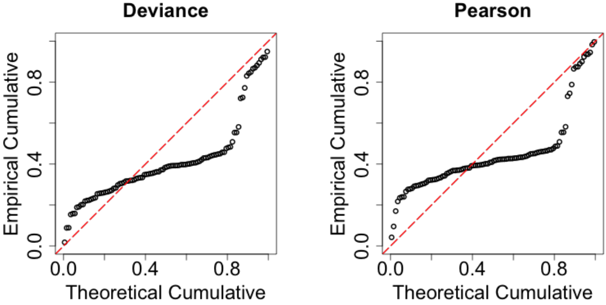


Figure 3: PP plots of residuals for a Poisson GLM under the **true model**.

m-Asymptotics

- m: the number of trials of binomial distributions, or the Poisson means, which controls the discreteness level
- m-asymptotics: deviance residuals are normally distributed with a discrepancy term of order at least $O_p(m^{-1/2})$ (Pierce and Schafer (1986))
- When m is small, deviance residuals and Pearson residuals could have a large discrepancy with the null pattern even under the true model, even with large n

Quasi-empirical residual distribution function (Yang (2021))

The Quasi-empirical residual distribution function, $\hat{U}(\cdot)$, should be close to the identity function under true model

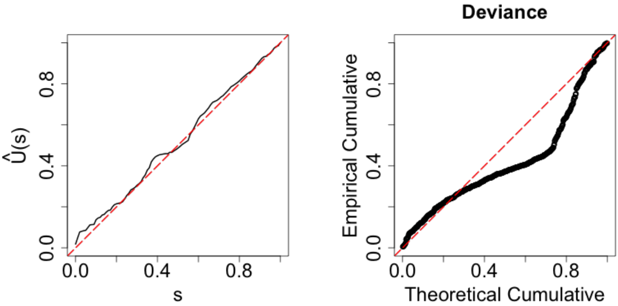


Figure 4: $\hat{U}$ and deviance residuals for a Poisson example under the true model.

Quasi-empirical residual distribution function

If Y is continuous, for any fixed value $s \in (0, 1)$,

$$\Pr(F(Y|\mathbf{X} = \mathbf{x}) \leq s) = s. \tag{1}$$

Conditioning on $\mathbf{X} = \mathbf{x}$, (1) holds for discrete Y if and only if $s = F(k|\mathbf{x})$ for some integer k, i.e.,

$$\Pr\left(Y \leq k | F(k|\mathbf{X}) = s\right) = s.$$

Yang (2021) proposed to use the subset of the data for which $F(k|\mathbf{X}) \approx s$ to estimate $\Pr(Y \leq k|F(k|\mathbf{X}) \approx s)$ instead

Define the grid point $F(k|\mathbf{x})$ closest to s as $H(s; X, \beta)$

A kernel function $K(\cdot)$ to is used assign weights to the observations depending on the distance of s and $H(s; X_i, \beta)$, $K((H(s; X_i, \beta) - s)/h_n)$, h_n is the bandwidth

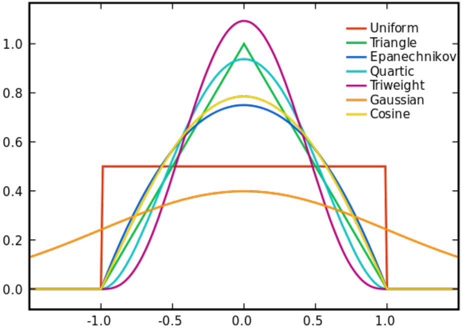


Figure 5: Kernel Functions

Define the quasi-empirical residual distribution function

$$\hat{U}(s) = \sum_{i=1}^{n} W_{ni} 1(F(Y_i|\mathbf{X}_i, \boldsymbol{\beta}) \leq H(s; X_i, \beta)), \qquad (2)$$

where

$$W_{ni} = \frac{K((H(s; X_i, \beta) - s)/h_n)}{\sum_{i=1}^{n} K((H(s; X_i, \beta) - s)/h_n)}$$

Comparison of empirical residual distribution function with $\hat{U}(s)$

Continuous	$\sum_{i=1}^{n} \frac{1}{n} 1(F(Y_i|X_i, \beta) \leq s)$
Discrete	$\sum_{i=1}^{n} W_{ni} 1(F(Y_i|\mathbf{X}_i, \boldsymbol{\beta}) \leq H(s; X_i, \beta))$

Model Diagnostics for LGPIF

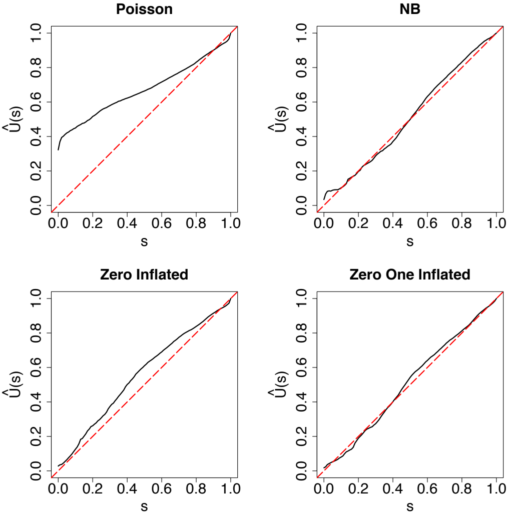


Figure 6: Plot of quasi-empirical residual distribution function $\hat{U}$ (Solid Line) for LGPIF data.

Quasi-empirical residual distribution function

- Pros
  - is principled
  - is close to the hypothesized pattern under the true model
  - under the misspecified model, shows a significant discrepancy
- Cons
  - does not produce residuals themselves and cannot identify causes of misspecification
  - requires tuning bandwidth
  - convergence rate $n^{-1/3}$

References

Cox, David R, and E Joyce Snell. 1968. "A General Definition of Residuals." *Journal of the Royal Statistical Society. Series B (Methodological)*, 248–75.

Pierce, Donald A, and Daniel W Schafer. 1986. "Residuals in Generalized Linear Models." *Journal of the American Statistical Association* 81 (396): 977–86.

Yang, Lu. 2021. "Assessment of Regression Models with Discrete Outcomes Using Quasi-Empirical Residual Distribution Functions." *Journal of Computational and Graphical Statistics* 30 (4): 1019–35.