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

- The diagnostic methods for GLMs mirror those used for Gaussian linear models.
- However, some adaptations are necessary and, depending on the type of GLM, not all diagnostic methods will be applicable.

Leverage and Influence

- Hat matrix
$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}$$
  where $\mathbf{W} = diag(w)$ are weights in IRLS.
- One important difference from the linear model case is that the leverages are no longer just a function of $\mathbf{X}$ and now depend on the response through the weights $\mathbf{W}$
- Cook's distance
$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'(\mathbf{X}'\mathbf{W}\mathbf{X})(\hat{\beta}_{(i)} - \hat{\beta})}{p'\hat{\phi}}$$

Galapagos data: large mean discrete outcomes

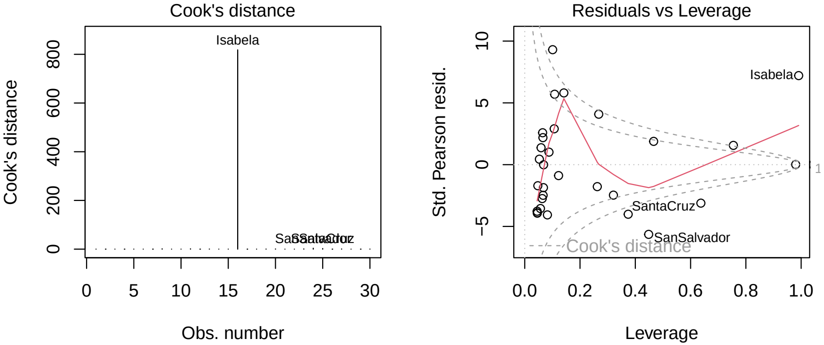
- There are 30 Galapagos islands and 7 variables in the data set. The relationship between the number of plant species and several geographic variables is of interest.

```
data(gala, package="faraway")
head(gala)
```

```
##             Species Endemics  Area Elevation Nearest Scruz Adjacent
## Baltra           58       23 25.09       346     0.6   0.6     1.84
## Bartolome        31       21  1.24       109     0.6  26.3   572.33
## Caldwell          3        3  0.21       114     2.8  58.7     0.78
## Champion         25        9  0.10        46     1.9  47.4     0.18
## Coamano           2        1  0.05        77     1.9   1.9   903.82
## Daphne.Major     18       11  0.34       119     8.0   8.0     1.84
```

```
modp <- glm(Species ~ .,family=poisson,gala)
```

```
par(mfrow=c(1,2))
plot(modp,which=4)
plot(modp,which=5)
```



Residuals

- Pearson residuals

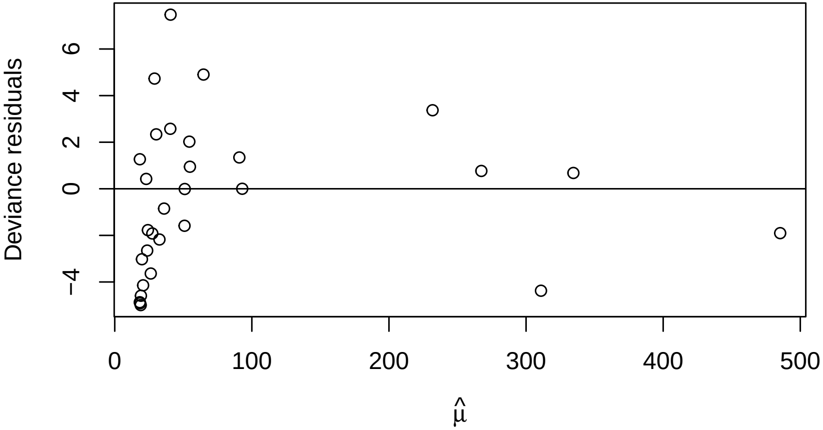
$$\hat{e}_{Pi} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

- $X^2 = \sum_i \hat{e}_{Pi}$
- Let deviance $D = \sum_i d_i$, deviance residuals

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

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

```
plot(residuals(modp,type="deviance") ~ predict(modp,type="response"),
xlab=expression(hat(mu)),ylab="Deviance residuals")
abline(h=0)
```

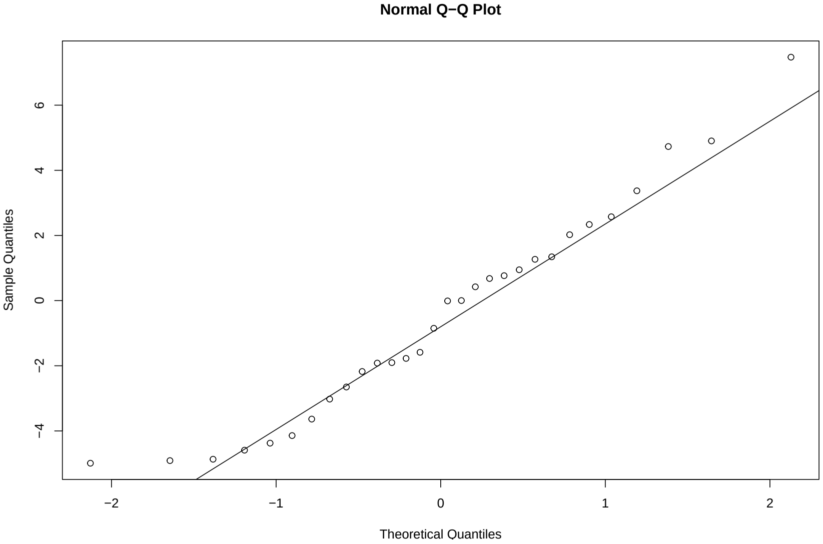


Potential remedies

- 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

QQ plot

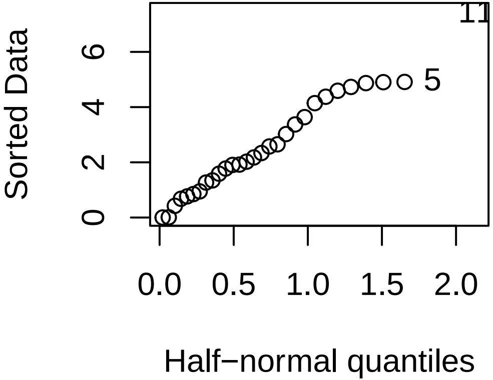
```
qqnorm(residuals(modp,type="deviance"))
qqline(residuals(modp,type="deviance"))
```



Half-normal plots

- One can use a half-normal plot that compares the sorted absolute residuals and the quantiles of the half-normal distribution: $\Phi^{-1}\left(\frac{n+i}{2n+1}\right) \quad i = 1, \ldots, n$
- We seek outliers which may be identified as points off the trend

```
library(faraway)
halfnorm(residuals(modp))
```



Small mean discrete outcomes: LGPIF data

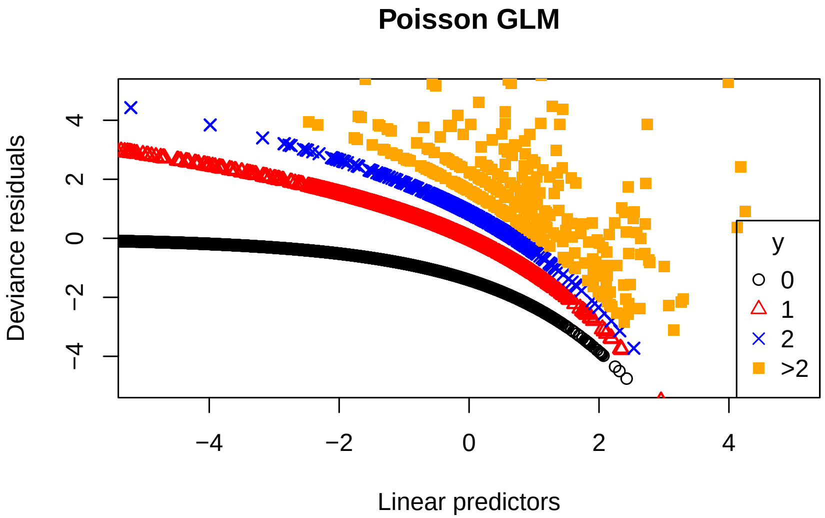
- In some cases, plots of the residuals are not particularly helpful.
- For a binary response, the residual can only take two possible values for given predicted response. This is the most extreme situation, but similar discreteness can occur for binomial responses with small group sizes and Poisson responses that are small.

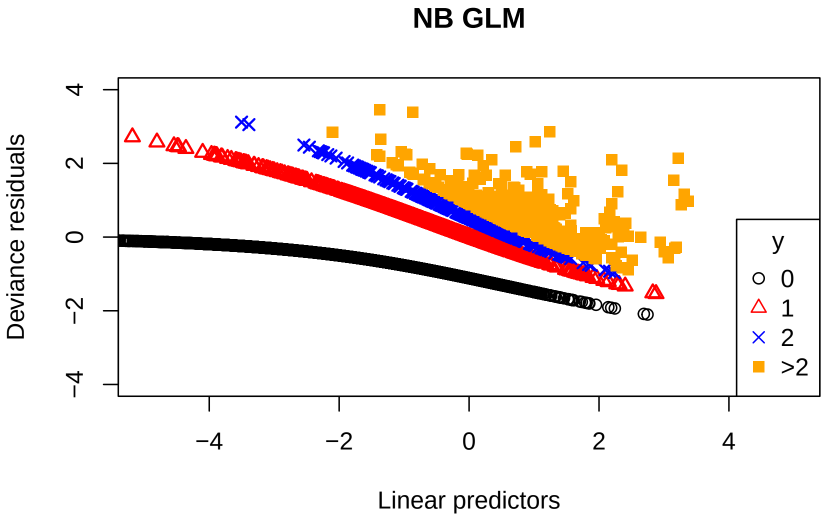
```
freqinBC <- readRDS("freqinBC.rds")
table(freqinBC$FreqBC)
```

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

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

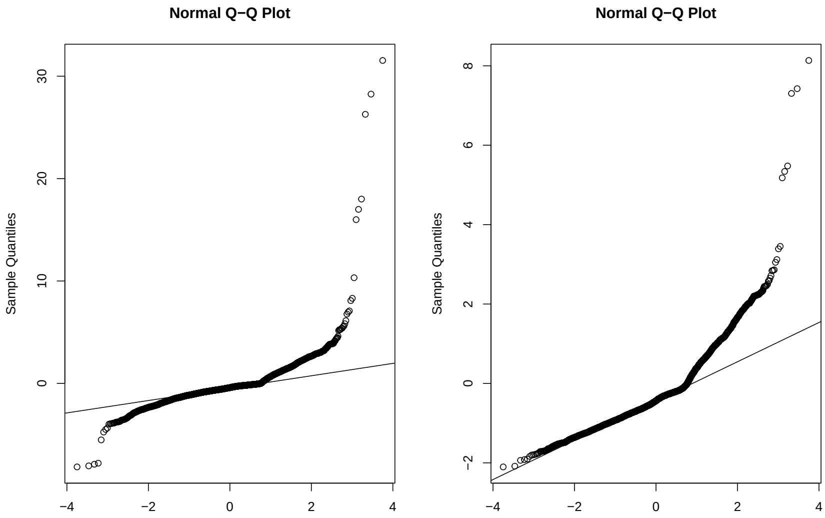
- Plots of residuals in these small mean cases tend to show curved lines of points corresponding to the limited number of observed responses. Such artifacts can obscure the main purpose of the plot.




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

QQ plots

```
par(mfrow=c(1,2))
qqnorm(residuals(freqmodelBC,type="deviance"))
qqline(residuals(freqmodelBC,type="deviance"))
qqnorm(residuals(freqmodelBCnb,type="deviance"))
qqline(residuals(freqmodelBCnb,type="deviance"))
```



Normal Q−Q Plot

Normal Q−Q Plot

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

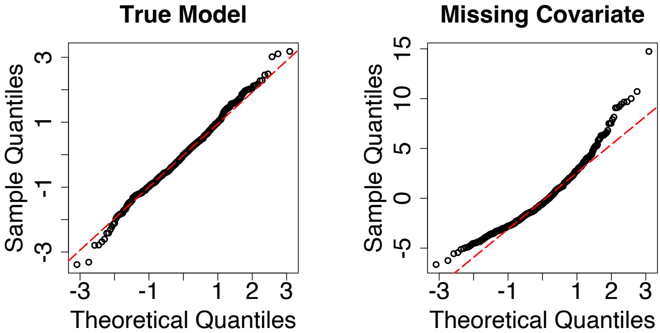


Figure 1: QQ plots for linear regression model residuals.

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

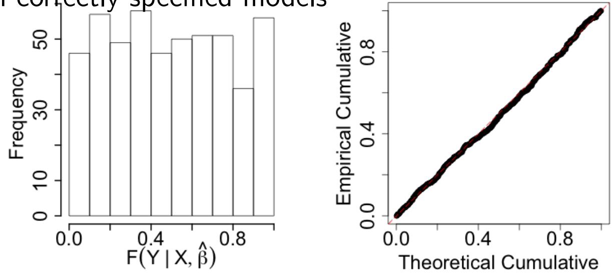


Figure 2: Histogram and PP plot of Cox-Snell residuals for a gamma example.

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

- How Good Is the Approximation?
  - A simulated example: Poisson GLM with log link

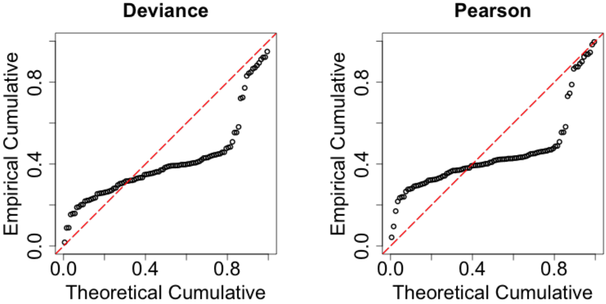


Figure 3: PP plots of residuals for a Poisson GLM under the **true model**.

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

Randomized Quantile Residuals (Dunn and Smyth (1996))

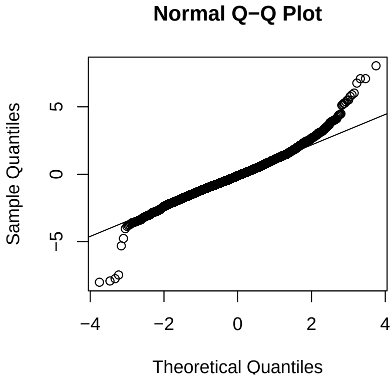
- Idea: transform discrete integer-valued data into continuous data by adding noise
- Let $a_i = \hat{F}_i(Y_i - 1)$ and $b_i = \hat{F}_i(Y_i)$, then the randomized quantile residual

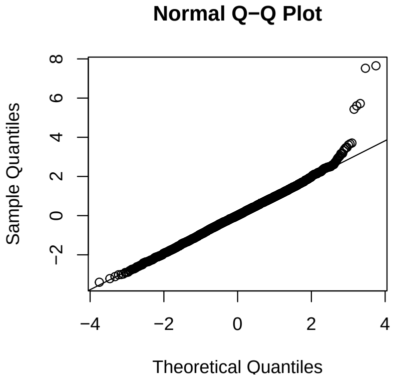
  $$\hat{e}_{Ri} = \Phi^{-1}(V_i),$$

  where V_i is a uniform random variable on the interval $(a_i, b_i]$ independent of Y_i.
- Null pattern: normality

```r
par(mfrow=c(1,2))
library(statmod)
resr <- qresid(freqmodelBC)
qqnorm(resr[-which(is.infinite(resr))])
qqline(resr[-which(is.infinite(resr))])

resrnb <- qresid(freqmodelBCnb)
qqnorm(resrnb[-which(is.infinite(resrnb))])
qqline(resrnb[-which(is.infinite(resrnb))])
```





Drawbacks of Randomized Quantile Residuals

- The procedure injects noise to the data
- The behavior of the residuals depend on the realization of the noise

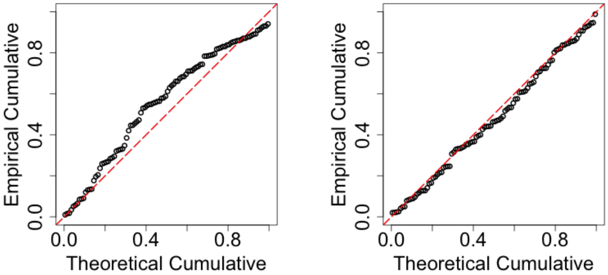


Figure 4: PP plot of randomized quantile residuals of a Poisson GLM example with two different random seeds.

- Not sensitive to misspecification

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

- The Quasi-empirical residual distribution function is an alternative to empirical distribution function of Cox-Snell residuals

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

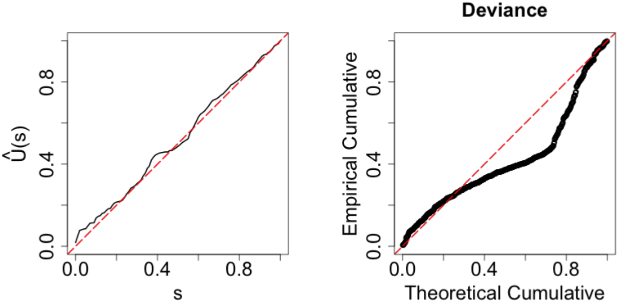


Figure 5: $\hat{U}$ and deviance residuals for a Poisson example under the true model.

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

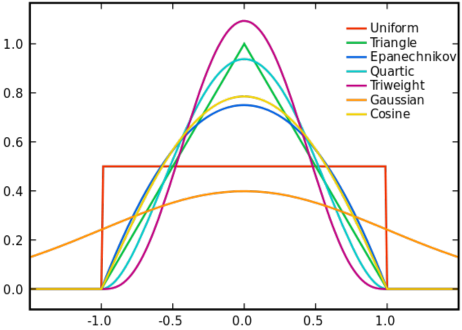


Figure 6: Kernel Functions

- 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

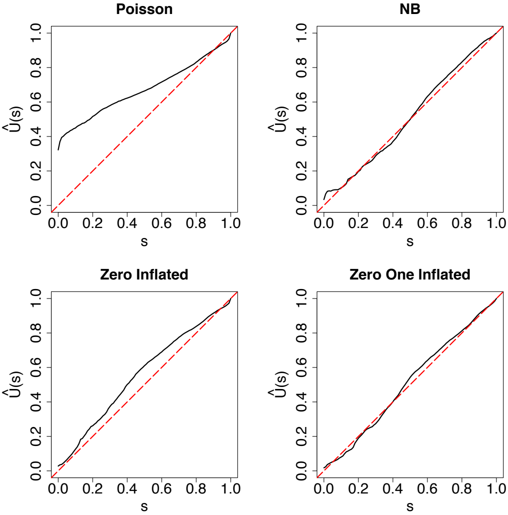


Figure 7: Plot of quasi-empirical residual distribution function $\hat{U}$ (Solid Line) for LGPIF data.

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 $n^{-1/3}$

Double probability integral transform residuals (in progress)

- $F(Y|\mathbf{X})$ itself is not uniformly distributed for discrete outcomes
- Another layer of probability integral transform, $G_0\left(F(Y|\mathbf{X})\right)$, yields a uniform variable under the true model, where G_0 is the distribution of $F(Y|\mathbf{X})$
- The *double probability integral transform residuals*

$$\hat{r}(Y_i|\mathbf{X}_i) = \hat{G}_i\left(F(Y_i|\mathbf{X}_i, \beta)\right)$$

where $\hat{G}_i$ is an estimator of G_0 suited to the *i*th observation

$$\hat{G}_i(s) = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} F\left(F^{(-1)}(s|\mathbf{X}_j, \hat{\beta})|\mathbf{X}_j, \hat{\beta}\right).$$

Causes of misspecification

- Overdispersion: S-shaped pattern

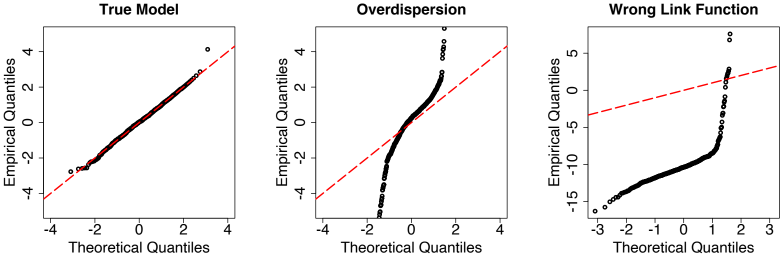


Figure 8: QQ plots of the double probability integral transform residuals under the correctly specified model (left) and models with overdispersion (middle) and an incorrect link function (right).

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.

Dunn, Peter K, and Gordon K Smyth. 1996. "Randomized Quantile Residuals." *Journal of Computational and Graphical Statistics* 5 (3): 236–44.

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.