ggESDA: An **R** Package for exploratory symbolic data analysis using ggplot2

Bo-Syue Jiang

National Taipei University

Abstract

This paper presents the **ggESDA** package,which we developed for exploratory symbolic data analysis in R.Based on **ggplot2** Wickham (2009),the **ggESDA** package which is familiar

Keywords: SDA, EDA, symbolic data analysis, exploratory data analysis, ggplot2 extensions, interval-valued data, R.

1. Introduction: xxx(wait for edit)

1..... (wait for edit)

2..... (wait for edit)

3..... (wait for edit)

4..... (wait for edit)

2. sec title (wait for edit)

(wait for edit)

(wait for edit)

R (wait for edit) `glm()` (Chambers and Hastie 1992) in the **stats** package.

```
glm(formula, data, subset, na.action, weights, offset,
  family = gaussian, start = NULL, control = glm.control(...),
  model = TRUE, y = TRUE, x = FALSE, ...)
```

Type	Distribution	Method	Description
GLM	Poisson	ML	Poisson regression: classical GLM, estimated by maximum likelihood (ML)
		Quasi	"Quasi-Poisson regression": same mean function, estimated by quasi-ML (QML) or equivalently generalized estimating equations (GEE), inference adjustment via estimated dispersion parameter
		Adjusted	"Adjusted Poisson regression": same mean function, estimated by QML/GEE, inference adjustment via sandwich covariances
	NB	ML	NB regression: extended GLM, estimated by ML including additional shape parameter
Zero-augmented	Poisson	ML	Zero-inflated Poisson (ZIP), hurdle Poisson
	NB	ML	Zero-inflated NB (ZINB), hurdle NB

Table 1: Overview of various count regression models. The table is usually placed at the top of the page (`[t!]`), centered (`centering`), has a caption below the table, column headers and captions are in sentence style, and if possible vertical lines should be avoided.

(wait for edit)

(wait for edit)

3. third title (wait for edit)

(wait for edit)

```
R> data("quine", package = "MASS")
```

and a basic frequency distribution of the response variable is displayed in Figure 1.

(wait for edit)

(wait for edit)

```
R> m_pois <- glm(Days ~ (Eth + Sex + Age + Lrn)^2, data = quine,
+    family = poisson)
```

To account for potential overdispersion we also consider a negative binomial GLM.

```
R> library("MASS")
R> m_nbin <- glm.nb(Days ~ (Eth + Sex + Age + Lrn)^2, data = quine)
```

In a comparison with the BIC the latter model is clearly preferred.

```
R> BIC(m_pois, m_nbin)
```
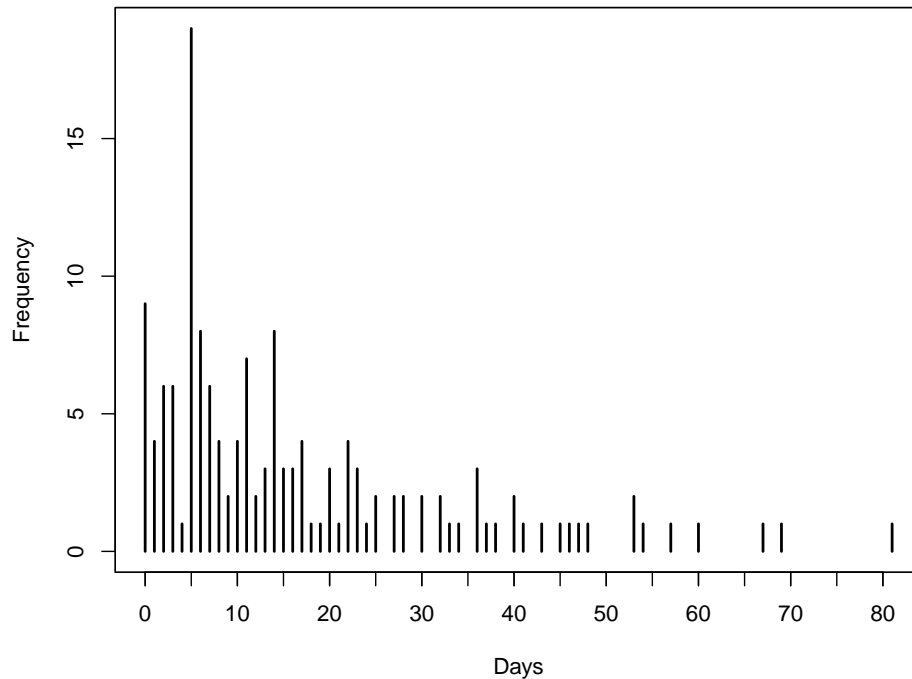
Figure 1: Frequency distribution for number of days absent from school.

```
          df       BIC
m_pois 18 2046.851
m_nbin 19 1157.235
```

Hence, the full summary of that model is shown below.

```
R> summary(m_nbin)

Call:
glm.nb(formula = Days ~ (Eth + Sex + Age + Lrn)^2, data = quine,
    init.theta = 1.60364105, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0857  -0.8306  -0.2620   0.4282   2.0898

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.00155    0.33709   8.904  < 2e-16 ***
EthN        -0.24591    0.39135  -0.628  0.52977
SexM        -0.77181    0.38021  -2.030  0.04236 *
AgeF1       -0.02546    0.41615  -0.061  0.95121
AgeF2       -0.54884    0.54393  -1.009  0.31296
AgeF3       -0.25735    0.40558  -0.635  0.52574
```

```
LrnSL        0.38919     0.48421    0.804   0.42153
EthN:SexM    0.36240     0.29430    1.231   0.21818
EthN:AgeF1  -0.70000     0.43646   -1.604   0.10876
EthN:AgeF2  -1.23283     0.42962   -2.870   0.00411 **
EthN:AgeF3   0.04721     0.44883    0.105   0.91622
EthN:LrnSL   0.06847     0.34040    0.201   0.84059
SexM:AgeF1   0.02257     0.47360    0.048   0.96198
SexM:AgeF2   1.55330     0.51325    3.026   0.00247 **
SexM:AgeF3   1.25227     0.45539    2.750   0.00596 **
SexM:LrnSL   0.07187     0.40805    0.176   0.86019
AgeF1:LrnSL -0.43101     0.47948   -0.899   0.36870
AgeF2:LrnSL  0.52074     0.48567    1.072   0.28363
AgeF3:LrnSL      NA          NA       NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.6036) family taken to be 1)

    Null deviance: 235.23  on 145  degrees of freedom
Residual deviance: 167.53  on 128  degrees of freedom
AIC: 1100.5

Number of Fisher Scoring iterations: 1


          Theta:  1.604
      Std. Err.:  0.214

 2 x log-likelihood:  -1062.546
```

4. Summary and discussion

▌ (wait for edit)

Computational details

▌ (wait for edit)

(wait for edit)

Acknowledgments

(wait for edit)

References

Chambers JM, Hastie TJ (eds.) (1992). *Statistical Models in S*. Chapman & Hall, London.

Wickham H (2009). "**ggplot2**: Elegant Graphics for Data Analysis." *Media*, **35**(211), 10–1007. doi:10.1007/978-0-387-98141-3.

(wait for edit)

Affiliation:

myaddress