Proceedings of the International Conference on Statistics: Theory and Applications (ICSTA'19)
Lisbon, Portugal – August 13 - 14, 2019
Paper No. ICSTA 21
DOI: 10.11159/icsta19.21

Joint Mean and Dispersion Effects Modelling in R: The jmdem Package

Ka Yui Karl Wu[1]
[1]Singapore University of Social Science
463 Clementi Road, Singapore 599494
karlwuky@suss.edu.sg

Abstract – The mean of a response variable is commonly modelled by linear regression or generalised linear model where the error distribution is either Gaussian, binomial, Poisson or one from the exponential family. Furthermore, as the error variance is assumed to be identical for all observed data, the same applies to the dispersion parameter, which is usually estimated by the scaled Pearson chi-squared statistic. The statistical inference on the model parameters becomes unreliable once these assumptions are violated. One way to address the issue of varying variance is to formulate dispersion models in which the expected dispersion is estimated by a generalised linear model. With the R package *jmdem*, we can fit the response mean and dispersion in a joint model. The advantage of this approach is that the two models are interlinked, and the estimates of both models are the arguments that maximises the joint likelihood function which reduces the computation effort significantly and enhances the quality of the estimator at the same time.

Keywords**:** Generalised linear models, R package, Dispersion models.

1. Introduction

Each probability distribution is characterised by its parameters, most commonly the location and dispersion parameters as in the case of the Gaussian distribution. The estimation of the location parameter for the ith observation, usually represented by the response mean $\mu_i, i = 1, \dots n$, using linear regression $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$, where $X_1, \dots, X_p$ are p different non-random independent variables and $\beta_0, \beta_1, \dots, \beta_p$ are the unknown regression coefficients, is one of the most essential and common statistical techniques. The model errors $\epsilon_1, \dots, \epsilon_n$ are assumed to be random and follow independently the same probability distribution as the response variable with zero expectation and constant variance σ^2. This also implies that the variance of the response variable ought to be constant for all observations in the entire sample. By the theory of generalised linear models (GLM) [3] [4], $\mathrm{var}(y_i)$ is a composition of $\phi V(\mu_i)$, where ϕ is the unknown dispersion parameter and $V(\mu_i)$ is the variance function well-defined by the distribution of the response variable y_i. To estimate ϕ, the Pearson chi-square statistic has been suggested

$$\hat{\phi} = \frac{1}{n-p-1}\sum \frac{(y_i - \mu_i)^2}{V(\mu_i)} \tag{1}$$

which is an aggregated statistic for the entire sample. The estimate $\hat{\phi}$ also indicates that the dispersion parameter does not vary from observation to observation. Nevertheless, if this assumption is violated, the estimation of the standard errors of β_j, $j = 0, \dots, p$ can be biased. An approach to overcome this problem is a double generalised linear model proposed by Smyth [5].

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \tag{2}$$

$$h(\phi_i) = \lambda_0 + \lambda_1 z_{i1} + \cdots + \lambda_q z_{iq} \tag{3}$$

The equations (2) and (3) together form the joint mean and dispersion effects model (*jmdem*). These are two generalised linear models in which the location and dispersion parameters are estimated by two sets of independent variables $X_1, \dots, X_p$ as well as $Z_1, \dots, Z_q$. The link functions $g(\cdot)$ and $h(\cdot)$ can be any arbitrary functions which map the response mean μ_i and

response dispersion ϕ_i to the real line. The parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$ and $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \dots, \lambda_q)^\top$ are estimated by maximising the log-likelihood function of the exponential family given the design data matrix $\boldsymbol{X}$, $\boldsymbol{Z}$ and response variable vector $\boldsymbol{Y}$:

$$\ell(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y}) = \sum_{i=1}^{n} \left\{ \frac{w_i}{h^{-1}(\boldsymbol{z}_i\boldsymbol{\lambda})} \left[y_i \theta(\mu_i) - b\big(\theta(\mu_i)\big) \right] + \log\left[c\big(y_i, h^{-1}(\boldsymbol{z}_i\boldsymbol{\lambda})\big) \right] \right\} \tag{4}$$

where $\theta_i = \theta(\mu_i)$ is the canonical parameter and $b(\cdot)$ is the cumulant function with the properties $\partial b/\partial\theta = b'(\cdot) = \mu = g^{-1}(x_i\beta)$ and $\partial^2 b/\partial\theta^2 = V(\mu_i)$. Note that $h^{-1}(\boldsymbol{z}_i\boldsymbol{\lambda})$ is another way to express the estimated individual dispersion parameter derived from (3). The function $c\big(y_i, h^{-1}(\boldsymbol{z}_i\boldsymbol{\lambda})\big)$ is an arbitrary function that depends only on the response variable and the dispersion parameter, and not the response mean μ_i.

2. The jmdem Package in R

The `jmdem` package [8] in R fits joint models for the mean and dispersion effects as defined (2) and (3) by maximising the log-likelihood function in (4). Suppose we have a data frame called "example.dat" that contains the variables y, $x1$, $x2$, $z1$ and $z2$. To fit a joint mean and dispersion effects model in which the response variable is normally distributed, we can use the following syntax in R:

```
model1 <- jmdem(mformula = y ~ x1 + x2, dformula = ~ z1 + z2,
                data = example.dat, mfamily = poisson(link = "log"))
```

The `jmdem` syntax as shown in the above example is constructed in a very similar way as `lm` or `glm` that are used to fit linear models or generalised linear models. The main difference here is that `jmdem` contains two model formulas: the argument `mformula` is responsible for the fitting of the mean effects model and `dformula` for the fitting of the dispersion effects model. The equations of both the mean and dispersion effect models are written after the "~" operator with the corresponding independent variables connected by mathematical operators such as "+" or "−", as well as ":", if the interaction effect of two or more variables should be included. The variable written before "~" in `mformula` is the response variable. Note that there is no response variable for `dformula` since it is determined by either one of the following formulas

$$d_i = D_i(y_i, \mu_i) = 2w_i \int_{\mu_i}^{y_i} \frac{y_i - t_i}{V(t_i)} dt_i \tag{5a}$$

$$d_i = r_{p_i}(y_i, \mu_i) = \frac{w_i (y_i - \mu_i)^2}{V(\mu_i)} \tag{5b}$$

where $D_i(y_i, \mu_i)$ is the so called deviance component computed by the quasi-likelihood function which measures the difference of the individual likelihood evaluated at the observed and expected values of the response variable, respectively. The individual Pearson residual $r_{p_i}(y_i, \mu_i)$ as given in (5b) takes the standardised square distances of the response mean from its observed value into account. While the deviance component in (5a) is the commonly proposed approach [2] [7], the use of the individual Pearson residuals is briefly introduced in [3] and studied in detail in [6]. To specify the type of d_i we can add the option "`dev.type = c("deviance", "pearson")`" into the `jmdem` function.

Same as `glm`, `jmdem` allows us to specify the distribution family and link function for both the mean and dispersion effect models according to the characteristic of the data. In the above example, we assume the response variable to be Poisson distributed with logarithm link function. Note that the Gaussian distribution with identity link function is the default setting for `mfamily` and gamma distribution with log-link for `dfamily`. It is noteworthy that the `jmdem` package only integrates the same family objects for models provided by base R, which includes those distributions that are also used in `glm` such as

`gaussian`, `binomial`, `Poisson`, etc. Other commonly used distributions such as the negative binomial or Tweedie distributions cannot be specified in `jmdem` yet. But these options will be available in the future version of the package.

The estimation mechanism in `jmdem` is to use `optim`, a general-purpose optimisation function integrated in R, to find the optimum of the likelihood function by trying various values for the target variables in an iterative process. In other words, initial values of the regression coefficients must be specified prior to the estimation, and the estimates of the parameters will be updated after each iteration of optimisation. By adding "`betastart = c(…)`" and "`lambdastart = c(…)`" as arguments to `jmdem` we can specify the initial values of $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$ and $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \dots, \lambda_q)^\top$. If they are omitted, `jmdem` will use the estimates of $\widehat{\boldsymbol{\beta}} = (\bar{y}, 0, \dots, 0)^\top$ and $\hat{\boldsymbol{\lambda}} = (\hat{\phi}_0, 0, \dots, 0)^\top$ as their initial values, where is $\bar{y}$ the sample mean of $y_1, \dots, y_n$ and $\hat{\phi}_0 = \widehat{\text{var}(y)}\,/V(\bar{y})$. Moreover, we can also add `method = c("Nelder-Mead", "BFGS", "CG", "L-BFGS-B", "SANN", "Brent")` to the `jmdem` syntax to specify the optimisation method which are integrated in the `optim` function. For detail explanation of these methods we refer to the R documentation [9].

3. Real Data Application

The following study was conducted by Long [1] in which the number of articles published by a sample of scholars should be analysed and explained. The explanatory variables here are given in the following table:

Table 1: Table of Independent Variables.

Variable names	Descriptions
mar	Martial status (1 = married, 0 = not married)
fem	Gender (1 = female, 0 = male)
phd	Prestige of the PhD department
ment	Number of citations received by the person's mentor
kid5	Number of children (0 = none, 1 = 1 child, 2 = 2 children, 3 = 3 children or more)

The number of published articles is obviously a Poisson count variable with sample mean 1.69 and sample variance 3.71. From the theory of GLM, we know that $V(\mu) = \mu$ for Poisson distributed random variable. Hence, the estimated dispersion here is $\hat{\phi} = \text{var}(Y)/V(\mu) = 3.71/1.69 = 2.19$. Though the data are overdispersed here and a standard Poisson log-linear model would not be suitable, we will fit one for our reference.

For the application of jmdem, we will fit a Poisson mean effects model and a gamma log-linear dispersion effect model jointly. The response variable of the dispersion effect model is computed by the quasi-likelihood function as given in (5a). For the optimisation, we choose the "`Nelder-Mead`" method which is the default setting of the `optim` function. Furthermore, we will add the argument `disp.adj = TRUE` into the `jmdem` syntax to correct the higher order cumulants of the dispersion parameter as suggested in [3] and [6]. The estimation results of the final model are:

Table 2: Main Effects Estimation Results.

Mean Effects	Poisson-Gamma jmdem			Poisson GLM		
(Intercept)	0.3007	***	(0.0827)	0.3477	***	(0.0601)
fem (fem = 1)	-0.202	***	(0.0736)	-0.2260	***	(0.0547)
mar (mar = 1)	0.1397	*	(0.0845)	0.1481	**	(0.0628)
kid5 (kid5 = 1)	-0.1617	*	(0.0955)	-0.1803	**	(0.0706)
kid5.2 (kid5 = 2)	-0.2956	**	(0.1237)	-0.3278	***	(0.0909)
kid5.3 (kid5 = 3)	-0.6735	*	(0.3583)	-0.8215	***	(0.2817)
ment	0.0287	***	(0.0032)	0.0256	***	(0.0020)

(Values in parentheses are standard errors, effects marked with * have p-value < 0.1, ** have p-value < 0.05 and *** have p-value < 0.01, respectively)

Table 3: Dispersion Effects Estimation Results.

Dispersion Effects	**Poisson-Gamma jmdem**	
(Intercept)	0.2964 ***	(0.065)
ment	0.0145 ***	(0.005)

(Values in parentheses are standard errors, effects marked with * have p-value < 0.1, ** have p-value < 0.05 and *** have p-value < 0.01, respectively)

The prestige of the PhD department has been identified as insignificant on the number of published articles by both approaches at the 5% significance level. It is therefore removed from the final model. For comparison, we keep all other independent variables as their effects have been identified as significant by either one of the models. As given in Table 2, the Poisson GLM identifies all effects as highly statistically significant while gender and the number of citations the mentor received are the only effects confirmed as such according to jmdem. Furthermore, significant difference in the number of published articles between scholars with and without children has been identified by the Poisson GLM, whereas jmdem could only confirm the same finding between scholars with two children and none. In general, standard GLM tends to underestimate the standard errors of the coefficients since it does not take over- or underdispersion into account. The joint mean and dispersion effects model makes the corresponding amendment by modelling the individual dispersion.

The final dispersion effects model, which is a gamma GLM with log-link, does not only indicate that there is an aggregated overdispersion in the sample as the intercept is positive and statistically significant, it also shows that the probability distribution of published articles are not the same for each individual since the dispersion parameter here depends on the individual citation record of the mentor. As a result, this set of data is not only overdispersed, the dispersion also varies from observation to observation.

4. Conclusion

The `jmdem` package facilitates the modelling of the mean and dispersion effects jointly for data that follow the same probability distribution, but with individually varying location and dispersion parameters. The assumption of constant dispersion in an independent sample can therefore be omitted, and the standard errors of the mean effect coefficients will not be over- or underestimated when varying dispersion occurs. This leads to more accurate identification of significant mean and dispersion effects. The `jmdem` package includes all the features of the `lm` and `glm` function in R, respectively, and extended it to the specific use for the joint mean and dispersion effect models, including the distribution family of the dispersion model, the optimisation method, the computation of the individual deviance as the response variable of the dispersion model as well as the adjustment of higher order cumulants of the dispersion estimator. The number of optional distribution families will be enlarged in the next major update of the package.

References

[1] J. S. Long, "The Origins of Sex Differences in Science," *Social Forces*, vol. 68, no. 4, pp. 1297-1316, 1990.
[2] P. McCullagh, "Quasi-Likelihood Functions," *The Annals of Statistics*, vol. 11, no. 1, pp. 59-67, 1983.
[3] P. McCullagh, J. A. Nelder, *Generalized Linear Models*. Second Edition. Chapman & Hall, 1989.
[4] J. A. Nelder, R. W. Wedderburn, "Generalized Linear Models," *Journal of the Royal Statistical Society: Series A*, vol. 135, no. 3, pp. 370-384, 1972.
[5] G.K. Smyth, "Generalized Linear Models with Varying Dispersion," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 51, no. 1, 47–60, 1989.

[6] K. Y. K. Wu, W. K. Li, "On a Dispersion Model with Pearson Residual Responses", *Computation Statistics and Data Analysis*, vol. 103, pp. 17-27, 2016.
[7] Wedderburn, R., "Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method," *Biometrika*, vol. 61, no. 3, pp. 439–447, 1974.
[8] K. Y. K. Wu (2018, April 27). Package 'jmdem' [Online]. Available: https://cran.r-project.org/web/packages/jmdem/jmdem.pdf.
[9] R Documentation, package *stats* version 3.5.0 (2019, April 15). General-purpose Optimization. Available: https://stat.ethz.ch/R-manual/R-patched/library/stats/html/optim.html.