MRCV: A Package for Analyzing Categorical Variables with Multiple Response Options

by Natalie A. Koziol and Christopher R. Bilder

Abstract Multiple response categorical variables (MRCVs), also known as "pick any" or "choose all that apply" variables, summarize survey questions for which respondents are allowed to select more than one category response option. Traditional methods for analyzing the association between categorical variables are not appropriate with MRCVs due to the within-subject dependence among responses. We have developed the **MRCV** package as the first R package available to correctly analyze MRCV data. Statistical methods offered by our package include counterparts to traditional Pearson chi-square tests for independence and loglinear models, where bootstrap methods and Rao-Scott adjustments are relied on to obtain valid inferences. We demonstrate the primary functions within the package by analyzing data from a survey assessing the swine waste management practices of Kansas farmers.

Introduction

Survey questions often instruct respondents to "choose all that apply" from a list of response categories. For example, surveys instituted by U.S. government agencies are mandated to ask race and ethnicity questions in this format (Office of Management and Budget, 1997, p. 58781). In medical applications, "choose all that apply" questions have been used for a variety of purposes, including gathering information about treatment and monitoring strategies (Kantarjian et al., 2007; Riegel et al., 2006). Outside of surveys, this format can appear in unexpected applications. For example, wildlife management researchers are often interested in the food habits of animal species. Traces of prey in scats provide these researchers with a "choose all that apply" type of response because multiple prey types may be present (Lemons et al., 2010; Riemer et al., 2011).

Variables that summarize data arising from a "choose all that apply" format are referred to as multiple response categorical variables (MRCVs), and the response categories within each MRCV are referred to as items (Bilder and Loughin, 2004). Because individual subjects are allowed to choose multiple items, the responses are likely dependent, and therefore traditional methods for analyzing categorical variables (e.g., Pearson chi-square tests for independence, loglinear models) are not appropriate. Unfortunately, numerous examples exist where these traditional methods are still used (see Wright, 2010 for a review), which can lead to erroneous results (Loughin and Scherer, 1998).

While MRCVs have been identified since at least Coombs (1964), methods for correctly analyzing MRCVs in the context of common categorical data analysis interests, such as examining associations between variables, have only been available for approximately 15 years (e.g., see Agresti and Liu, 1999). Our **MRCV** package (Koziol and Bilder, 2014) is the first R package available to implement valid inference procedures for this type of data. The functions within the package can be used by researchers who want to examine the relationship among items from up to three MRCVs.

We begin this paper by first illustrating functions within the package for summarizing MRCV data and testing for independence. Then, we illustrate functions for fitting a generalized loglinear model to MRCV data and for performing follow-up analyses using method functions. Our examples focus on only two MRCVs for brevity reasons, but we discuss extensions in the conclusion.

Test for independence

We begin with an example from Bilder and Loughin (2007) involving a simple random sample of Kansas swine farmers. There are two MRCVs to be examined here, and we denote them generically as W and Y. The first MRCV (W) corresponds to a survey question that asked farmers to state which contaminants they tested for from the items "nitrogen", "phosphorous", and "salt" (W_1, W_2, W_3, respectively). The second MRCV (Y) corresponds to a survey question that asked farmers to identify their swine waste storage methods from the items "lagoon", "pit", "natural drainage", and "holding tank" (Y_1, Y_2, Y_3, Y_4, respectively). Farmers were instructed to "choose all that apply" from each of these predefined lists. By using a 0 to denote an item not chosen (negative response) and a 1 to denote an item chosen (positive response), each observation consists of a set of correlated binary responses, as shown below:

```
> head(farmer2, n = 3)
  w1 w2 w3 y1 y2 y3 y4
1  0  0  0  0  0  0  0
2  0  0  0  0  0  0  1
3  0  0  0  0  0  0  1

> tail(farmer2, n = 3)
    w1 w2 w3 y1 y2 y3 y4
277  1  1  1  1  1  0  0
278  1  1  1  1  1  0  0
279  1  1  1  1  1  1  0
```

We see, for example, that the third farmer does not test for any contaminants and uses only a holding tank for waste storage.

Contingency table-like summaries of MRCV data are often given in papers. In particular, marginal counts for all pairwise positive responses between items in *W* and *Y* are shown in Table 1. This display format can lead researchers to want to apply Pearson chi-square tests (or other simple categorical measures) to the table of counts in order to understand associations between the MRCVs. However, this approach is not correct because it does not take into account the fact that an individual subject can contribute to multiple counts in the table, which violates any type of multinomial distribution underlying assumption for these specific counts. Furthermore, three other tables summarizing the pairwise positive/negative responses (e.g., summarizing responses for items "not" chosen) of this type could also be constructed. Agresti and Liu (1999) and Bilder and Loughin (2001) show that testing procedures are not invariant to whether positive or negative responses are summarized and that different conclusions about the data can be reached depending on the types of responses summarized.

Examining all possible combinations of the positive/negative item responses between MRCVs is the preferred way to display and subsequently analyze MRCV data. The item.response.table() function provides this summary for each (W_i, Y_j) pair:

```
> item.response.table(data = farmer2, I = 3, J = 4)

        y1          y2          y3          y4
        0    1      0    1      0    1      0    1
   w1 0 123  116    175  64     156  83     228  11
      1  13   27     24  16      38   2      38   2
   w2 0 128  121    181  68     165  84     237  12
      1   8   22     18  12      29   1      29   1
   w3 0 134  124    184  74     174  84     245  13
      1   2   19     15   6      20   1      21   0
```

where I is the number of items for *W* and J is the number of items for *Y*. The pairwise item-response table indicates, for example, that 27 farmers tested for nitrogen and used lagoon as a waste storage method (i.e., $W_1 = 1$, $Y_1 = 1$). Furthermore, 123 farmers did not test for nitrogen and did not use a lagoon, 13 farmers tested for nitrogen without using a lagoon, and 116 farmers did not test for nitrogen while using a lagoon. In total, $27 + 123 + 13 + 116 = 279$ farmers participated in the survey (there are no missing responses to any item).

Agresti and Liu (1999) provided the MRCV extension to testing for independence between single response categorical variables (SRCVs). This test, known as a test for simultaneous pairwise marginal independence (SPMI), involves determining whether each $W_1, \ldots, W_I$ is pairwise independent of each $Y_1, \ldots, Y_J$. Our MI.test() function calculates their modified Pearson statistic as $X_S^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} X_{S,i,j}^2$ where $X_{S,i,j}^2$ is the usual Pearson chi-square statistic used in this situation to test for independence in the 2×2 tables formed by each (W_i, Y_j) response combination. In our example, X_S^2 is the sum of 12

		Waste storage method			
		Lagoon	Pit	Natural Drainage	Holding tank
	Nitrogen	27	16	2	2
Contaminant	Phosphorous	22	12	1	1
	Salt	19	6	1	0

Table 1: Pairwise positive responses for the data in farmer2. While not recommended, this summary format is available through the marginal.table() function of **MRCV**.

pairwise marginal tests for independence. In general, X_S^2 does not have an asymptotic χ_{IJ}^2 distribution due to dependency among the $X_{S,i,j}^2$. Rather, the asymptotic distribution is a linear combination of independent χ_1^2 random variables (Bilder and Loughin, 2004).

The MI.test() function offers three methods, available through its type argument, that can be used with X_S^2 or the $X_{S,i,j}^2$ individual statistics to perform valid tests for SPMI. The type = "boot" argument value specifies the use of the nonparametric bootstrap to estimate the sampling distribution of X_S^2 under SPMI and to calculate an appropriate p-value using B resamples. In addition, two p-value combination methods—the product and minimum of p-values—are implemented to combine the p-values obtained from $X_{S,i,j}^2$ and a χ_1^2 approximation. Details on these bootstrap approaches are given in Bilder and Loughin (2004). The type = "rs2" argument value applies a Rao-Scott second-order adjustment to X_S^2 and its sampling distribution. This procedure adjusts X_S^2 to match the first two moments of a chi-square random variable, asymptotically. Details on this approach are provided in Bilder and Loughin (2004) and Thomas and Decady (2004). Finally, the type = "bon" argument value simply applies a Bonferroni adjustment with each $X_{S,i,j}^2$ and a χ_1^2 approximation. To implement all three methods, we can use the type = "all" argument value:

```
> set.seed(102211) # Set seed to replicate bootstrap results
> MI.test(data = farmer2, I = 3, J = 4, type = "all", B = 1999, plot.hist = TRUE)

Test for Simultaneous Pairwise Marginal Independence (SPMI)

Unadjusted Pearson Chi-Square Tests for Independence:
X^2_S = 64.03
X^2_S.ij =
      y1   y2    y3    y4
w1  4.93 2.93 14.29 0.01
w2  6.56 2.11 11.68 0.13
w3 13.98 0.00  7.08 0.32

Bootstrap Results:
Final results based on 1999 resamples
p.boot = 0.0005
p.combo.prod = 0.0005
p.combo.min = 0.001

Second-Order Rao-Scott Adjusted Results:
X^2_S.adj = 36.17
df.adj = 6.78
p.adj < 0.0001

Bonferroni Adjusted Results:
p.adj = 0.0019
p.ij.adj =
     y1     y2     y3     y4
w1 0.3163 1.0000 0.0019 1.0000
w2 0.1253 1.0000 0.0076 1.0000
w3 0.0022 1.0000 0.0934 1.0000
```

Figure 1 shows histograms from the bootstrap implementations. All of the methods provide strong evidence for rejecting SPMI. The $X_{S,i,j}^2$ and corresponding Bonferroni adjusted p-values indicate a significant association for the (W_1, Y_3), (W_2, Y_3), and (W_3, Y_1) combinations.

Loglinear modeling

SPMI is only one possible association structure between MRCVs. Bilder and Loughin (2007) introduced a flexible loglinear modeling approach that allows researchers to consider alternative association structures somewhere between SPMI and complete dependence. Within this framework, a model under SPMI is given as

$$\log\left(\mu_{ab(ij)}\right) = \gamma_{ij} + \eta_{a(ij)}^W + \eta_{b(ij)}^Y \tag{1}$$

where $\mu_{ab(ij)}$ is the expected number of subjects who responded ($W_i = a$, $Y_j = b$) for $a, b \in \{0, 1\}$.
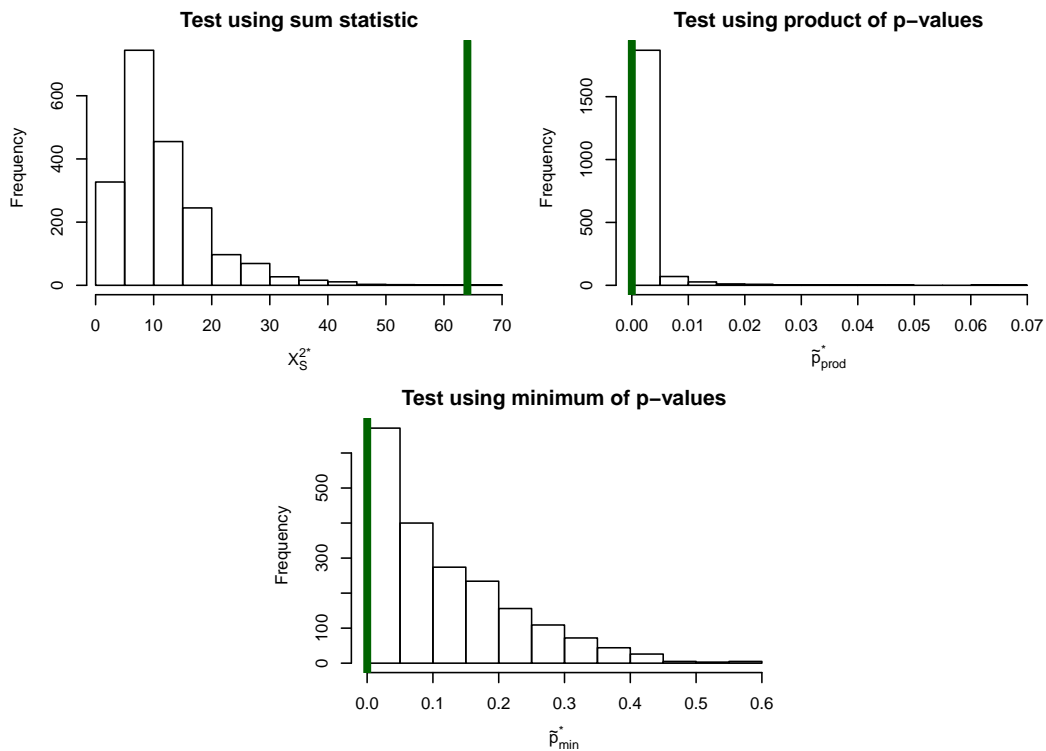
Figure 1: Histograms of the bootstrap estimated sampling distributions, where X_S^{2*} is the modified Pearson statistic calculated for a resample, and $\tilde{p}_{prod}^*$ and $\tilde{p}_{min}^*$ are the product and minimum p-value combination statistics, respectively, calculated for a resample. The vertical lines correspond to the statistics for the observed data.

The terms on the right side of the model are the same as for a loglinear model under independence between two SRCVs, where we have added a subscript (ij) to indicate a particular 2×2 table for (W_i, Y_j) within the pairwise item-response table. The usual constraints are placed on the model parameters to maintain identifiability.

Adding additional terms to Equation (1) leads to different types of association structures between the MRCVs. For example, adding λ_{ab} to Equation (1) produces a homogeneous association model (i.e., a model that implies equal odds ratios, not necessarily equal to 1, for each 2×2 table), adding $\lambda_{ab(i)}^W$ or $\lambda_{ab(j)}^Y$ to the homogeneous association model produces a W- or Y-main effects model, respectively, and adding both of these terms to the homogeneous association model produces a W- and Y-main effects model. The addition of a WY interaction term, $\lambda_{ab(ij)}^{WY}$, produces the saturated model.

The `genloglin()` function estimates the above models through a marginal estimation approach. Within `genloglin()`, a new data frame is created by converting the raw data into the pairwise item-response counts:

```
> item.response.table(data = farmer2, I = 3, J = 4, create.dataframe = TRUE)
   W  Y wi yj count
1 w1 y1  0  0   123
2 w1 y1  0  1   116
3 w1 y1  1  0    13
4 w1 y1  1  1    27
5 w1 y2  0  0   175

< output omitted >

48 w3 y4  1  1     0
```

The `glm()` function is subsequently called from within `genloglin()` to estimate a loglinear model to these counts. Rao-Scott adjustments are then applied to obtain valid large-sample standard error estimates. The `model` argument of `genloglin()` can take the names of `"spmi"`, `"homogeneous"`, `"w.main"`, `"y.main"`, `"wy.main"`, and `"saturated"` to specify a particular model. Alternatively, a user-supplied formula allows for more flexibility by specifying the model in terms of `W`, `Y`, `wi`, `yj`, `count`, `W1`, ...,

WI, and Y1,...,YJ, which we illustrate shortly. The boot = TRUE (the default) value for genloglin() specifies that resamples should be taken under the fitted model. We use the method of Gange (1995) for generating correlated binary data to perform semi-parametric bootstrap resampling in this case. These resamples are subsequently used for hypothesis tests, confidence intervals, and/or standardized residuals with our related method functions for objects returned by genloglin().

We demonstrate the genloglin() function by estimating the Y-main effects model to the farmer2 data, and then summarize the results using our summary() method function:

```
> set.seed(499077) # Set seed to replicate bootstrap results
> mod.fit <- genloglin(data = farmer2, I = 3, J = 4, model = "y.main", B = 1999,
+                      print.status = FALSE)
> summary(mod.fit)

Call:
glm(formula = count ~ -1 + W:Y + wi %in% W:Y + yj %in% W:Y + wi:yj + wi:yj %in% Y,
  family = poisson(link = log), data = model.data)

Deviance Residuals:
     Min       1Q    Median       3Q      Max
-1.58007  -0.13272  0.00043  0.10282  0.79587

Coefficients:
          Estimate    RS SE z value Pr(>|z|)
Ww1:Yy1    4.83360  0.06535  73.969  < 2e-16 ***
Ww2:Yy1    4.85571  0.06387  76.023  < 2e-16 ***
Ww3:Yy1    4.87418  0.06314  77.199  < 2e-16 ***

< output omitted >

        Null deviance: 25401.0663   Residual deviance:     5.8825
        Number of Fisher Scoring iterations: 4
```

The print.status argument can be changed to TRUE (default) in order to print model fitting information while the function is running. Information typically provided by the glm() function can be extracted from mod.fit.

The formula argument within the Call: portion of the output displays an alternative way that the Y-main effects model could have been specified using variable names. For a model under SPMI (Equation (1)), the syntax -1 + W:Y + wi %in% W:Y + yj %in% W:Y specifies an ordinary loglinear model under independence *within* each 2×2 table formed by the (W_i, Y_j) pairs; i.e., the intercept (W:Y), "row effect" (wi %in% W:Y), and "column effect" (yj %in% W:Y) terms. Note that the addition of wi:yj %in% W:Y would then lead to a saturated loglinear model within the 2×2 tables. Instead, the addition of wi:yj + wi:yj %in% Y allows for the associations to vary across the items in Y (waste storage) but to be the same across items in W (contaminant).

The deviance values in the output should not be used with chi-square distributional approximations to construct traditional model comparison tests. Instead, our anova() method function offers bootstrap and Rao-Scott second-order adjustments (type = "boot" and type = "rs2", respectively, or type = "all" for both methods) to obtain appropriate tests for comparing the model specified in genloglin() to an alternative model given by its model.HA argument. Comparing the Y-main effects model to the saturated model shows moderate evidence of lack-of-fit:

```
> anova(object = mod.fit, model.HA = "saturated", type = "all")

Model comparison statistics for
H0 = y.main
HA = saturated

Pearson chi-square statistic = 5.34
LRT statistic = 5.88

Second-Order Rao-Scott Adjusted Results:
Rao-Scott Pearson chi-square statistic = 10.85, df = 5.23, p = 0.0624
Rao-Scott LRT statistic = 11.96, df = 5.23, p = 0.0409

Bootstrap Results:
```

```
Final results based on 1999 resamples
Pearson chi-square p-value = 0.0385
LRT p-value = 0.0255
```

Our `residuals()` method function provides Pearson standardized residuals, where bootstrap or asymptotic standard errors can be used in their formation. For the Y-main effects model, we find that lack-of-fit occurs for the (W_3, Y_1) association. This suggests the need to estimate a new model that explicitly accounts for the heterogeneity:

```
mod.fit.w3y1 <- genloglin(data = farmer2, I = 3, J = 4, model = count ~ -1 + W:Y +
                          wi %in% W:Y + yj %in% W:Y + wi:yj + wi:yj %in% Y +
                          wi:yj %in% W3:Y1, B = 1999)
```

where the `wi:yj %in% W3:Y1` term forces a perfect fit to the (W_3, Y_1) association while still maintaining a Y-main effects model elsewhere.

Once an appropriate model has been identified, our `predict()` method function can be used to obtain observed and model-estimated odds ratios with corresponding asymptotic and bootstrap BC$_a$ (Efron, 1987) confidence intervals. These odds ratios help to facilitate interpretation of the association among items between the two MRCVs.

Summary

The equivalents of many traditional categorical data analysis methods are implemented within our package in the context of MRCVs. We demonstrated a few of the package's primary functions for analyzing the association between two MRCVs. While not shown here, these functions can be used to analyze MRCVs in other settings. For instance, tests for *multiple marginal independence* (MMI; Agresti and Liu, 1999) between an MRCV and an SRCV can be performed by `MI.test()`, where the I argument is set to a value of 1. An example is given within the help file for this function. Additionally, the MRCV package can be used to analyze the association between three MRCVs. For example, Bilder and Loughin (2007) discuss a third "choose all that apply" question asked of the swine farmers that relates to the farmers' sources of veterinary information. We show in the help file for `genloglin()` how to estimate a generalized loglinear model for this setting.

Agresti and Liu (1999, 2001) show how to take advantage of many commonly used modeling methods (e.g., generalized linear mixed models) for MRCV data. Most of these methods have disadvantages to their use—for example, standard generalized linear mixed models induce a positive correlation between binary responses within subjects, but a negative correlation can occur with MRCV data. Their recommended modeling method, a generalized loglinear model fit through generalized estimating equation (GEE) methodology, can work reasonably well in very large sample sizes (Bilder et al., 2000). The help file for `MI.test()` shows how to use functions in the **geepack** package (Yan et al., 2012) to estimate this model and then subsequently test for MMI via a Wald test.

We envision future additions to the package that will allow for extensions to other situations. For example, "choose all that apply" questions are often asked in complex survey sampling settings. Bilder and Loughin (2009) propose using the same generalized loglinear models, where now different Rao-Scott adjustments are needed to take into account the sampling design. Also, MRCV data can arise over a longitudinal setting, and Suesse and Liu (2013) propose the use of GEE methodology to fit models for this situation. Finally, Nandram et al. (2009) offer a Bayesian perspective to the MMI testing problem. Due to the ubiquitous nature of "choose all that apply" type data formats, we expect there to be many other unique settings where new statistical methods need to be developed. We encourage readers to contact us about their novel methods and/or interest in collaboration.

Bibliography

A. Agresti and I. Liu. Modeling a categorical variable allowing arbitrarily many category choices. *Biometrics*, 55(3):936–943, 1999. [p144, 145, 149]

A. Agresti and I. Liu. Strategies for modeling a categorical variable allowing multiple category choices. *Sociological Methods & Research*, 29(4):403–434, 2001. [p149]

C. Bilder and T. Loughin. On the first-order Rao-Scott correction of the Umesh-Loughin-Scherer statistic. *Biometrics*, 57(4):1253–1255, 2001. [p145]

C. Bilder and T. Loughin. Testing for marginal independence between two categorical variables with multiple responses. *Biometrics*, 60(1):241–248, 2004. [p144, 146]

C. Bilder and T. Loughin. Modeling association between two or more categorical variables that allow for multiple category choices. *Communications in Statistics–Theory and Methods*, 36(2):433–451, 2007. [p144, 146, 149]

C. Bilder and T. Loughin. Modeling multiple-response categorical data from complex surveys. *The Canadian Journal of Statistics*, 37(4):553–570, 2009. [p149]

C. Bilder, T. Loughin, and D. Nettleton. Multiple marginal independence testing for pick any/c variables. *Communications in Statistics–Simulation and Computation*, 29(4):1285–1316, 2000. [p149]

C. Coombs. *A Theory of Data*. John Wiley & Sons, New York, 1964. [p144]

B. Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397): 171–185, 1987. [p149]

S. Gange. Generating multivariate categorical variates using the iterative proportional fitting algorithm. *The American Statistician*, 49(2):134–138, 1995. [p148]

H. Kantarjian, J. Cortes, F. Guilhot, A. Hochhaus, M. Baccarani, and L. Lokey. Diagnosis and management of chronic myeloid leukemia: A survey of American and European practice patterns. *Cancer*, 109(7):1365–1375, 2007. [p144]

N. Koziol and C. Bilder. *MRCV: Methods for Analyzing Multiple Response Categorical Variables*, 2014. URL http://CRAN.R-project.org/package=MRCV. R package version 0.3-1. [p144]

P. Lemons, J. Sedinger, M. Herzog, P. Gipson, and R. Gilliland. Landscape effects on diets of two canids in northwestern Texas: A multinomial modeling approach. *Journal of Mammalogy*, 91(1):66–78, 2010. [p144]

T. Loughin and P. Scherer. Testing for association in contingency tables with multiple column responses. *Biometrics*, 54(2):630–637, 1998. [p144]

B. Nandram, M. Toto, and M. Katzoff. Bayesian inference for a stratified categorical variable allowing all possible category choices. *Journal of Statistical Computation and Simulation*, 79(2):161–179, 2009. [p149]

Office of Management and Budget. Revisions to the standards for the classification of federal data on race and ethnicity. *Federal Register*, 62:58781–58790, 1997. [p144]

B. Riegel, D. Moser, M. Powell, T. Rector, and E. Havranek. Nonpharmacologic care by heart failure experts. *Journal of Cardiac Failure*, 12(2):149–153, 2006. [p144]

S. Riemer, B. Wright, and R. Brown. Food habits of Steller sea lions (*Eumetopias jubatus*) off Oregon and northern California, 1986–2007. *Fishery Bulletin*, 109(4):369–381, 2011. [p144]

T. Suesse and I. Liu. Modelling strategies for repeated multiple response data. *International Statistical Review*, 81(2):230–248, 2013. [p149]

D. Thomas and Y. Decady. Testing for association using multiple response survey data: Approximate procedures based on the Rao-Scott approach. *International Journal of Testing*, 4(1):43–59, 2004. [p146]

B. Wright. Use of chi-square tests to analyze scat-derived diet composition data. *Marine Mammal Science*, 26(2):395–401, 2010. [p144]

J. Yan, S. Hojsgaard, and U. Halekoh. *geepack: Generalized Estimating Equation Package*, 2012. URL http://CRAN.R-project.org/package=geepack. R package version 1.1-6. [p149]

Natalie A. Koziol
Department of Statistics; Department of Educational Psychology
University of Nebraska-Lincoln
Lincoln, NE, United States
nak371@gmail.com

Christopher R. Bilder
Department of Statistics
University of Nebraska-Lincoln
Lincoln, NE, United States
chris@chrisbilder.com