Methods and Overview of Using EdSurvey for Multivariate Regression*†

Developed by Alex Lishinski

October 31, 2018

Multivariate Regression With mvrlm.sdf

Multivariate multiple regression (MMR) is a technique that extends multiple linear regression to include models with multiple outcome variables. In EdSurvey, this technique can be used with `mvrlm.sdf`, which accounts for the complex sample design of National Center for Education Statistics (NCES) data, similar to the lm.sdf function in `EdSurvey`.

This document presents an example of fitting an MMR model to the National Assessment of Educational Progress (NAEP) primer data with students' sex (`dsex`) and the frequency of talking about studies at home (`b017451`) as predictors. The outcome variables are the `algebra` and `geometry` subscales, each with a set of five plausible values, and the weight variable is the full sample weight `origwt`. Taylor series variance estimation is not currently available for multivariate regression, so variance is estimated using the jackknife method, and the jackknife replicate weights are read in as well.

The `|` symbol is used in the model formula to separate the multiple outcome variables. The sampling weight is the default weight for the dataset, unless otherwise specified. In the example below, the `mvrlm.sdf` function uses `origwt` as the default weight for the NAEP primer data. The `weightVar` argument can be used to manually specify the sampling weight. The `mvrlm.sdf` function also allows users to recode a predictor or reset the reference level of a categorical predictor. Please consult the manual for more information about these features.

```
library(EdSurvey)

sdf <- readNAEP(system.file("extdata/data", "M36NT2PM.dat", package = "NAEPprimer"))

mlm1 <- mvrlm.sdf(algebra | geometry ~ dsex + b017451, data = sdf)
summary(mlm1)
```

```
##
## Formula: algebra | geometry ~ dsex + b017451
##
## jrrIMax:
## Weight variable: 'origwt'
## Variance method:
## JK replicates: 62
## full data n: 17606
## n used: 16331
##
## Coefficients:
##
## algebra
##                                coef        se         t    dof  Pr(>|t|)
## (Intercept)                 272.20056   1.04042 261.62628 50.062 < 2.2e-16
```

*This publication was prepared for NCES under Contract No. ED-IES-12-D-0002 with the American Institutes for Research. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

†The author would like to thank Dan Sherman and Mike Cohen for reviewing this document.

```
## dsexFemale                    -0.87023   0.60461  -1.43933 59.273 0.1553167
## b017451Once every few weeks    4.56270   1.26649   3.60263 51.008 0.0007139
## b017451About once a week      11.57610   1.35723   8.52921 54.296 1.335e-11
## b0174512 or 3 times a week    14.79453   1.22471  12.08004 60.169 < 2.2e-16
## b017451Every day               8.20413   1.29110   6.35435 48.096 7.137e-08
##
## (Intercept)                 ***
## dsexFemale
## b017451Once every few weeks ***
## b017451About once a week    ***
## b0174512 or 3 times a week  ***
## b017451Every day            ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## geometry
##                                 coef        se         t    dof  Pr(>|t|)
## (Intercept)                 268.15888   1.02654 261.22555 43.224 < 2.2e-16
## dsexFemale                   -1.46909   0.70813  -2.07460 54.956  0.042720
## b017451Once every few weeks   3.89595   1.15020   3.38718 66.796  0.001187
## b017451About once a week      8.92464   1.28604   6.93961 63.715 2.422e-09
## b0174512 or 3 times a week   12.73099   1.12302  11.33636 64.863 < 2.2e-16
## b017451Every day              5.91482   1.29863   4.55465 45.929 3.860e-05
##
## (Intercept)                 ***
## dsexFemale                  *
## b017451Once every few weeks **
## b017451About once a week    ***
## b0174512 or 3 times a week  ***
## b017451Every day            ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual correlation matrix:
##
##          algebra geometry
## algebra    1.000    0.853
## geometry   0.853    1.000
##
## Multiple R-squared by dependent variable:
##
##  algebra geometry
##   0.0215   0.0172
```

The summary method outputs a separate coefficient table for each outcome variable. The residual correlation matrix is also provided, as well as the R-squared value for each outcome variable.

Method Details

In the case of multivariate regression of the form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$$

where $\boldsymbol{Y}$ is a matrix of n observations on s dependent variables; $\boldsymbol{X}$ is a matrix with columns for k+1

independent variables; $\boldsymbol{B}$ is a matrix of regression coefficients, one column for each dependent variable; $\boldsymbol{E}$ is a matrix of errors, a weighted regression is used so that the estimated coefficients ($\hat{\boldsymbol{B}}$) minimize the trace of the weighted residual sum of squares and cross products matrix:

$$\hat{\boldsymbol{B}} = \text{ArgMin}_{\boldsymbol{B}} \ \ tr((\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})^T \boldsymbol{W} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}))$$

where $\boldsymbol{X}_i$ is the ith row of $\boldsymbol{X}$, $\boldsymbol{Y}_i$ is the ith row of $\boldsymbol{Y}$, $\boldsymbol{W}$ is a diagonal matrix of the weights, and $\text{ArgMin}_{\boldsymbol{B}}$ means the value of $\boldsymbol{B}$ that minimizes the expression that follows it.

Estimation

The methods used to estimate coefficients, variance, and covariance for multivariate multiple regression are largely similar to those used in univariate multiple regression.

Coefficient Estimation

The coefficient estimation in `mvrlm.sdf` produces the same coefficient estimates as when the regressions are run separately using `lm.sdf`, and the details of these methods can be found in the vignette titled Statistics.

Variance Estimation

The variance estimation in `mvrlm.sdf` produces the same standard error estimates as when the regressions are run separately using `lm.sdf`, and the details of these methods can be found in the vignette titled Statistics.

When the predicted value does not have plausible values, the variance of the coefficients is estimated according to the section, "Estimation of Standard Errors of Weighted Means When Plausible Values Are Not Present, Using the Jackknife Method."

When plausible values are present, the variance of the coefficients is estimated according to the section "Estimation of Standard Errors of Weighted Means When Plausible Values Are Present, Using the Jackknife Method."

Residual Variance-Covariance Matrix Estimation

In addition to estimation of the regression coefficients for each dependent variable, the MMR model also produces residual covariance estimates for the dependent variables. The residual variance-covariance matrix is a $s \times s$ matrix for a model with s dependent variables that summarizes residuals within and between dependent variables.

The residuals for the ith dependent variable are calculated as follows:

$$\boldsymbol{R_i} = \boldsymbol{Y_i} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$$

where $\boldsymbol{Y_i}$ is the $p \times n$ matrix of plausible values for the ith dependent variable, $\boldsymbol{X}$ is the $k \times n$ matrix of independent variables, and $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ is the $p \times k$ matrix of estimated coefficients for the p plausible values and the k independent variables. When the ith dependent variable has no plausible values, $\boldsymbol{R_i}$ is simply the vector of residuals for that variable.

To calculate the residual variance-covariance matrix, residuals must be summarized across plausible values. For dependent variables with plausible values, the mean residual is taken across the plausible values for each observation, and the residual value is simply taken for dependent variables without plausible values. The residual vector for the ith dependent variable is calculated as follows:

$$E_i = \frac{1}{p}\sum_{a=1}^{p} r_a$$

where r_a is the ath column of the matrix of residuals $\boldsymbol{R_i}$ for the ith dependent variable. When the ith dependent variable has no plausible values, $\boldsymbol{E_i}$ is simply the vector of residuals for that variable.

The $s \times s$ residual variance-covariance matrix is then calculated from the residual vectors for each dependent variable as follows:

$$\begin{bmatrix} E_1^T E_1 & E_1^T E_2 & \dots & E_1^T E_s \\ E_2^T E_1 & E_2^T E_2 & \dots & E_2^T E_s \\ \vdots & \vdots & \ddots & \vdots \\ E_s^T E_1 & E_s^T E_2 & \dots & E_s^T E_s \end{bmatrix}$$

Coefficient Variance-Covariance Matrix Estimation

The `vcov` method can be used to find the coefficient variance-covariance matrix for MMR models. The coefficient variance-covariance matrix is calculated using the methods detailed in the vignette titled Statistics, where the imputation and sampling variance components are calculated separately and then summed to form the variance-covariance matrix.

In the univariate case, the coefficient matrix is a $k \times k$ symmetric matrix for a model with k regression coefficients, whereas the variance-covariance matrix for the multivariate case is a $sk \times sk$ symmetric block matrix, where the $k \times k$ blocks on the diagonal represent the variance-covariance values within each dependent variable (these values match those in the variance-covariance matrix from a corresponding univariate model), whereas the $k \times k$ off-diagonal blocks represent the variance-covariance values across dependent variables.

$$\begin{bmatrix} \boldsymbol{V_1} & \boldsymbol{C_{1,2}} & \dots & \boldsymbol{C_{1,s}} \\ \boldsymbol{C_{2,1}} & \boldsymbol{V_2} & \dots & \boldsymbol{C_{2,s}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{C_{s,1}} & \boldsymbol{C_{s,2}} & \dots & \boldsymbol{V_s} \end{bmatrix}$$

The diagonal blocks $\boldsymbol{V_i}$ are $k \times k$ matrices of the following form for the ith dependent variable:

$$V_i = V_{jrr} + V_{imp}$$

When the variable does not have plausible values, V_{imp} is 0. The imputation and sampling variance components are calculated as indicated by the "Estimation of Covariances" section in the vignette titled Statistics.

The off-diagonal blocks $C_{a,b}$ are $k \times k$ matrices of the following form for dependent variables a and b:

$$C_{a,b} = C_{jrr} + C_{imp}$$

When one variable does not have plausible values, C_{imp} is 0. The imputation and sampling covariance components are calculated as indicated by the "Estimation of Covariances" section in the vignette titled Statistics.