Using the Inverse Transform to Specify Contrasts in Regression and Latent Curve Structural Equation Models

Thomas N Templin

Wayne State University, College of Nursing, Office of Health Research,
5557 Cass Ave. Detroit, MI 48202

Abstract

A simple yet general method for specifying contrasts to test hypotheses in regression and latent curve structural equation models is presented. The traditional qualitative variable coding schemes used in multiple regression (e.g., dummy coding) have a more general formulation. Five matrices are involved: The coding scheme, A. The matrix which gives the distribution and ordering of cases, W; WA = X; X is the design matrix. The contrast coefficient matrix C; and C^{-1} = A. In practice, only C, C^{-1}, and A are necessary because the statistical software generates the design matrix. This method has great generality because the same coding matrix, A, is used in multiple regression, multilevel modeling, and latent curve structural equal models. Starting with the contrasts allows one to compute the coding matrix, A, for a wide variety specific hypothesis.

Key Words: Structural Equation Models, Categorical Predictors, Growth Curve Models, Contrasts for Linear Models

1. Introduction

A well rehearsed line from regression methodology is that the number of contrasts is one less than the number of categories. Thus a coding scheme for a categorical variable with m levels would be displayed in a matrix with m rows representing the categories and $m-1$ columns representing variables (e.g., Cohn and Cohn, 1983, p. 183). If a linearly independent column vector is added to the coding scheme, the resulting matrix, say A, is $m\,x\,m$ invertible. It turns out that for certain choices of the column vector, say the constant 1, the rows of A^{-1} correspond to coefficients for linear combinations of cell means. Reversing this operation makes it possible to determine a coding schemes from a set of contrasts defined in terms of cell means, $C\ (m\,x\,m)$. There is some ambiguity in how the coefficients in the first row of C are to be determined and the effect of this coding on the obtained contrasts.

1.1 Example: Contrast Definitions from Dummy Codes

A command in R can be used to generate dummy codes. Some documentation refers to these as “treatment contrasts” and to the resulting matrix as a “contrast matrix”. This terminology is contradictory to the definitions being offered here. The treatment contrasts created in R with the addition of a column vector of 1’s is an example of the regression variable coding matrix, A. I show below that A^{-1} = C, the contrasts of interest.

```
> contr.treatment(4)
          2 3 4
1         0 0 0
2         1 0 0
3         0 1 0
4         0 0 1
```

If we add the intercept we get an m x m coding scheme, A, that can be used to generate a design matrix:

```
>A = matrix ( c(
1,  0,  0,  0,
1,  1,  0,  0,
1,  0,  1,  0,
1,  0,  0,  1), byrow=T, ncol=4)

>A
  [,1] [,2] [,3] [,4]
[1,]   1   0   0   0
[2,]   1   1   0   0
[3,]   1   0   1   0
[4,]   1   0   0   1

> C = solve(A)
    [,1] [,2] [,3] [,4]
[1,]   1    0    0    0
[2,]  -1    1    0    0
[3,]  -1    0    1    0
[4,]  -1    0    0    1
```

This is clearly the coefficient matrix of interest. It shows that the dummy coding scheme compares the first factor level with each succeeding level. The meaning of the first row is ambiguous. It appears that this code identifies the first cell as the reference cell.

1.2 Coding schemes for longitudinal data

The coding schemes that are used to create contrasts of category means can be used to create contrasts for longitudinal or repeated measures data as well. For example, the coding scheme for dummy variables when applied to longitudinal data can be made to compare each successive wave of data with the first wave or the last wave.

1.3 Novel Codes from Contrasts: Latent Curve Multiple Baseline Model

Using the fact that $C^{-1} = A$, the desired coding matrix for a wide variety of different hypotheses can be determined. For example, suppose a multiple baseline design with two baseline and one follow-up measure. The hypothesis is that the average of the baseline assessment differs from the follow-up assessment. The desired contrast is in the third row of C below:

$$C = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -.5 & -.5 & 1 \end{pmatrix}$$

The mean of the first two periods is compared to the mean of the third period. The second contrast is also of some interest because it compares the first two periods which are expected to be equal.

The coding matrix, A, for this is: $C^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & .5 & 1 \end{pmatrix}$.

The model and codes (in red) are shown in Figure 1 below. The estimated coefficient for the first column of A was the mean of ξ_1. This is shown in the figure as μ_{ξ_1} (= 1.45). The estimated coefficient for the third column of A, the contrast of interest, was μ_{ξ_3} (= -.01). This coefficient is the difference in latent means expected from the contrast specified in row 3 of C above. The latent means were 1.4745, 1.481, 1.4665, respectively. The value of the contrast computed from the latent means was exactly as expected. Data published in Steyer, et al (2000) was used in this example.

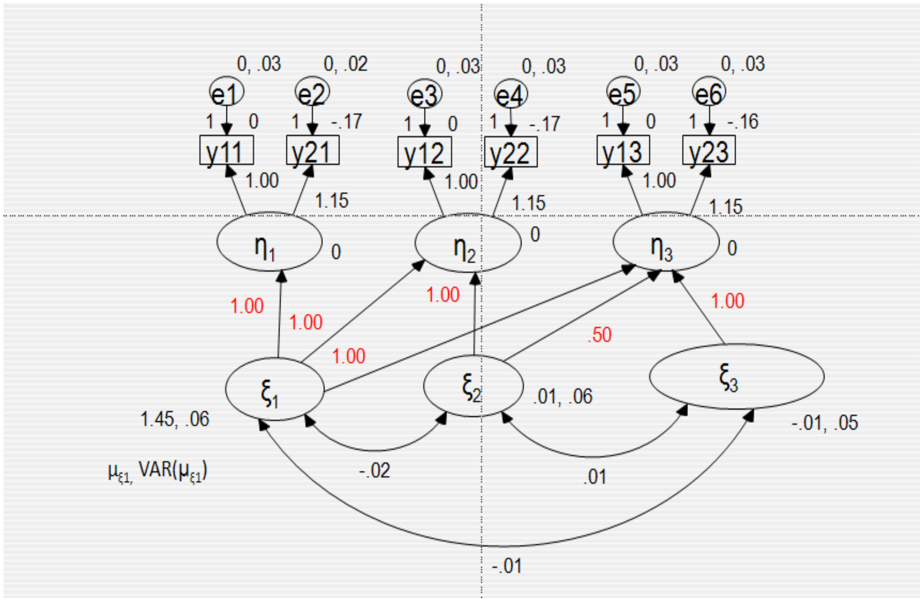

Figure 1. Latent curve multiple baseline model showing coefficients from the coding matrix, A, in red. Note that columns in A correspond to ξ variables and rows, to η variables so $a_{33} = 1$ is the regression of η_3 *on* ξ_3.

1.4 Additional Background

I explored these relationship and described a procedure to create customized contrasts in longitudinal SEM models (2007, 2008a, 2008b) including MR and latent curve models (LCMs). Recently, Axel Meyer and colleagues (Mayer, Steyer, & Mueller, 2012) described the use of the inverse transform to code quantitative (e.g., orthogonal polynomials) and qualitative research factors for first and second order LCMs. I could not find an explanation of why the inverse transform worked that was not tautological. In the remainder of this article I develop a more general formulation of contrast coding using the inverse transform.

2. The General Formulation

The regression variable coding schemes are not intuitive. David Nichols (1997) gave the following explanation:

$$C^* = (X'X)^{-1}X'$$

where X is the design matrix and C^* is the contrast for an individual case.

This suggests that a general formulation will need to define X in addition to C, and A. The design matrix is determined by the coding scheme, A, and the ordering and distribution of cases, W. Ordering of cases is design dependent. For a between group design with equal n,

$$W \ = I_n \ \otimes 1_n.$$

To make computations simple, assume three groups with 2 cases in Group 1 and one case in each of the other groups.

$$W = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \ (4x3)$$

Dummy coding for a three group design, with the third group as the reference cell is as follows:

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

So the design matrix is:

$$X = WA = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \ (4x3)$$

The R code follows:

```
> A = matrix( c(1,1,0,  1,0,1,   1,0,0), nrow=3, byrow=T)
> A
    [,1] [,2] [,3]
[1,]    1    1    0
[2,]    1    0    1
[3,]    1    0    0

> X = W%*%A
> X
    [,1] [,2] [,3]
[1,]    1    1    0
[2,]    1    1    0
[3,]    1    0    1
[4,]    1    0    0
```

The contrast coefficients that apply to each y_i are given by:

C* = $(X'X)^{-1}X'$

```
> C* = solve(t(X)%*%X)%*%t(X) , the asterisk was added for consistency in notation
> C*
    [,1] [,2] [,3] [,4]
[1,]  0.0  0.0    0    1
[2,]  0.5  0.5    0   -1
[3,]  0.0  0.0    1   -1
```

These are the weights applied to y that give the contrasts. To see the actual contrast coefficients, post multiply by W.

```
> (C*)%*%W
    [,1] [,2] [,3]
[1,]    0    0    1
[2,]    1    0   -1
[3,]    0    1   -1
```

C = the contrast matrix, the weights applied to cell means.

$$C = \begin{array}{cccc} . & Group\ 1 & Group\ 2 & Group\ 3 \\ \mathrm{C}_1 & 0 & 0 & 1 \\ \mathrm{C}_2 & 1 & 0 & -1 \\ \mathrm{C}_3 & 0 & 1 & -1 \end{array}$$

3. Discussion and Conclusions

The inverse transform can be used to determine the coding scheme from meaningful contrast specifications for a wide variety of linear models. To facilitate further investigations of this approach to coding linear models, I developed a general formulation. Five matrices are involved: The coding scheme, A. The matrix which gives the distribution and ordering of cases, W; WA = X; X is the design matrix. The contrast coefficient matrix C; and C^{-1} = A. In practice, only C, C^{-1}, and A are necessary because the statistical software generates the design matrix.

In the example, W was used to give the distribution of cases within groups. It is expected that the form of W will depend on the nature of the design: In a within-subjects design, the rows of A will be distributed in X in a case-wise fashion, the codes are nested within person. In this case, W takes the form of stacked identity matrix.

Further research is needed: What is the effect of the row 1 non-contrast coefficients in C in determining what effects are estimated? It is expected that unequal n per group/level will estimate different effects than the equal n case. How can unequal n be taken into account? And finally, how should contrast definitions be applied to individual cases and groups simultaneously?

A. Starting with the observed data and design matrix, X, show that $\boldsymbol{C = A^{-1}}$

Rewrite the equation, C* = $(X'X)^{-1}X'$, with X = WA and multiple by W to show that C = A^{-1} and consequently that AC = I.

$$C = C^{*}W = \left((WA)'(WA)\right)^{-1}(WA)'W$$

$$C = C^{*}W = \left((A'W')(WA)\right)^{-1}(A'W')W$$

The transpose of a product is the transpose of the factors in reverse order

$$C = C^{*}W = (A'W'WA)^{-1}A'W'W$$

Clearing parentheses

$$C = C^{*}W = (A'DA)^{-1}A'D$$

Replacing $W'W$ with D a diagonal matrix of cell sizes

$$C = C^{*}W = (A^{-1}\,D^{-1}(A')^{-1})\,A'D$$

The inverse of a product is the product of the inverses in reverse order.

$$C = C^{*}W = A^{-1}\,D^{-1}(A')^{-1}\;\;A'D$$

Clear parentheses and identify cancelations

$$C = C^{*}W = A^{-1}$$

$AC = I$ Q.E.D.

Acknowledgements

I gratefully acknowledge the Wayne State University College of Nursing Office of Health Research and the Dean of Research, Nancy Artinian, for support of this research.

This research was supported by grants from the National Institutes of Health: 5P30 AG015281 (Jackson, PI), NIH/NIDA R15 DA03282201 (Rice, PI); and NIH/NCI 1R01CA149432-01A1 (Vallerand, PI) and by the Michigan Center for Urban African American Aging Research (Jackson and Lichtenberg, PIs)

References

Cohen, J., & Cohen, P. (1983). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (Second ed.). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Mayer, A., Steyer, R., & Mueller, H. (2012). A general approach to defining latent growth components. *Structural Equation Modeling: A Multidisciplinary Journal, 19(4), 513-533.*

Nichols, D. P. (1997). What kind of contrasts are these? UCLA: Statistical Consulting Group. Downloaded from the web page, http://www.ats.ucla.edu/stat/stata/ado/analysis/. (July 25, 2013). Originally from SPSS Keywords, Number 63, 1997.

Steyer, R., Partchev, I., & Shanahan, M. J. (2000). Modeling true intradindividual change in structural equation models: The case of poverty and children's psychosocial adjustment. In T. D. Little, K. U. Schnabel & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples.* (pp. 109-126, 269-281). Mahwah: Lawrence Erlbaum Associates.

Templin, T. N. (2008). A Simple Way to Specify Contrasts in Structural Equation Models of Longitudinal Data. *Paper presented at the Annual Research Conference of the Midwest Nursing Research Society*, Indianapolis.

Templin, T. N. (2008). Using the Inverse Transform for Coding Specific Longitudinal SEM Hypotheses. *Paper presented at the Annual Convention of the American Psychological Association*, Boston.

Templin, T. N. (2007). Structural Equation Models for Longitudinal Analysis of Health Disparities Data. *Michigan Center for African American Aging Research Workshop* (2hr). Institute for Social Research, University of Michigan. http://mcuaaar.wayne.edu/mmcdoc/mcuaaar5b.ppt