

Multidimensional Latent Regression

Ray Adams and Margaret Wu, 29 August 2010

In tutorial seven, we illustrated how ConQuest can be used to fit multidimensional item response models; and in tutorial five, we illustrated how ConQuest can be used to estimate latent regression models. In this tutorial, we bring these two functions together, using ConQuest to fit multidimensional latent regression models.

In the first half of the tutorial, we fit multidimensional latent regression models of two and five dimensions. Some output that is standard for regression analysis is not available in this version of ConQuest; but in the second half of the tutorial, we illustrate how plausible values can be drawn. The plausible values can be analysed, using traditional regression techniques, to produce further regression statistics.

FITTING A TWO-DIMENSIONAL LATENT REGRESSION

The data we are analysing were collected by Adams, Doig and Rosier (1991) as part of their study of science achievement in Victorian schools. In their study, Adams et al. used a battery of multiple choice and extended response written tests. In this sample analysis, we will consider ability as assessed by the multiple choice test as one latent outcome and ability as assessed by the first of the extended response tests as a second latent outcome. Then we will regress these two outcomes onto three background variables: student grade, student gender and an indicator of socio-economic status.

The data set that we use contains the responses of 2564 students to the battery of tests; all of the items have been prescored. The multiple choice items are located in columns 50 through 114, and the extended response test that we will use is located in columns 1 through 9. If students were administered a test but did not respond to an item, a code of 9 has been entered into the file. If a student was not administered an item, then the file contains a blank character. We will be treating the 9 as an incorrect response and the blanks as missing-response data. The student's grade code is located in column 118, the gender code is located in column 119, and the indicator of socio-economic status is in columns 122 through 127.¹ The gender variable is coded 0 for female and 1 for male, the grade variable is coded 1 for the lower grade and 2 for the upper grade, and the socio-economic indicator is a composite that represents a student's socio-economic status.

¹ See Adams, Doig and Rosier (1991) for how the socio-economic indicator was constructed.

Multidimensional Latent Regression

The files that will be used in this sample analysis are:

ex8a.cqc	The command statements that we use.
ex6.dat	The data.
ex8a.prm	An initial set of item parameter estimates.
ex8a.reg	An initial set of regression coefficient parameter estimates.
ex8a.cov	An initial set of variance-covariance parameter estimates.
ex8a.shw	The population model parameter estimates.

The contents of file `ex8a.cqc` are shown in Figure 1.

```
1. datafile ex6.dat;
2. format responses 1-9,50-114 grade 118 gender 119
   ses 122-127!tasks(74);
3. model tasks+tasks*step;
4. recode (9) (0);
5. score (0,1,2,3,4) (0,1,2,3,4) ( ) ! tasks(1-9);
6. score (0,1,2,3,4) ( ) (0,1,2,3,4) ! tasks(10-74);
7. regression grade,gender,ses;
8. export parameters >> ex8a.prm;
9. export reg_coefficients >> ex8a.reg;
10. export covariance >> ex8a.cov;
11. import init_parameters << ex8a.prm;
12. import init_reg_coefficients << ex8a.reg;
13. import init_covariance << ex8a.cov;
14. set update=yes,warnings=no;
15. estimate!fit=no,converge=.002;
16. show ! tables=3 >> ex8a.shw;
```

Figure 1 Command File for a Two-dimensional Latent Regression

1. We are analysing data in the file `ex6.dat`.
2. The `format` statement is reading 74 responses; assigning the label `tasks` to those responses; and reading `grade`, `gender` and `ses` data. The column specifications for the responses are made up of two separate response blocks. The first nine items are read from columns 1 through 9 (these are the extended response items that we are using), and the remaining 65 items are read from columns 50 through 114 (these are the multiple choice items).
3. We are using the partial credit model because the items are a mixture of polytomous and dichotomous items.
4. A code of 9 has been used for missing-response data caused by the student not responding to an item. We want to treat this as though it were identical to an incorrect response, so we recode it to 0.
- 5-6. We use two `score` statements, one for each dimension. The first statement scores the first nine tasks on the first dimension, and the second statement scores the remaining 65 tasks on the second dimension.
7. This regression statement specifies a population model that regresses the two latent variables onto `grade`, `gender` and `ses`.
- 8.-10. These `export` statements result in the parameter estimates being written to the files `ex8a.prm`, `ex8a.reg` and `ex8a.cov`. In conjunction with the `set` statement (line 14),

these `export` statements result in updated parameter estimates being written to these files after each iteration.

- 11.-13. Initial values of all parameter estimates are read from the files `ex8a.prm`, `ex8a.reg` and `ex8a.cov`. These initial values have been provided to speed up the analyses.
14. In conjunction with the `export` statements (lines 8 through 10), this `set` statement results in updated parameter estimates being written to the files after each iteration, and it turns off warning messages.
15. Begins estimation of the model. The options turn off calculation of the fit tests and instruct estimation to terminate when the change in the parameter estimates from one iteration to the next is less than 0.002.
16. Writes the estimates of the population model parameter estimates to `ex8a.shw`.

RUNNING THE TWO-DIMENSIONAL LATENT REGRESSION

To run this sample analysis, launch the console version of ConQuest by typing the command

```
ConQuestCMD ex8a.cqc
```

ConQuest will begin executing the statements that are in the file `ex8a.cqc`; and as they are executed, they will be echoed on the screen. When ConQuest reaches the `estimate` statement, it will begin fitting the two-dimensional latent multiple regression. This particular sample analysis will converge after a single iteration, because we have provided very accurate initial values.

<p>NOTE: <i>If you run this sample analysis without the initial values, it will take in excess of 1000 iterations to converge. While fitting multidimensional models can take a substantial amount of computing time, this particular analysis will take an unusually large number of iterations because of the sparse nature of the data set. In these data, just 40% of the students responded to items on the first dimension; and the first 50 multiple choice items were responded to by only 25% of the sample. All students responded to the last 15 items.</i></p>

In Figure 2, we report the parameter estimates for the population model used in this analysis. In this case, we have two sets of four regression coefficients—a constant and one for each of the three regressors. The conditional variance-covariance matrix is also reported.

```

=====
ConQuest: Generalised Item Response Modelling Software      Mon Jan 06 06:48:23
TABLES OF POPULATION MODEL PARAMETER ESTIMATES
=====
REGRESSION COEFFICIENTS
-----
Regression Variable      Dimension
                        -----
                        1          2
CONSTANT                 -1.431   -1.727
grade                    0.700    1.391
gender                   0.072    0.229
ses                      0.366    0.479
-----
An asterisk next to a parameter estimate indicates that it is constrained
=====
COVARIANCE/CORRELATION MATRIX
-----
Dimension                Dimension
                        -----
                        1          2
1                        0.411
2                        0.747
-----
Variance                 0.410    0.738
-----
An asterisk next to a parameter estimate indicates that it is constrained
Values below the diagonal are correlations and values above are covariances
=====

```

These are the two sets of regression coefficients, one for each latent dimension.

This is the conditional covariance/correlation matrix for the two-dimensional latent space.

Figure 2 Population Parameter Estimates for a Two-dimensional Latent Multiple Regression

All of the results reported here are in their natural metrics (logits). For example, on the first dimension, the difference between the performances of the lower grade and upper grades is 0.700 logits, the male students outperform the female students by 0.072, and a unit increase in the socio-economic status indicator predicts an increase of 0.366 logits in the latent variable. For the second dimension, the difference between the performances of the lower grade and upper grades is 1.391 logits, the male students outperform the female students by 0.229, and a unit increase in the socio-economic status indicator predicts an increase of 0.479 logits in the latent variable.²

To aid in the interpretation of these results, it is useful to fit a model without the regressors to obtain estimates of the variance of the two latent variables in this model, the multiple choice items and the extended response item. The command file `ex8b.cqc` is provided with the samples for this purpose. If this command file is executed, it will provide estimates of 0.601 (extended response) and 1.348 (multiple choice) for the variances of the two latent variables.

² The current version of ConQuest does not report standardised regression coefficients or standard errors for the regression parameter estimates. Plausible values can be generated (as explained later in this chapter) and analysed to obtain estimates of standard errors and to obtain standardised coefficients.

EXTENSION: *The model fitted in `ex8b.cqc` has the item response model parameters anchored at the values that were obtained from the model that is fit with `ex8a.cqc`. In general, the item response parameter estimates obtained from fitting a model with regressors will produce item parameter estimates that have smaller standard errors, although the gain in efficiency is generally very small. More importantly, there are occasions when item response model parameters estimated without the use of regressors will be inconsistent. This data set provides such a case, because some of the multiple choice items were administered only to students in the upper grade, while others were administered only to students in the lower grade. Readers interested in this issue are referred to Mislevy and Sheehan (1989) and Adams, Wilson and Wu (1997).*

In Table 1, we report the R^2 for each of the dimensions in the latent regression, and we report the grade, gender and socio-economic status (SES) regression coefficients as effect sizes that have been computed by dividing the estimate of the regression coefficients by the unconditional standard deviation of the respective latent variables. The results in the table show that the regression model explains marginally more variance for the multiple choice items than it does for the extended response items. Interestingly, the grade and SES effects are similar for the item types, but the gender effect is larger for the multiple choice items. For the extended response items, the gender difference is 9% of a student standard deviation, whereas for the multiple choice it is 19.7%.

Table 1 Effect Size Estimates for the Two-dimensional Latent Multiple Regression

<i>Estimate</i>	<i>Extended Response</i>	<i>Multiple Choice</i>
R^2	31.8%	45.3%
Grade Effect	0.903	1.198
Gender Effect	0.090	0.197
SES Effect	0.472	0.407

HIGHER-DIMENSIONAL MULTIPLE REGRESSION

In the Adams, Doig and Rosier (1991) battery of tests, four extended response tests and a set of 15 multiple choice were administered to students in both the upper and lower grades. In this higher-dimensional sample analysis, we are interested in grade, gender and SES effects for the five latent dimensions that are assumed to be assessed by these instruments. First, we will run an unconditional model (using the command file in Figure 3) to obtain initial values for a conditional model. Then we will run the conditional model and will also have ConQuest draw plausible values, using the command file in Figure 5.

Because of the high dimensionality, the analysis that is required here is best undertaken with Monte Carlo integration; and as this will need a large number of nodes, the model without regressors (the unconditional model) is fitted in two stages. In the first stage, a small number of nodes with a moderate convergence criterion is used to produce initial values. In the second stage, the initial values are read back into an analysis that uses more nodes and a more stringent convergence criterion.

The contents of the command file, `ex8c.cqc`, are shown in Figure 3.

```
1.  datafile ex6.dat;
2.  format responses 1-18,31-49,100-114 grade 118
      gender 119 ses 122-127!tasks(52);
3.  model tasks+tasks*step;
4.  recode (9) (0);
5.  score (0,1,2,3,4) (0,1,2,3,4) ( ) ( ) ( ) ( ) ! tasks(1-9);
6.  score (0,1,2,3,4) ( ) (0,1,2,3,4) ( ) ( ) ( ) ! tasks(10-18);
7.  score (0,1,2,3,4) ( ) ( ) (0,1,2,3,4) ( ) ( ) ! tasks(19-28);
8.  score (0,1,2,3,4) ( ) ( ) ( ) (0,1,2,3,4) ( ) ! tasks(29-37);
9.  score (0,1,2,3,4) ( ) ( ) ( ) ( ) (0,1,2,3,4) ! tasks(38-52);
10. export reg_coefficients >> ex8c.reg;
11. export covariance >> ex8c.cov;
12. export parameters >> ex8c.prm;
13. set update=yes,warnings=no;
14. estimate!fit=no,method=montecarlo, nodes=400, converge=.01;
15. reset;
16. datafile ex6.dat;
17. format responses 1-18,31-49,100-114 grade 118 gender 119
      ses 122-127!tasks(52);
18. model tasks+tasks*step;
19. recode (9) (0);
20. score (0,1,2,3,4) (0,1,2,3,4) ( ) ( ) ( ) ( ) ! tasks(1-9);
21. score (0,1,2,3,4) ( ) (0,1,2,3,4) ( ) ( ) ( ) ! tasks(10-18);
22. score (0,1,2,3,4) ( ) ( ) (0,1,2,3,4) ( ) ( ) ! tasks(19-28);
23. score (0,1,2,3,4) ( ) ( ) ( ) (0,1,2,3,4) ( ) ! tasks(29-37);
24. score (0,1,2,3,4) ( ) ( ) ( ) ( ) (0,1,2,3,4) ! tasks(38-52);
25. export parameters >> ex8c.prm;
26. import init_reg_coefficients << ex8c.reg;
27. import init_covariance << ex8c.cov;
28. import init_parameters << ex8c.prm;
29. set update=yes,warnings=no;
30. estimate!method=montecarlo, nodes=2000, converge=.002;
31. show !tables=1:3:5 >> ex8c.shw;
```

Figure 3 Command File for Fitting the Five-dimensional Latent Unconditional Model

1. We are using the data in `ex6.dat`.
2. The responses to the four extended response instruments administered to all the students are in columns 1 through 18 and 31 through 49; and the responses to the 15 multiple choice items administered to all the students are in columns 100 through 114. Columns 19 through 30 contain the responses to an instrument that was administered to the lower grade students only, and columns 50 through 99 contain the responses to multiple choice items that were administered to students in one of the grades only. We have decided not to include those data in these analyses.
3. We are using the partial credit model.
4. Any code of 9 (item not responded to by the student) will be recoded to 0 and therefore scored as 0.
- 5.-9. These five `score` statements allocate the items that make up the five instruments to the five different dimensions.

- 10.-13. The `export` statements, in conjunction with the `set` statement, ensure that the parameter estimates are written to the files `ex8c.reg`, `ex8c.cov` and `ex8c.prm` after each iteration. This is useful if you want to use the values generated by the final iteration as initial values in a further analysis, as we will do here.
14. Initiates the estimation of a partial credit model using the Monte Carlo method to approximate multidimensional integrals. This estimation is done with 400 nodes, a value that will probably lead to good estimates of the item parameters, but the latent variance-covariance matrix may not be well estimated.³ We are using 400 nodes here to obtain initial values for input into the second analysis that uses 2000 nodes. We have specified `fit=no` because we will not be generating any displays and thus have no need for this data at this time. We are also using a convergence criteria of just 0.01, which is appropriate for the first stage of a two-stage estimation.
15. The `reset` statement resets all variables to their initial values and is used to separate distinct analyses that are in a single command file.
- 16.-24. As for lines 1 through 9 above.
25. We are exporting only the item response model parameter estimates.
- 26.-28. Initial values for all of the parameter estimates are being read from the files that were written in the previous analysis.
29. Used in conjunction with line 25 to ensure that the item response model parameter estimates are written after each iteration.
30. The estimation method is Monte Carlo, but this time we are using 2000 nodes and a convergence criterion of 0.002. This should be sufficient to produce accurate estimates for all of the parameters.
31. Writes selected tables to the output file `ex8c.shw`.

RUNNING THE FIVE-DIMENSIONAL LATENT UNCONDITIONAL SAMPLE ANALYSIS

To run this sample analysis, launch the console version of ConQuest by typing the command

```
ConQuestCMD ex8c.cqc
```

ConQuest will begin executing the statements in the file `ex8c.cqc`; and as they are executed, they will be echoed on the screen. When ConQuest reaches the first `estimate` statement, it will begin fitting the five-dimensional model using a 400-node Monte Carlo integration. It will execute for 50 iterations and then terminate because the deviance is no longer improving. ConQuest will then proceed to analyse the data again using a 2000-node Monte Carlo integration, reading initial values from the export files produced by the previous 400-node analysis. It will take 15 iterations for the convergence criterion of 0.002 to be attained.

Figure 4 shows the estimated population parameters for the unconditional five-dimensional latent space. The analysis shows that the correlation between these latent dimensions is moderately high but unlikely to be high enough to justify the use of a unidimensional model.

³ Simulation studies (Volodin and Adams, 1995) suggest that 1000 to 2000 nodes may be needed for accurate estimation of the variance-covariance matrix.

EXTENSION: *To test whether a unidimensional model would be appropriate for these data, you could fit a unidimensional model and then compare the deviance of the unidimensional model with that of the five-dimensional model. (For an example of a population model parameter estimate for a unidimensional model, see Figure 7-3.) The difference in the deviance of the models can be assumed to follow a chi-squared distribution with 14 degrees of freedom. The 14-parameter difference comes from having four additional means, four fewer item parameters and 14 additional elements in the variance-covariance matrix.*

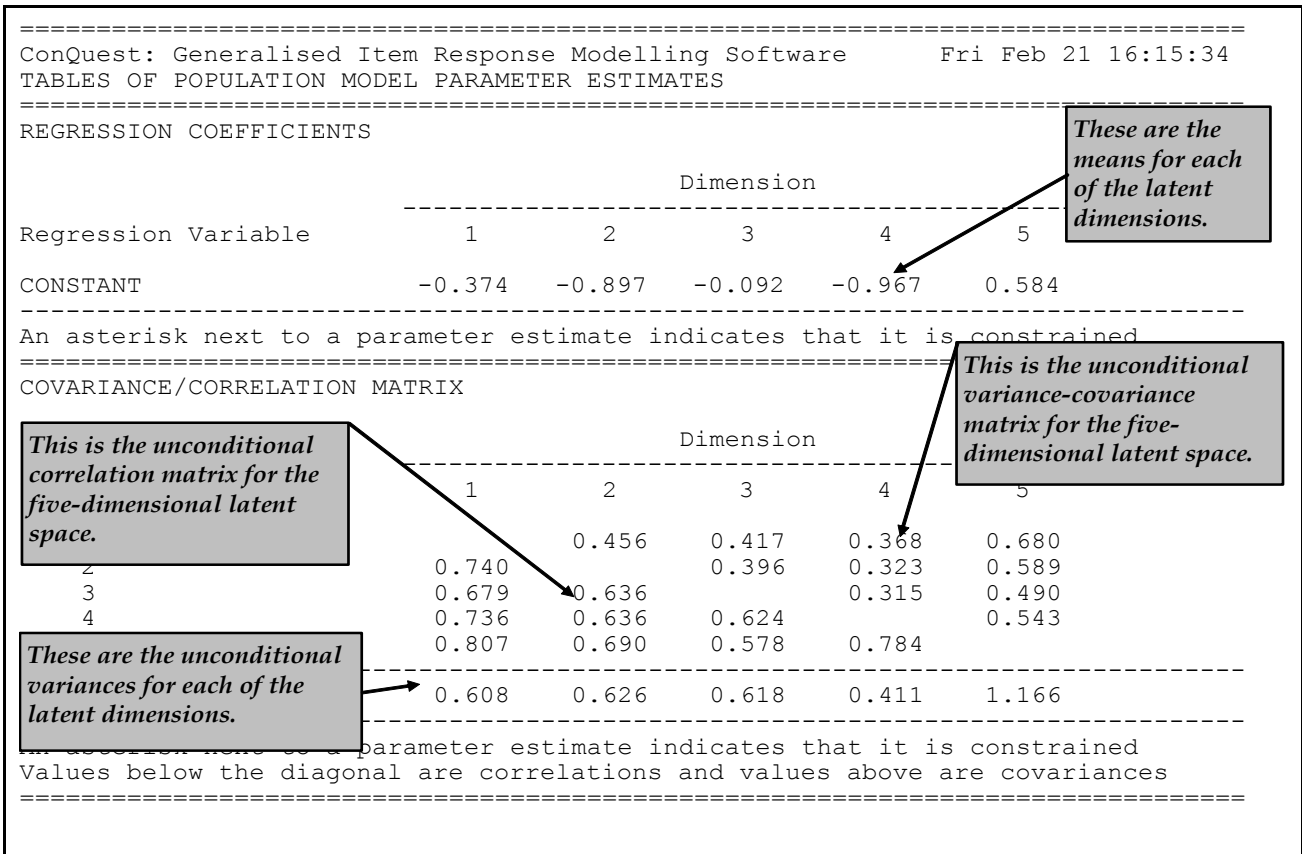


Figure 4 Population Parameter Estimates for the Unconditional Five-dimensional Model

The command file for fitting the five-dimensional latent regression model (the conditional model) is given in Figure 5. The command file is very similar to that used for the unconditional analysis, so in describing the code we will focus only on the differences.


```

1. datafile ex6.dat;
2. format responses 1-18,31-49,100-114 grade 118
   gender 119 ses 122-127!tasks(52);
3. regression grade, gender, ses;
4. model tasks+tasks*step;
5. recode (9) (0);
6. score (0,1,2,3,4) (0,1,2,3,4) ( ) ( ) ( ) ( )
   ! tasks(1-9);
7. score (0,1,2,3,4) ( ) (0,1,2,3,4) ( ) ( ) ( )
   ! tasks(10-18);
8. score (0,1,2,3,4) ( ) ( ) (0,1,2,3,4) ( ) ( )
   ! tasks(19-28);
9. score (0,1,2,3,4) ( ) ( ) ( ) (0,1,2,3,4) ( )
   ! tasks(29-37);
10. score (0,1,2,3,4) ( ) ( ) ( ) ( ) (0,1,2,3,4)
   ! tasks(38-52);
11. import anchor_parameters << ex8c.prm;
12. import init_covariance << ex8c.cov;
13. estimate!method=montecarlo,nodes=2000, conv=.002;
14. show cases ! estimates=latent >> ex8d.pls;
15. show cases ! estimates=eap >> ex8d.eap;
16. show !tables=1:3:5>>ex8d.shw;
17. quit;

```

Figure 5 Command File for Fitting the Five-dimensional Latent Regression Model and Drawing Plausible Values (ex8d.cqc)

- 1.-2. No comment.
3. The third statement in this command file specifies the regression variables that are to be used in the model (in this case, grade, gender and ses).
- 4.-10. No comment.
11. This `import` statement requests that item response model parameter values be read from the file `ex8c.prm` (created by the five-dimensional unconditional model) and be anchored at the values specified in that file. This means that, in this analysis, we will not be estimating item parameters.

WARNING: *The current version of ConQuest is unable to estimate both item response model parameters and population model parameters in a conditional model (that is, a model with regressors) when the Monte Carlo method is used. This will not usually be a severe limitation because you can generally obtain consistent estimates of the item parameters by fitting an unconditional model and then entering those estimates as anchored values in a conditional model.*

12. This `import` statement uses the estimated unconditional variance-covariance matrix as an initial value. This is done in this sample analysis so that the analysis will be performed more quickly.

Multidimensional Latent Regression

13. The estimation will be done with the Monte Carlo method, using 2000 nodes and a convergence criterion of 0.002.
- 14.-15. These show statements result in plausible values and expected a-posteriori estimates being written to the files ex8d.pls and ex8d.eap respectively.
16. The final show statement requests tables 1, 3 and 5 be written to file ex8d.shw.

RUNNING THE FIVE-DIMENSIONAL LATENT REGRESSION SAMPLE ANALYSIS

To run this sample analysis, launch the console version of ConQuest by typing the command

```
ConQuestCMD ex8d.cqc
```

ConQuest will begin executing the statements in the file `ex8d.cqc`; and as they are executed, they will be echoed on the screen. When ConQuest reaches the `estimate` statement, it will begin fitting the five-dimensional model, using a 2000-node Monte Carlo integration. This analysis will take 29 iterations for the convergence criterion of 0.002 to be attained. The `show` statements will then be executed, producing files of plausible values, expected a-posteriori ability estimates and output tables. Extracts from the first two files are shown in Figures 6 and 7.

NOTE:

The expected a-posteriori and plausible value files contain values for all cases on all dimensions—even for latent dimensions on which the cases have not responded to any questions. If there are dimensions for which one or more cases have not made any response, then maximum likelihood ability estimates of the latent variable cannot be calculated.

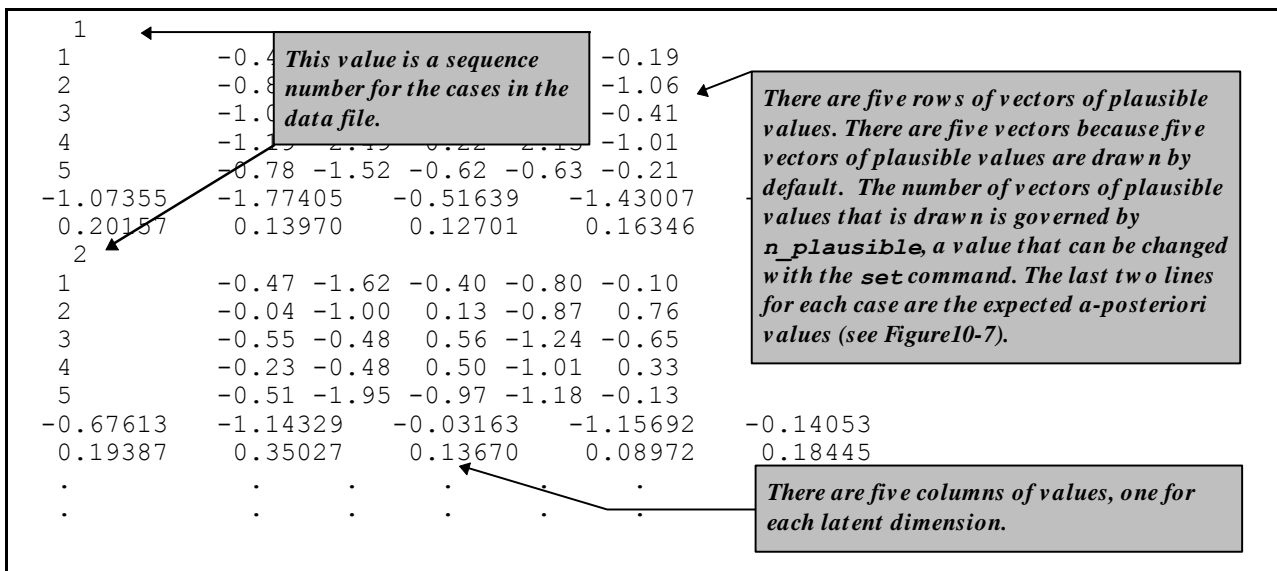


Figure 6 Extract from the File of Plausible Values

1	-1.09143	-1.759	0.17039	0.135	0.52452	0.19183
2	-0.67869	-1.12420	0.19116	0.32623	-0.02075	-1.15043
3	-1.04968	-1.55069	0.11481	0.35155	-0.58795	-1.61956
.
.

This value is a sequence number for the cases in the file.

There are two rows of values. The first contains the EAP ability estimates for each dimension, and the second contains the error variance of the posterior distributions.

There are five columns of values, one for each latent dimension.

Figure 7 Extract from the File of Expected A-posteriori Values⁴

Figure 8 shows the estimates of the parameters of the population model. It contains estimates of the four regression coefficients for each of the latent dimensions and the estimate of the conditional variance-covariance matrix between the dimensions. This variance-covariance matrix is also expressed as a correlation matrix.

```

=====
ConQuest: Generalised Item Response Modelling Software      Thu Jan 09 11:31:07
TABLES OF POPULATION MODEL PARAMETER ESTIMATES
=====
REGRESSION COEFFICIENTS

```

Regression Variable	Dimension				
	1	2	3	4	5
CONSTANT	-1.459	-1.959	-0.921	-1.951	-1.459
grade	0.726	0.705	0.556	0.701	1.334
gender	0.100	0.076	0.047	-0.009	0.248
ses	0.367	0.265	0.260	0.207	0.461

These are the regression coefficients for each of the latent dimensions.

An asterisk next to a parameter estimate indicates that it is constrained

```

=====
COVARIANCE/CORRELATION MATRIX

```

Dimension	Dimension				
	1	2	3	4	5
5					
4	0.652	0.290	0.305	0.188	0.365
3	0.656	0.550	0.284	0.166	0.311
2	0.582	0.462	0.476	0.179	0.264
1	0.722	0.553	0.449	0.653	0.266
	0.400	0.493	0.542	0.261	0.638

This is the conditional correlation matrix for the five-dimensional latent space.

These are the conditional variances for each of the latent dimensions.

This is the conditional variance-covariance matrix for the five-dimensional latent space.

a parameter estimate indicates that it is constrained
values below the diagonal are correlations and values above are covariances
=====

Figure 8 Population Model Parameter Estimates for the Five-dimensional Latent Regression

⁴ The EAP values in Figures 6 and 7 are not the same, because ConQuest selects a different random number generator seed each time EAP values are generated.

In Table 2, the estimates of the regression coefficients have been divided by the estimate of the unconditional standard deviation of the respective latent variables to provide effect size estimates. Combining the unconditional results that were obtained from analysing the data with the command file `ex8c.cqc` and were reported in Figure 4 with the latent regression results produced using the command file `ex8d.cqc` and reported in Figure 8, we obtain the effect size estimates reported in Table 2. Additional analyses of this latent regression model can be obtained by merging the EAP ability estimates and the plausible values with the background variables (such as gender or grade) and undertaking conventional analyses.

Table 2 Effect Size Estimates for the Five-dimensional Latent Multiple Regression

<i>Estimate</i>	<i>Force & Motion</i>	<i>Light & Sight</i>	<i>Matter</i>	<i>Earth & Space</i>	<i>Multiple Choice</i>
R^2	34%	21%	12%	36%	45%
Grade Effect	1.188	1.134	0.900	0.586	1.143
Gender Effect	0.160	0.125	0.083	-0.039	0.213
SES Effect	0.600	0.425	0.417	0.511	0.396

SUMMARY

In this chapter, we have seen how ConQuest can be used to fit multidimensional latent regression models. The fitting of multidimensional latent regression models brings together two sets of functionality that we have demonstrated in previous chapters: the facility to estimate latent regression models and the facility to fit multidimensional item response models.

REFERENCES

- Adams, R. J., Doig, B. A., and Rosier, M. 1991. *Science Learning in Victorian Schools*. Hawthorn, Victoria: Australian Council for Educational Research.
- Adams, R. J., Wilson, M. R., and Wu, M. L. 1997. Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioural Statistics*, 22(1), 46-75.
- Mislevy, R. J., and Sheehan, K. M. 1989. The role of collateral information about examinees in item parameter estimation. *Psychometrika*, 54, 661-79.
- Volodin, N. A., and Adams, R. J. 1995. Identifying and estimating a D-dimensional item response model. Paper presented at the International Objective Measurement Workshop, University of California. April, Berkeley, California.