Three-level analysis of health expenditure data

Contents

1 The data

The data set used here is the same as that used in Section 4.2 of the *Generalized Linear Modeling Guide*, and forms part of the data library of the Medical Expenditure Panel Survey (MEPS). Collected in 1999, these data from a longitudinal national survey were used to obtain regional and national estimates of health care use and expenditure based on the health expenditures of a sample of U.S. civilian non-institutionalized participants. The survey sample design utilized stratification, clustering, multiple stages of selection, and disproportionate sampling. The sample was drawn from 143 strata, divided into 460 PSUs. Information on 23,565 participants included positive person-level weights and forms the data set used here, excluding the 1,053 participants in the original data with zero person-level weights. Data for the first 10 participants on most of the variables used in this section are shown below in the form of a LISREL spreadsheet file, named **meps2.lsf**.

meps2.lsf

	VARSTR99	VARPSU99	PERWT99F	TOTEXP99	RACE	GENDER
2	1.00	1.00	4495.33	4.20	0.00	1.00
3	1.00	1.00	7689.24	4.98	0.00	0.00
4	1.00	2.00	11069.78	5.18	0.00	1.00
5	1.00	2.00	9288.98	8.11	0.00	0.00
6	1.00	2.00	12651.61	10.09	1.00	0.00
7	1.00	2.00	6156.33	6.64	1.00	1.00
8	1.00	2.00	31359.46	8.29	0.00	0.00
9	1.00	2.00	29082.10	7.85	0.00	1.00
10	1.00	2.00	14474.96	8.47	0.00	1.00

The variables of interest are:

- VARSTR99 is the stratum identification variable (143 strata in total).
- VARPSU99 is the PSU identification variable (460 PSUs in total).
- PERWT99F represents the final sample weight, with weights ranging between 307.16 and 80061.61, correcting for both non-response and adjustments to population control totals from the Current Population Survey.
- TOTEXP99 is the natural logarithm of the total health expenditure of a respondent in 1999, ranging between 0 and 12.24 and representing actual expenditure of between $0 and $206,721.
- RACE is an ethnicity indicator, with a value of 1 indicating white respondents, and 0 denoting all other ethnic groups as well as respondents for which ethnicity is not known. This variable was recoded from the original MEPS variable RACEX.
- GENDER is a gender indicator, with a value of 0 indicating a male participant and 1 a female participant; recoded from the original MEPS variable RGENDER.
- INSCOV is an indicator of the level of insurance coverage, where 0 indicates private coverage any time during 1999, and 1 indicates public coverage or no insurance at all during 1999.
- RPOVC991 to RPOVC995 are five indicator variables, each associated with a category of the original MEPS variable RPOVC99 which was constructed by dividing family income by the applicable poverty line (selection of which depended on family size and composition), expressed as a percentage.

Income is a variable that is often transformed using its natural log. Doing so in effect causes the impact of each additional dollar to decrease as income increases. Logarithmic transformation is also useful in lessening the influence of outliers, as the natural logarithm of a variable is much less sensitive to extreme observations than is the variable itself.

The original MEPS variable RPOVC99 assumed a value of 1 for a family with "high" income level where family income was equal to or greater than 400% of the applicable poverty line, and a value of 2 for those with a "low income" level (associated with 125% to 200% of the poverty line). Families with "middle income", "near poor" and "negative or poor" levels of income relative to poverty line income were coded 3, 4 and 5 respectively. For the "middle income" category, the ratio (as percentage) of family income to poverty line was 200% to less than 400%. In the case of "near poor" families, the percentages ranged between 100% and 125%, and for "negative or poor," the family income was less than 100% of the relevant poverty line. Thus, a value of 1 on the indicator variable RPOVC991 indicates a family with income at the "high " level, while a value of 1 on the variable RPOVC995 indicates a family with "negative or poor" income level. The variables RPOVC992, RPOVC993, and RPOVC994 are associated with the categories "low income", "middle income" and "near poor" respectively.

Note that as each of the five indicator variables for categories of RPOVC99 is coded 1 if a participant responded in that category and 0 otherwise, only four of the five indicator variables can be used in a model where an intercept is included. Indicator variables of this type can easily be created by using the **Create Dummies for** option on the **Select Response and Fixed variables** dialog box. Here, we opted to create them prior to analysis as illustration of that feature is not relevant to the example at hand.

2 The model

The multilevel model does not make provision for the specification of design related variables such as stratum or PSU. Instead, these design variables are used to define the hierarchical structure of the data. In this example, the stratum identification variable VARSTR99 is used as the level-3 identifier and the PSU identification variable VARPSU99 serves to identify level-2 units (*i.e.*, PSUs) nested within a given stratum. We thus use the design variables to define a three-level hierarchical structure, with participants as level-1 observations nested within PSUs, in turn nested within strata. While not explicitly acknowledging the survey design or offering a conventional design effect estimate to measure the difference in estimates obtained when implementing this design compared to estimates obtained under a simple random sample, a multilevel model offers the advantage of estimating the variation in total health care expenditure within and between PSUs.

A general three-level model for a response variable y depending on a set of r predictors $x_1, x_2, \ldots, x_r$ can be written in the form

$$y_{ijk} = \mathbf{x}'_{(f)ijk}\boldsymbol{\beta} + \mathbf{x}'_{(3)ijk}\mathbf{v}_i + \mathbf{x}'_{(2)ijk}\mathbf{u}_{ij} + \mathbf{x}'_{(1)ijk}\mathbf{e}_{ijk}$$

where $i = 1, 2, \ldots, N$ denotes the level-3 units, $j = 1, 2, \ldots, n_i$ the level-2 units, and $k = 1, 2, \ldots, n_{ij}$ the level-1 units. In this context, y_{ijk} represents the response of individual k, nested within level-2 unit j and level-3 unit i. The model shown here consists of a fixed and a random part. The fixed part of the model is represented by the vector product $\mathbf{x}'_{(f)ijk}\boldsymbol{\beta}$, where $\mathbf{x}'_{(f)ijk}$ is a typical row of the design matrix of the fixed part of the model with, as elements, a subset of the r predictors. The vector $\boldsymbol{\beta}$ contains the fixed, but unknown parameters to be estimated. The vector products $\mathbf{x}'_{(3)ijk}\mathbf{v}_i$, $\mathbf{x}'_{(2)ijk}\mathbf{u}_{ij}$, and $\mathbf{x}'_{(1)ijk}\mathbf{e}_{ijk}$ denote the random part of the model at levels 3, 2, and 1 respectively. For example, $\mathbf{x}'_{(3)ijk}$ represents a typical row of the design matrix of the random part at level-3, and $\mathbf{v}_i$ the vector of random level-3 coefficients to be estimated. The products $\mathbf{x}'_{(2)ijk}\mathbf{u}_{ij}$ and $\mathbf{x}'_{(1)ijk}\mathbf{e}_{ijk}$ serve the same purpose at levels 2 and 1 respectively. It is assumed that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N$ are independently and identically distributed (i.i.d.) with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Phi}_{(3)}$. Similarly, $\mathbf{u}_{i1}, \mathbf{u}_{i2}, \ldots, \mathbf{u}_{in_i}$ are assumed i.i.d., with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Phi}_{(2)}$, and $\mathbf{e}_{ij1}, \mathbf{e}_{ij2}, \ldots, \mathbf{e}_{ijn_{ij}}$ are assumed i.i.d., with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Phi}_{(1)}$.

Within this hierarchical framework, the model fitted to the data uses the participant's gender, ethnicity, type of health insurance cover, and measure of income relative to poverty level to predict the total expenditure on health care in 1999, the latter transformed to the natural logarithm of the actual expenses incurred.

$$\begin{aligned}\text{TOTEXP99}_{ijk} = \beta_0 + \beta_1 * \text{GENDER}_{ijk} + \beta_2 * \text{RACE}_{ijk} + \beta_3 * \text{INSCOV}_{ijk} + \\ \beta_4 * \text{RPOVC991}_{ijk} + \beta_5 * \text{RPOVC992}_{ijk} + \beta_6 * \text{RPOVC993}_{ijk} + \\ \beta_7 * \text{RPOVC994}_{ijk} + v_{i0} + u_{ij0} + e_{ijk}\end{aligned}$$

where β_0 denotes the average expected total expenditure on health care in 1999, and $\beta_1, \beta_2, \ldots, \beta_7$ indicate the estimated coefficients associated with the fixed part of the model which contains the predictor variables GENDER, RACE, INSCOV and the indicator variables for categories of income relative to the poverty level. The random part of the model is represented by v_{i0}, u_{ij0} and e_{ijk}, which denote the variation in average total health related expenditure over strata, between PSUs (or, in other words, over PSUs nested within strata) and between participants at the lowest level of the hierarchy.

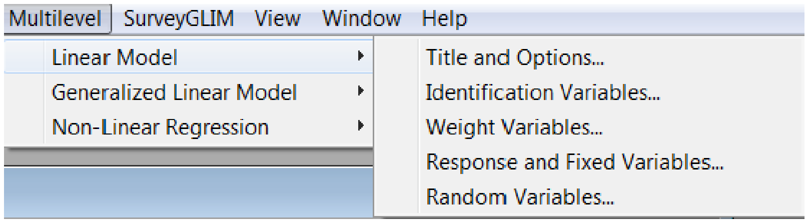
3 Multilevel analysis with sampling weights

The model is fitted to the data in **meps2.lsf** by using the sequence of four dialog boxes accessed via the **Multilevel**, **Linear Model** option from the main menu bar in LISREL. Note that options such as **Multilevel** and **SurveyGLIM** are only available on the main menu bar when a ***.lsf** file is open.

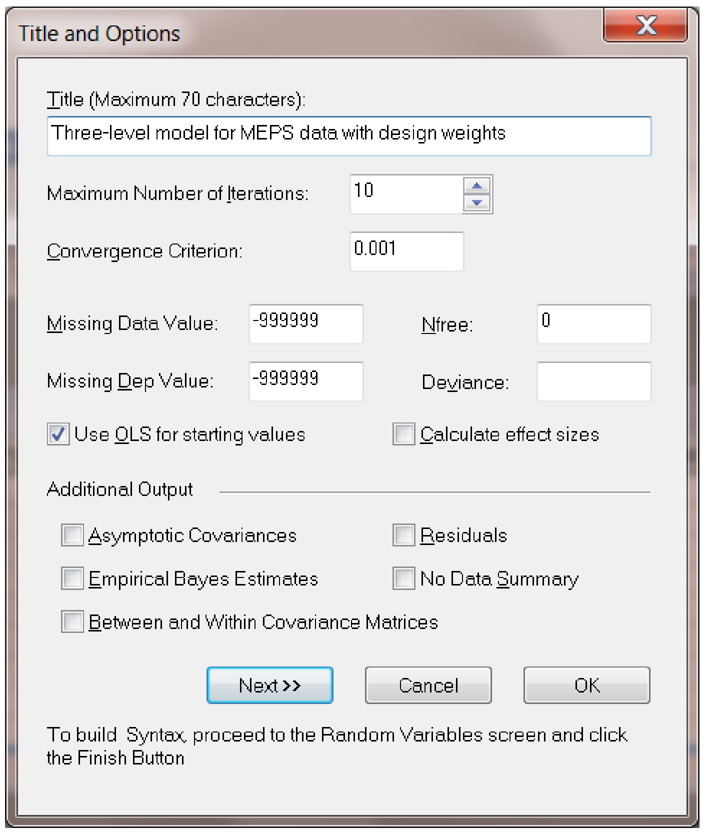
The first step is to open the LSF shown above, which is accomplished as follows:

Use the **File**, **Open** option to activate the display of an **Open** dialog box. Set the **Files of type** drop-down list box to **LISREL System Data (*.lsf)** and browse for the file **meps2.lsf** in the **Multilevel Examples** folder.

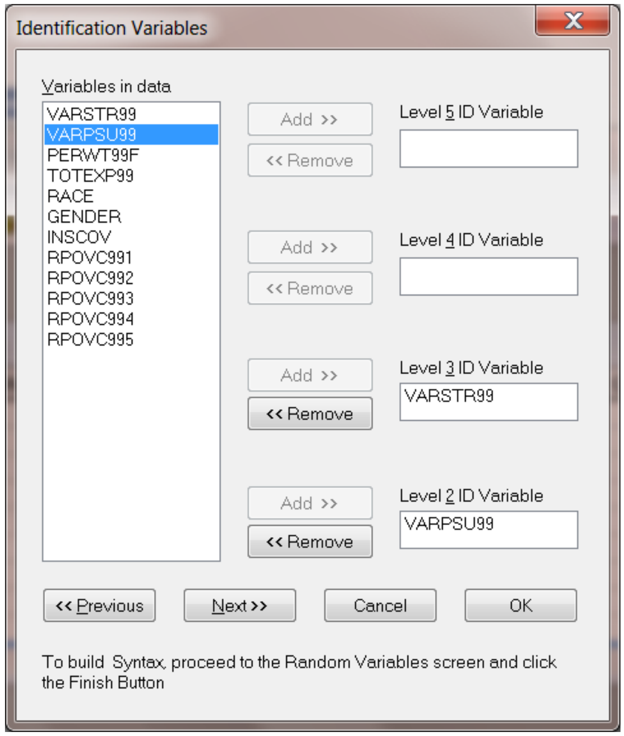
The next step is to describe the model to be fitted using the multilevel module in LISREL. From the main menu bar, select the **Multilevel** option. Here we limit our discussion to linear models, and thus the **Linear Model** option will be used throughout.

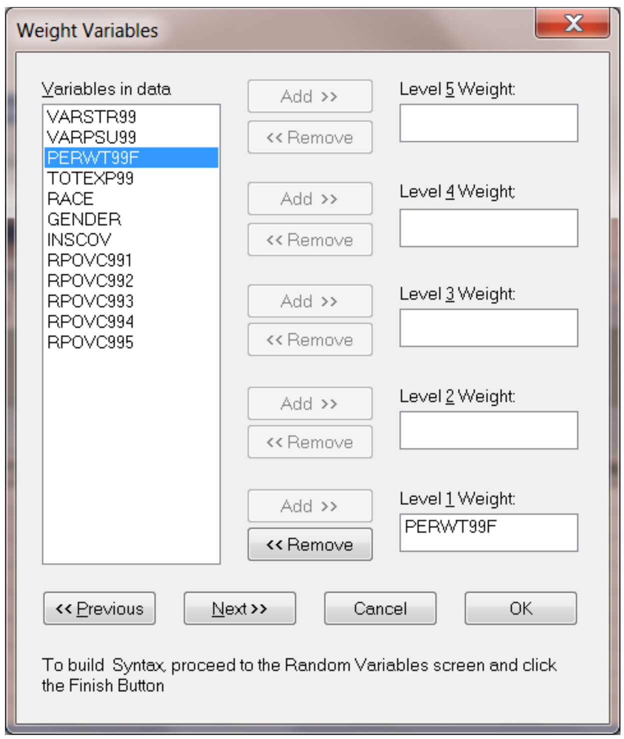


The first of the four options on the pop-up menu provide access to the **Title and Options** dialog. Start by providing a title for the analysis in the **Title** field. In this example, default settings for all other options associated with this dialog box are used. Click the **Next** button to go to the **Identification Variables** dialog box.



On the **Identification Variables** dialog box, enter the variables defining the hierarchical structure as ID variables. As mentioned before, the stratum identification variable is used to indicate the level-3 units in the hierarchical structure, and the PSU identification variable serves a similar purpose at level-2. Select the variables VARSTR99 and VARPSU99 as Level-3 ID variable and Level-2 ID variable respectively by clicking on the variable names in the **Variables in data** field at the left of the dialog box. Add them to the ID variable fields by clicking the appropriate **Add** button for each.



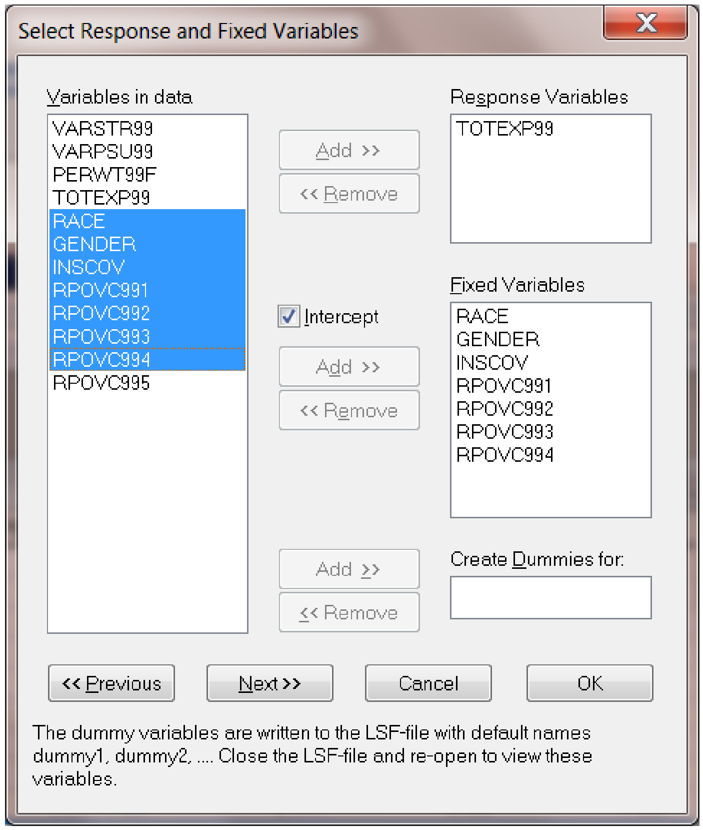


The **Weight Variables** dialog box is used to provide information on weight variables, if any. In our case, only one weight, denoted by the variable PERWT99F, is available. Select this variable from the **Variables in data** field and add it to the **Level-1 weight** field as shown below. As all available information is now entered on this dialog box, click the **Next** button to proceed to the **Select Response and Fixed Variables** dialog box.

The **Select Response and Fixed Variables** dialog box is used to identify the outcome variable and predictor variables, if any. Select and add the outcome variable TOTEXP99 to the **Response Variables** field in the same way as described for the previous dialog box. Next, select the variables starting from RACE to RPOVC994 by dragging the mouse over them and click the **Add** button next to the **Fixed Variables** field to include these variables as predictors in the model. This completes the specification of the response and fixed variables.

Before moving to the next dialog box, two other options available on this dialog box are worth noting.

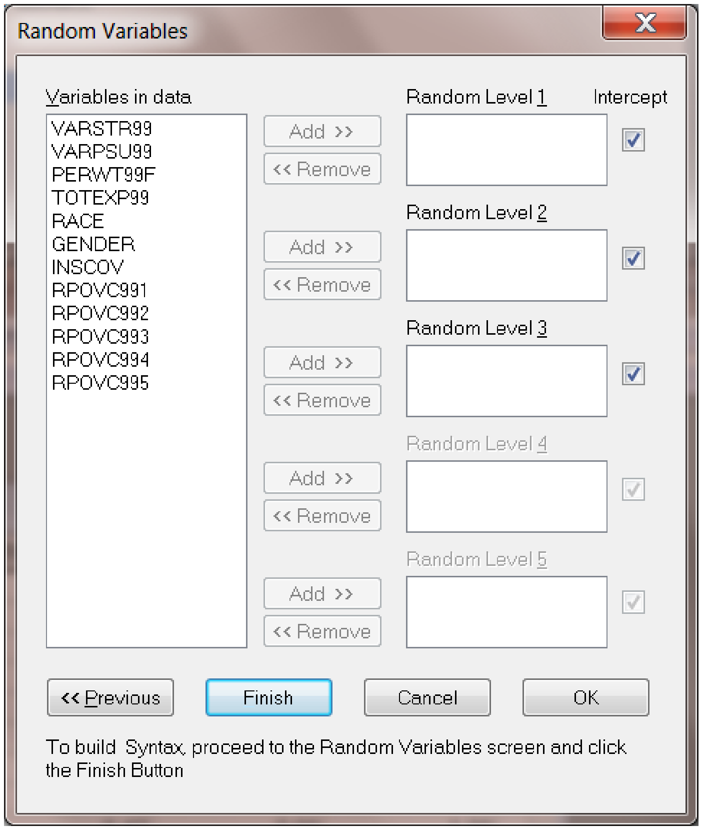
As previously discussed, the indicator variable associated with the highest level of income relative to the poverty line income is not selected for inclusion as the model fitted to the data has an intercept. Because of the intercept term, inclusion of all five indicator variables would lead to a design matrix of less than full rank and is bound to cause problems during the iterative procedure. An alternative approach would be to use all five indicator variables in a model without an intercept term. This can be achieved by deselecting the intercept term by unchecking the box next to Intercept.



The **Create Dummies for** option available on the **Select Response and Fixed** dialog box can be used to create indicator variables for the categories of a categorical variable such as RPOVC99. In fact, the indicator variables RPOVC991 to RPOVC995 were created in precisely this way for inclusion in the present analysis, and simply renamed from their default names of DUMMY1 to DUMMY5 using the **Define Variable** option from the **Data** menu accessed from the main menu bar in LISREL.

That said; proceed to the **Random Variables** dialog box by clicking the **Next** button.

The **Random Variables** dialog box shown below displays the default settings associated with this dialog box. In the current model, only the intercept coefficients are allowed to vary randomly at the various levels of the hierarchy. As this corresponds to the default settings shown on the dialog box, click the **Finish** button to generate the syntax for the model.



The syntax shown below corresponds to the information entered via the dialog boxes above. Run the model by clicking the **Run Prelis** icon on the main menu bar.

meps2.PRL

```
OPTIONS OLS=YES CONVERGE=0.001000 MAXITER=10 OUTPUT=STANDARD ;
TITLE=Three-level model for MEPS data with design weights;
SY='C:\LISREL9 Examples\MLEVELEX\meps2.lsf';
ID3=VARSTR99;
ID2=VARPSU99;
WEIGHT1=PERWT99F;
RESPONSE=TOTEXP99;
FIXED=intcept RACE GENDER INSCOV RPOVC991 RPOVC992 RPOVC993 RPOVC994;
RANDOM1=intcept;
RANDOM2=intcept;
RANDOM3=intcept;
```

Portions of the output file **meps2.out** are shown below.

```
meps2.OUT

                         +--------------+
                         | DATA SUMMARY |
                         +--------------+

 NUMBER OF LEVEL 3 UNITS :      143
 NUMBER OF LEVEL 2 UNITS :      460
 NUMBER OF LEVEL 1 UNITS :    23564

 ID3  :      1      2      3      4      5      6      7      8
 N2   :      2      2      2      2      2      3      2      2
 N1   :     29     85     55     86    159     48     48     78

 ID3  :      9     10     11     12     13     14     15     16
 N2   :      2      2      2     11      2      2      2      3
 N1   :    114     23     68    408     68     62    168     59

 ID3  :     17     18     19     20     21     22     23     24
 N2   :      2      5      2      2      2      2      2      2
 N1   :    364    215     69     44     40     59    171    151

 ID3  :     25     26     27     28     29     30     31     32
 N2   :      3      2      2      2      2      2     10      2
 N1   :    225     18     86     25     58     26    417     26
```

In the first section of the output file a description of the hierarchical structure is provided in the **Data Summary** section. A total of 143 strata, 460 PSUs and information from 23,564 individual participants were included at levels 3, 2 and 1 of the multilevel model. This corresponds to the survey design described earlier. In addition, a summary of the number of PSUs and participants nested within each stratum is provided. For stratum number 1 (ID3: 1), data are available from only 29 participants nested within 2 primary sampling units (N2: 2). By contrast, for stratum number 12 (ID3: 12), data are available from 408 participants (N1: 408) nested within 11 primary sampling units (N2: 11).

The output describing the estimated **fixed effects** after convergence is shown next. The estimates are shown in the column with heading BETA-HAT and correspond to the coefficients $\beta_0, \beta_2, \ldots, \beta_7$ in the model specification. From the z-values and associated exceedance probabilities, we see that the coefficients associated with gender, ethnicity and insurance coverage type were all highly significant. Recall that a value of 1 for the ethnicity indicator variable RACE indicated that a participant was white, with a value of 0 assigned to participants from all other ethnic groups. The positive estimated coefficient for this variable indicates an increase of 0.94298 units in the logarithm of total health expenditure, holding all other predictors constant. Similarly, female participants (coded "1" on the gender indicator GENDER), are expected to have a total health expenditure 0.91057 higher than male participants if all other variables are held constant. In contrast, participants with public coverage or no coverage have a lower expected total expenditure, as indicated by the negative estimated coefficient -0.65109.

```
meps2.OUT
 Three-level model for MEPS data with design weights

 ITERATION NUMBER     6

                          +-----------------------+
                          | FIXED PART OF MODEL   |
                          +-----------------------+

-----------------------------------------------------------------------------
  COEFFICIENTS              BETA-HAT      STD.ERR.      Z-VALUE      PR > |Z|
-----------------------------------------------------------------------------
  intcept                   4.39189       0.11481      38.25193      0.00000
  RACE                      0.94272       0.08693      10.84511      0.00000
  GENDER                    0.91059       0.03677      24.76122      0.00000
  INSCOV                   -0.65101       0.07624      -8.53904      0.00000
  RPOVC991                  0.35725       0.11520       3.10110      0.00193
  RPOVC992                 -0.13848       0.10523      -1.31593      0.18820
  RPOVC993                  0.07022       0.11782       0.59595      0.55121
  RPOVC994                 -0.32929       0.14183      -2.32166      0.02025

                          +-----------------------+
                          |   -2 LOG-LIKELIHOOD   |
                          +-----------------------+

  DEVIANCE= -2*LOG(LIKELIHOOD) =    119135.095120823
  NUMBER OF FREE PARAMETERS =               11

  CHI-SQUARE SCALE FACTOR =      0.25580
```

Turning to the indicator variables associated with income relative to the poverty line income, it can be seen that only two of the indicator variables, RPOVC991 and RPOVC994, have estimated coefficients that are significantly different from zero at a 5% level of significance. In the case of families with a "high" income, the estimate of 0.35750 for RPOVC991 indicates an expected increase in expenditure, while for "near poor" families, the estimate of -0.32939 indicates an expected decrease in expenditure, holding all other variables constant.

Estimated outcomes for different groups

To evaluate the expected effect of the measure of a family's income to the corresponding poverty line income, suppose that the variables RACE, GENDER, and INSCOV are held at zero, as would be the case for a nonwhite male participant with private insurance coverage. If such a participant originates from a family with "high" income, the logarithm of total health expenditure is expected to be

$$\begin{aligned} &\beta_0 + \beta_4(\text{RPOVC991}) + \beta_5(\text{RPOVC992}) + \beta_6(\text{RPOVC993}) + \beta_7(\text{RPOVC994}) \\ &= \beta_0 + \beta_4 \\ &= 4.39123 + 0.35750 \\ &= 4.74873 \end{aligned}$$

which translates to a projected total expenditure of $e^{4.74873} = \$115.437$. In contrast, for a participant with similar demographic background and coverage from a "near poor" family, we obtain a projected total expenditure of

$$\begin{aligned} &e^{\beta_0 + \beta_7} \\ &= e^{4.39123 - 0.32929} \\ &= \$58.086 \end{aligned}$$

The predicted total expenditure (as natural logarithm) for similar participants from "low", "middle" or "negative or poor" families are similarly obtained by calculating $e^{\beta_0+\beta_5}$, $e^{\beta_0+\beta_6}$ and e^{β_0} respectively.

Table 2.1: Predicted total health expenditure for various subgroups

Respondents with high family income (RPOVC991 = 1)	**Male** (GENDER = 0)		**Female** (GENDER = 1)	
	Insurance coverage:		**Insurance coverage:**	
	Private (INSCOV = 0)	**Public/none** (INSCOV = 1)	**Private** (INSCOV = 0)	**Public/none** (INSCOV = 1)
Nonwhite (RACE = 0)	$115	$60	$287	$150
White (RACE = 1)	$296	$155	$737	$384
Respondents with near poor income (RPOVC994 = 1)				
Nonwhite (RACE = 0)	$58	$30	$145	$75
White (RACE = 1)	$149	$78	$370	$193

In Table 2.1, the predicted total health expenditure is given for respondents with high or near poor family income, for each of the subpopulations formed by gender, ethnicity and insurance coverage. For purposes of the comparison, results are expressed in U.S. dollars, rather than in the natural logarithmic units of the outcome variable TOTEXP99. Respondents from families with high income consistently outspend their near poor counterparts by approximately 100%, regardless of gender, ethnicity or level of insurance coverage. In families with high income, female respondents spent more in 1999 than their male counterparts, regardless of ethnicity. This is generally also true for near poor respondents. It is also apparent that the total health expenditure in 1999 was higher for respondents with private insurance than for respondents with public or no coverage, and that white respondents spent more than respondents from other ethnic groups, regardless of gender or the level of family income. From exploratory analyses, we know that the outcome variable TOTEXP99 is highly skewed, with median 1999 expenditure of $ 377.41. When this is taken in account, we can conclude that, generally speaking, white females spent more on health in 1999 than 50% of all respondents in the sample.

The output for the **random part** of the model follows and is shown in the image above. There is significant variation in the average estimated total health expenditure at all levels, with the most variation over the participants (level-1), and the least variation over strata (level-3).

An estimate of the level-2 cluster effect, for example, is obtained as

$$\frac{0.17706}{0.07305+0.17706+7.00628}\times 100\% = 2.41\%$$

indicating that only 2.41% of the total variance explained is at level-2 of the model.

meps2.OUT

RANDOM PART OF MODEL

LEVEL 3	TAU-HAT	STD.ERR.	Z-VALUE	PR > |Z|
intcept /intcept	0.07303	0.04571	1.59773	0.11010

LEVEL 2	TAU-HAT	STD.ERR.	Z-VALUE	PR > |Z|
intcept /intcept	0.17695	0.05128	3.45046	0.00056

LEVEL 1	TAU-HAT	STD.ERR.	Z-VALUE	PR > |Z|
intcept /intcept	7.00626	0.22390	31.29193	0.00000

4 Multilevel analysis without sampling weights

To evaluate the effect on the estimated coefficients if the sampling weights are ignored for data known to come from a disproportionally sampled survey, we fit the same model without a WEIGHT command.

To fit the unweighted model, the syntax file from the previous analysis can be edited by simply deleting the WEIGHT1 command from the syntax file.

Alternatively, the **Level-1 Weight** field on the **Weight Variables** dialog box can be cleared by clicking on this field and then clicking the **Remove** button next to this field.

Clicking **Next** on this and the next two dialog boxes, followed by clicking the **Finish** button on the **Random Variables** dialog box will generate a revised syntax file.

After running the analysis by clicking the **Run Prelis** icon on the main menu bar, the following output is obtained for the fixed and random parts of the unweighted model.

meps2.OUT

ITERATION NUMBER 4

FIXED PART OF MODEL

COEFFICIENTS	BETA-HAT	STD.ERR.	Z-VALUE	PR > |Z|
intcept	4.45841	0.08350	53.39310	0.00000
RACE	0.68364	0.04971	13.75292	0.00000
GENDER	0.93063	0.03581	25.98628	0.00000
INSCOV	-0.61785	0.04571	-13.51696	0.00000
RPOVC991	0.49302	0.06489	7.59791	0.00000
RPOVC992	-0.15390	0.06677	-2.30502	0.02117
RPOVC993	0.10053	0.06193	1.62342	0.10450
RPOVC994	-0.34592	0.08978	-3.85321	0.00012

-2 LOG-LIKELIHOOD

DEVIANCE= -2*LOG(LIKELIHOOD) = 115546.006821173

NUMBER OF FREE PARAMETERS = 11

meps2.OUT

RANDOM PART OF MODEL

LEVEL 3	TAU-HAT	STD.ERR.	Z-VALUE	PR > |Z|
intcept /intcept	0.14587	0.03648	3.99824	0.00006

LEVEL 2	TAU-HAT	STD.ERR.	Z-VALUE	PR > |Z|
intcept /intcept	0.17442	0.02767	6.30265	0.00000

LEVEL 1	TAU-HAT	STD.ERR.	Z-VALUE	PR > |Z|
intcept /intcept	7.46282	0.06938	107.56310	0.00000

In Table 2.2, the predicted total health expenditure is given for respondents with high or near poor family income, for each of the subpopulations formed by gender, ethnicity and insurance coverage. When compared to Table 2.1, where similar results were given for the weighted analysis, no difference in the overall pattern of expenditure is detected. Note, however, that the predicted expenditure for Nonwhite respondents (RACE = 0) are consistently higher in Table 2.2 than was the case in Table 2.1. For white respondents, the unweighted results shown in Table 2.2 are consistently lower than the corresponding results in Table 2.1. If sample weights are not used in the analysis, it may lead to a consistent, although small, overestimation of the health expenditure of nonwhite respondents, and to an underestimation of the health expenditures of their white counterparts.

Table 2.2: Predicted total health expenditure for various subgroups

Respondents with high family income (RPOVC991 = 1)	**Male** (GENDER = 0)		**Female** (GENDER = 1)	
	Insurance coverage:		**Insurance coverage:**	
	Private (INSCOV = 0)	**Public/none** (INSCOV = 1)	**Private** (INSCOV = 0)	**Public/none** (INSCOV = 1)
Nonwhite (RACE = 0)	$141	$76	$359	$193
White (RACE = 1)	$280	$151	$710	$383
Respondents with near poor income (RPOVC994 = 1)				
Nonwhite (RACE = 0)	$61	$33	$154	$83
White (RACE = 1)	$120	$65	$304	$164

Results for the two models (weighted and unweighted) are summarized in Table 2.3. While results for the models fitted in this case are not dramatically different, we observe that while some coefficients are larger for the unweighted model (for example, the estimates for intcept, GENDER, INSCOV, and most markedly for RPOVC991), coefficients for RPOVC992 and RACE are larger for the weighted model. The largest difference observed is in the case of ethnicity, where an estimated increase of 0.94 in expenditure is associated with a white respondent under the weighted model, compared to only 0.68 for a white respondent in the unweighted model (holding all other variables constant). As this translates to a difference of $e^{0.26} = 1{,}296$ in total health expenditure for 1999, this difference is more important than it seems at first glance. In addition, the models are sufficiently different in that coefficients statistically significant in one model are no longer significant in the other, as illustrated by the estimated coefficients for the indicator variable RPOVC992. In the weighted model, low income respondents are not expected to have a significantly different expected total expenditure, while the estimated coefficient under the unweighted model indicates a statistically significant decrease of -0.15 units in the total expected expenditure.

Table 2.3: Results of weighted and unweighted level-3 models for the MEPS data

Coefficient	Estimate (weighted)	Estimate (unweighted)
intcept	4.39123	4.45841
RACE	0.94298	0.68364
GENDER	0.91057	0.93063
INSCOV	-0.65109	-0.61785
RPOVC991	0.35750	0.49302
RPOVC992	-0.13832*	-0.15390
RPOVC993	0.07036*	0.10053*
RPOVC994	-0.32929	-0.34592
Level-1 variance	7.00628	7.46282
Level-2 variance	0.17706	0.17442
Level-3 variance	0.07305	0.14587

* Not significant at a 5% level of significance.

5 Comparison with SurveyGLIM model

A similar model was fitted to the data using the SurveyGLIM module and a Normal-Identity model. Results are summarized in Table 2.4. In general, results obtained for the two models are similar.

Table 2.4: Results of weighted multilevel and SurveyGLIM models for the MEPS data

Coefficient	Multilevel model	SurveyGLIM model
intcept	4.39123	4.2771
RACE	0.94298	0.9393
GENDER	0.91057	0.9204
INSCOV	-0.65109	-0.6952
RPOVC991	0.35750	0.4319
RPOVC992	-0.13832*	-0.1415*
RPOVC993	0.07036*	0.1186*
RPOVC994	-0.32929	-0.3433

* Not significant at a 5% level of significance.

We conclude that, where weight variables are available for survey data, these should be included in the model as neglecting to do so can have a definite impact on the estimated coefficients. In the current example, results for the two models were not dramatically different, but comparison of predicted expenditure indicated the risk of consistently over- or underestimating the total health expenditure for groups with different levels of family poverty. From the results it seems reasonable to assume that it included a component to adjust for the over/undersampling of ethnic and gender groups, a procedure commonly used in survey design to ensure representativeness. This is in agreement with the fact that, according to the MEPS HC-054: 1999 report, Hispanic and black households were oversampled at rates of approximately 2 and 1.5 times the rate of remaining households.