Analysis of CPC survey data

Contents

1. Introduction

In this section, data from the March 1995 Current Population Survey are used. The data set is a subset of data obtained from the Data Library at the Department of Statistics at UCLA. A small number of demographic variables for two occupation groups was extracted, and all analyses are based on unweighted data.

Only respondents between the ages of twenty-one and sixty-five, who held full time positions in 1994 and had an annual income of US $1 or more were considered. The two groups we will focus on here are defined as follows:

Educational sector	Respondents with professional specialty in the educational sector
Construction sector	Operators, fabricators, and laborers in the construction sector

The variable GROUP in the PRELIS system file **INCOME.LSF** represents the groups, with GROUP = 0 for the respondents in the construction sector and GROUP = 1 for respondents in the educational sector.

Other demographic variables and their codes are:

GENDER	0 = female; 1 = male
AGE	Age in single years
MARITAL	1 = married; 0 = other
HOURS	Hours worked during last week at all jobs
CITIZEN	1 for native Americans, 0 for all foreign born respondents
INCOME	The natural logarithm of the personal income during 1994
DEGREE	1 for respondents with master's degrees, professional school degree, or doctoral degree; 0 otherwise

Respondents were from 9 regions of the USA, and the state of residence was also given. The variables REGION and STATE represent this information. The full description of the regions and states within regions is presented in Table 2.2. On the respondent level, the variable PERSON is a respondent identity number. The variable INCOME will be used as response variable in all analyses.

The models considered here are:

- A 3-level model for the combined group, using **INCOME.LSF**
- A similar model for the education sector only, using a subset of the data
- A similar model for the construction sector only, using a subset of the data

Table 2.2: Region and State codes

Region	State Code	State Name
New England region	11	Maine
(REGION = 1)	12	New Hampshire
	13	Vermont
	14	Massachusetts
	15	Rhode Island
	16	Connecticut
Middle Atlantic region	21	New York
(REGION=2)	22	New Jersey
	23	Pennsylvania
East North Central region	31	Ohio
(REGION=3)	32	Indiana
	33	Illinois
	34	Michigan
	35	Wisconsin
West North Central region	41	Minnesota
(REGION=4)	42	Iowa
	43	Missouri
	44	North Dakota
	45	South Dakota
	46	Nebraska
	47	Kansas
South Atlantic region	51	Delaware
(REGION=5)	52	Maryland
	53	District of Columbia
	54	Virginia
	55	West Virginia
	56	North Carolina
	57	South Carolina
	58	Georgia
	59	Florida
East South Central region	61	Kentucky
(REGION=6)	62	Tennessee

	63	Alabama
	64	Mississippi
West South Central region	71	Arkansas
(REGION=7)	72	Louisiana
	73	Oklahoma
	74	Texas
Mountain region	81	Montana
(REGION=8)	82	Idaho
	83	Wyoming
	84	Colorado
	85	New Mexico
	86	Arizona
	87	Utah
	88	Nevada
Pacific region	91	Washington
(REGION=9)	92	Oregon
	93	California
	94	Alaska
	95	Hawaii

1. 3-level model for subset of CPC survey data

The data set used, as described in the previous section, is contained in the LSF file **INCOME.LSF** and is in the **Multilevel Examples** folder. The variable labels and the first fifteen data records of this file are shown below.

INCOME.lsf

	region	state	age	gender	marital	hours	citizen	person	constant	degree	group	income
1	1.00	11.00	59.00	0.00	0.00	40.00	1.00	14742.00	1.00	0.00	1.00	10.11
2	1.00	11.00	56.00	1.00	1.00	40.00	1.00	14743.00	1.00	0.00	1.00	10.34
3	1.00	11.00	64.00	0.00	1.00	12.00	1.00	14750.00	1.00	0.00	1.00	10.43
4	1.00	11.00	30.00	1.00	1.00	40.00	1.00	14751.00	1.00	0.00	0.00	9.63
5	1.00	11.00	27.00	0.00	1.00	40.00	1.00	14752.00	1.00	0.00	1.00	9.98
6	1.00	11.00	49.00	0.00	1.00	40.00	1.00	14767.00	1.00	1.00	1.00	10.68
7	1.00	11.00	41.00	1.00	1.00	40.00	1.00	14768.00	1.00	0.00	0.00	10.60
8	1.00	11.00	36.00	0.00	1.00	40.00	1.00	14769.00	1.00	0.00	1.00	10.01
9	1.00	11.00	46.00	0.00	1.00	36.00	1.00	14781.00	1.00	0.00	0.00	10.19
10	1.00	11.00	39.00	0.00	1.00	55.00	1.00	14785.00	1.00	0.00	1.00	10.02
11	1.00	11.00	30.00	1.00	1.00	40.00	1.00	14813.00	1.00	0.00	0.00	10.96
12	1.00	11.00	46.00	1.00	1.00	10.00	0.00	14825.00	1.00	0.00	0.00	9.80
13	1.00	11.00	63.00	1.00	1.00	25.00	1.00	14830.00	1.00	0.00	0.00	10.21
14	1.00	11.00	38.00	1.00	0.00	75.00	1.00	14833.00	1.00	0.00	0.00	10.06
	1.00	11.00	38.00	0.00	1.00	40.00	1.00	14845.00	1.00	0.00	1.00	7.19

We start creating the input file by accepting the defaults for the maximum number of iterations, the convergence criterion, and the output options, then provide the title for the analysis (optional). As all respondents are nested within state of residence, and states are in turn nested within the nine regions, we select REGION as the variable for the level-3 identification variable field. The variables STATE and PERSON are selected as level-2 and level-1 identification variables, respectively.

Next, we select INCOME, representing the natural logarithm of personal income, as the response variable for this analysis. The variables AGE, GENDER, MARITAL, HOURS, CITIZEN, DEGREE, and GROUP are all entered into the model as fixed effects.

Finally, the intercept term is identified as a random effect on all levels of the hierarchy.

As a result, the input file (**INCOME1.PRL**) looks like this:

```
OPTIONS ;
TITLE=Analysis of CPC data: combined group;
SY=INCOME.LSF;
ID2=state;
ID3=region;
WEIGHT2=intcept;
RESPONSE=income;
FIXED=age gender marital hours citizen intcept degree group;
RANDOM1=intcept;
RANDOM2=intcept;
RANDOM3=intcept;
```

The data summary and output for the final iteration are given below.

```
ITERATION NUMBER     3

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

-------------------------------------------------------------------------------
  COEFFICIENTS             BETA-HAT      STD.ERR.      Z-VALUE       PR > |Z|
-------------------------------------------------------------------------------
  age                      0.01636       0.00115      14.25735       0.00000
  gender                   0.23710       0.01147      20.66457       0.00000
  marital                  0.08456       0.01145       7.38619       0.00000
  hours                    0.01344       0.00129      10.44586       0.00000
  citizen                  0.28652       0.06014       4.76384       0.00000
  intcept                  8.19488       0.07520     108.97096       0.00000
  degree                   0.41226       0.03654      11.28208       0.00000
  group                    0.19798       0.04229       4.68100       0.00000

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

  DEVIANCE= -2*LOG(LIKELIHOOD) =    14222.6159132370
  NUMBER OF FREE PARAMETERS =              11

  CHI-SQUARE SCALE FACTOR =      0.28969

                        +-----------------------+
                        | RANDOM PART OF MODEL  |
                        +-----------------------+

-------------------------------------------------------------------------------
  LEVEL 3                       TAU-HAT      STD.ERR.     Z-VALUE   PR > |Z|
-------------------------------------------------------------------------------
  intcept /intcept              0.00783      0.00264      2.96527   0.00302

-------------------------------------------------------------------------------
  LEVEL 2                       TAU-HAT      STD.ERR.     Z-VALUE   PR > |Z|
-------------------------------------------------------------------------------
  intcept /intcept              0.00522      0.00285      1.83055   0.06717
```

```
------------------------------------------------------------------------------
  LEVEL 1                        TAU-HAT      STD.ERR.     Z-VALUE   PR > |Z|
------------------------------------------------------------------------------
  intcept /intcept               0.60688      0.05422     11.19244   0.00000


        LEVEL 3 COVARIANCE MATRIX

              intcept

 intcept      0.00783



LEVEL 3 CORRELATION MATRIX

              intcept

 intcept       1.0000


        LEVEL 2 COVARIANCE MATRIX

              intcept

 intcept      0.00522


        LEVEL 2 CORRELATION MATRIX

              intcept

 intcept       1.0000


        LEVEL 1 COVARIANCE MATRIX

              intcept

 intcept       0.6069


        LEVEL 1 CORRELATION MATRIX

              intcept

 intcept       1.0000
```

From the output given above, we see that:

The nine regions had between 291 and 1095 respondents, nested within states. The smallest number of level-2 units within a level-3 unit was 3, for the middle Atlantic region which included only New York, New Jersey, and Pennsylvania.

All the fixed effects were highly significant. The coefficient for the mean income, was 8.19488. Since the response variable is the natural logarithm of a respondent’s annual income, this number translates to a mean income of

$$\exp(8.19488 + 21(0.01636) + 40(0.01344)) = \$8{,}743$$

for a respondent from the construction sector who is 21 years of age, working 40 hours per week, unmarried, without a higher degree, and not a USA citizen. Although the size of the coefficients is quite small, it should be kept in mind that the natural logarithm of income is used as response variable. The relatively large positive coefficients for GENDER (0.23710), CITIZEN (0.28652), and DEGREE (0.41226) indicate that males, citizens of the USA, and respondents with a high education level tend to earn more when other variables are held constant. A comparison of two respondents with different demographic profiles as given below illustrates this point.

Respondent 1	**Respondent 2**
AGE=30	AGE=30
HOURS=40	HOURS=40
GROUP=1	GROUP=1
MARITAL=0	MARITAL=0
GENDER=0	GENDER=1
CITIZEN=0	CITIZEN=1
DEGREE=0	DEGREE=1

The first respondent's expected income is calculated as

$$\text{Expected income} = \exp[8.19488 + 30(0.01636) + 40(0.01344) + 0.19798] = \exp[9.42126] = \$12{,}348$$

while the expected income of the second respondent is

$$\text{Expected income} = \exp[8.19488 + 30(0.01636) + 40(0.01344) + 0.19798 + 0.23710 + 0.28652 + 0.41226] = \exp[10.35714] = \$31{,}481$$

Income varies most over the respondents (level-1 units), and least over the nine regions (level-3) units as we can see from the variances at these levels, given as 0.60688 and 0.00783, respectively.

In order to take a closer look at the relationships within the construction and educational sectors, two separate data sets will be created for these groups and similar models fitted in the next two examples. In the next section, a model for respondents from the education sector will be considered.

2. Three-level model for the educational sector

In the previous example, a 3-level model for the combined education and construction sector respondents from the 1995 CPC survey data was considered. In order to study effects for the educational sector only, a subset of the data in the file **INCOME.LSF** is used.

We select respondents belonging to the educational sector by using the PRELIS SC (select cases) command and select only those cases for which GROUP = 1. The new dataset is saved as **EDUC.LSF**.

```
CREATE A SUBSET OF THE FULL DATASET
SY=INCOME.PSF
```

```
SC GROUP=1
OU XM RA=EDUC.PSF
```

The input file for the analysis is exactly the same as in the previous example, with one exception : the variable GROUP is not included as a fixed effect, as this variable now has the value 1 for all respondents in the data.

The input file for this analysis (**EDUC.PRL**) is shown below.

```
OPTIONS OLS=YES CONVERGE=0.001000 MAXITER=10 OUTPUT=STANDARD ;
TITLE=Analysis of CPC data: educational sector;
SY=EDUC.LSF;
ID1=person;
ID2=state;
ID3=region;
RESPONSE=income;
FIXED=age gender marital hours citizen intcept degree;
RANDOM1=constant;
RANDOM2=constant;
RANDOM3=constant;
```

Partial output for this analysis follows.

```
ITERATION NUMBER     4

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


-------------------------------------------------------------------------------
  COEFFICIENTS             BETA-HAT      STD.ERR.      Z-VALUE       PR > |Z|
-------------------------------------------------------------------------------
  age                       0.02001       0.00129      15.49123       0.00000
  gender                    0.20943       0.02759       7.58961       0.00000
  marital                  -0.01506       0.02851      -0.52812       0.59741
  hours                     0.01458       0.00079      18.55257       0.00000
  citizen                   0.17746       0.05042       3.51950       0.00043
  intcept                   8.38120       0.07850     106.76220       0.00000
  degree                    0.39622       0.02693      14.71524       0.00000


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

  DEVIANCE= -2*LOG(LIKELIHOOD) =    6991.19480162653
  NUMBER OF FREE PARAMETERS =              10


                       +-----------------------+
                       | RANDOM PART OF MODEL  |
                       +-----------------------+


-------------------------------------------------------------------------------
```

```
  LEVEL 3                       TAU-HAT      STD.ERR.     Z-VALUE   PR > |Z|
--------------------------------------------------------------------------------
  constant/constant             0.00502      0.00445      1.12840   0.25915

--------------------------------------------------------------------------------
  LEVEL 2                       TAU-HAT      STD.ERR.     Z-VALUE   PR > |Z|
--------------------------------------------------------------------------------
  constant/constant             0.01283      0.00498      2.57744   0.00995

--------------------------------------------------------------------------------
  LEVEL 1                       TAU-HAT      STD.ERR.     Z-VALUE   PR > |Z|
--------------------------------------------------------------------------------
  constant/constant             0.50458      0.01267     39.83364   0.00000


       LEVEL 3 COVARIANCE MATRIX

            constant

constant     0.00502


       LEVEL 3 CORRELATION MATRIX

            constant

constant      1.0000


       LEVEL 2 COVARIANCE MATRIX

            constant

constant     0.01283


       LEVEL 2 CORRELATION MATRIX

            constant

constant      1.0000


       LEVEL 1 COVARIANCE MATRIX

            constant

constant      0.5046


       LEVEL 1 CORRELATION MATRIX

            constant

constant      1.0000
```

For the education sector, the following results are obtained.

All fixed effects are highly significant and positive, with the exception of the coefficient for marital status (MARITAL). For this group, the coefficient for the intercept is 8.38120. From the results of the previous analysis, the intercept for the group of respondents with GROUP = 1 was 8.19488 + 0.19798 = 8.39286, with all other variables held constant. In general, the same trends are observed for the combined and educational sector only groups: larger positive coefficients are obtained for the variables GENDER, CITIZEN, and DEGREE. Using the same two respondent profiles as in the previous example, with the exception of the GROUP variable which was not included in this analysis, we calculate the expected incomes of the two respondents.

Respondent 1	**Respondent 2**
AGE=30	AGE=30
HOURS=40	HOURS=40
GROUP=1	GROUP=1
MARITAL=0	MARITAL=0
GENDER=0	GENDER=1
CITIZEN=0	CITIZEN=1
DEGREE=0	DEGREE=1

The first respondent's expected income is calculated as

$$\text{Expected income} = \exp[8.38120 + 30(0.02001) + 40(0.01458)] = \exp[9.5647] = \$14{,}252$$

while the expected income of the second respondent is

$$\text{Expected income} = \exp[8.38120 + 30(0.02001) + 40(0.01458) + 0.20943 + 0.17746 + 0.39622] =$$

$$\exp[10.34781] = \$31{,}118$$

The difference between the expected income of these respondents is slightly smaller when only the educational sector is considered.

For this sector, the mean income varies little over the nine regions. The variation at level 3 of the model is smaller than for the combined model (0.00783 versus 0.00502) and is not significant at any commonly used level of significance. The conclusion may be reached that most of the variation previously observed at a region level (level 3) was due to differences between the two sectors. Variation at levels 1 and 2 remained significant.

In the last example, we will consider a similar model for the construction sector only.

3. Three-level model for the construction sector

In the previous two examples, a model for the combined education and construction sectors and a model for the educational sector only were fitted to the 1995 CPC survey data. As a final example, we consider a separate model for those respondents active in the construction sector during 1994.

In order to study effects for the construction sector only, a subset of the data in the file **INCOME.LSF** is used. As before, we select respondents belonging to the construction sector by running a small PRELIS input file using the select cases (SC) command:

```
CREATE A SUBSET OF THE FULL DATASET
SY=INCOME.PSF
SC GROUP=0
OU XM RA=CONS.PSF
```

The resulting file **CONS.LSF** contains only those cases with GROUP = 0.

The input file is exactly the same as in the previous example, with one exception: the variable GROUP is not included as a fixed effect, as this variable now has the value 0 for all respondents in the data set **CONS.LSF**. The input file (**CONS.PRL**) for this analysis is shown below.

```
OPTIONS OLS=YES CONVERGE=0.001000 MAXITER=10 OUTPUT=STANDARD ;
TITLE=Analysis of CPC data: construction sector ;
SY=CONS.LSF;
ID2=state;
ID3=region;
RESPONSE=income;
FIXED=age gender marital hours citizen intcept degree;
RANDOM1=constant;
RANDOM2=constant;
RANDOM3=constant;
```

Partial output for this analysis is given below.

```
ITERATION NUMBER     6


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


-------------------------------------------------------------------------------
  COEFFICIENTS             BETA-HAT      STD.ERR.      Z-VALUE       PR > |Z|
-------------------------------------------------------------------------------
  age                      0.01208       0.00157       7.69715       0.00000
  gender                   0.39847       0.09723       4.09819       0.00004
  marital                  0.20337       0.03505       5.80180       0.00000
  hours                    0.01183       0.00110      10.73244       0.00000
  citizen                  0.32688       0.04805       6.80231       0.00000
  intcept                  8.15061       0.13001      62.69199       0.00000
  degree                   0.21725       0.22626       0.96018       0.33697


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

  DEVIANCE= -2*LOG(LIKELIHOOD) =    7087.74887649088
  NUMBER OF FREE PARAMETERS =              10
```

+------------------------+
| RANDOM PART OF MODEL |
+------------------------+

LEVEL 3	TAU-HAT	STD.ERR.	Z-VALUE	PR > |Z|
constant/constant	0.01203	0.00747	1.61013	0.10737

LEVEL 2	TAU-HAT	STD.ERR.	Z-VALUE	PR > |Z|
constant/constant	0.00486	0.00408	1.18984	0.23411

LEVEL 1	TAU-HAT	STD.ERR.	Z-VALUE	PR > |Z|
constant/constant	0.70303	0.01880	37.39738	0.00000

LEVEL 3 COVARIANCE MATRIX

	constant
constant	0.01203

LEVEL 3 CORRELATION MATRIX

	constant
constant	1.0000

LEVEL 2 COVARIANCE MATRIX

	constant
constant	0.00486

LEVEL 2 CORRELATION MATRIX

	constant
constant	1.0000

LEVEL 1 COVARIANCE MATRIX

	constant
constant	0.7030

```
        LEVEL 1 CORRELATION MATRIX

             constant

constant      1.0000
```

The following conclusions may be reached from the output given above:

- When the fixed effects for this model is compared to those obtained for the education sector, the coefficients for AGE and HOURS are smaller. The age of a respondent in the construction sector and the number of hours worked will result in a smaller expected increase in annual personal income. In contrast with the education sector, where the effect of marital status (MARITAL) was not significant, a respondent in the construction sector is likely to earn more when the respondent is married, with all other variables held constant.

- The coefficient for CITIZEN is approximately twice that of a respondent from the education sector (0.32688 versus 0.17746). The mean income, with all other variables held fixed at 0, is 8.15061 (as natural logarithm). With all other variables held constant, this translates into a $1,339 difference in baseline income between citizens and non-citizens in the construction sector. The baseline expected income for a US citizen working in the construction sector can be calculated as

  $$\text{Expected baseline} = \exp(8.15061 + 0.32688) = \$4{,}805$$

  For a US citizen in the education sector, the expected baseline income is calculated as

  $$\text{Expected baseline} = \exp(8.38120 + 0.17746) = \$5{,}211$$

- Again, the largest coefficients obtained are for GENDER, CITIZEN, and DEGREE. Where the coefficient for GENDER was 0.20943 in the education sector, the coefficient for the construction sector is approximately twice that, at 0.39847. From the output above, it is seen that the coefficient for DEGREE is not significant. A closer examination of the data, using PRELIS data screening features, reveals that only 14 respondents have a high level of education (masters, professional, or Ph.D degree). If the same model is fitted without the degree predictor, all other estimated parameter values basically remain unchanged.
- Turning to the random effects, we see that the only significant variation is over respondents. At both state and division level, the variation is not significant. From this, combined with the results of the analysis for the education sector, we conclude that the significant variation at level 2 seen in the combined model is probably due to the differences between these two groups.