

Supplement for
“A constrained regression model for an ordinal response with ordinal
predictors”

by
Javier Espinosa and Christian Hennig

In Section 7 of the paper, we present a real data application to quality of life self assessment in Chile. In order to assist reproducibility, we provide the criteria used in the data preprocessing stage to get the final data set from the raw data that is publicly available online at: <http://observatorio.ministeriodesarrollosocial.gob.cl/casen-multidimensional/casen/basedatos.php>.

1 Response variable and sample definition

The response variable is a self assessment of the quality of life (QoL), the name of this variable in the original data set is `r20` and its possible values are integers from 1 to 10, representing the possible answers: ‘1 Completely Unsatisfied’, ‘2’, ..., ‘9’, ‘10 Completely Satisfied’ correspondingly.

The sample is defined as those householders who live in the capital and reported the quality of life self assessment. In the original data set, householders are identified with the value 1 of the variable `pco`, whereas the capital corresponds to `region=13` and the valid responses of QoL lie between 1 and 10.

2 Predictors

2.1 Ordinal predictors

Educational Level: This variable takes into account the educational level, years of schooling, and whether the householder knows how to read and/or write. All these variables are treated in the following sequential steps:

1. Variable `educ` is grouped into four categories: values 0 and 99 in “Not educated”, values 1 and 2 in “Primary”, values from 3 to 6 in “Secondary”, and from 7 to 12 in “Higher”.

2. Those classified as “Secondary” are moved to “Primary” if their years of schooling are less than 9 (variable `ESC>9`).
3. Those classified as “Not educated” and with `educ=99` are moved to “Primary” if their years of schooling are more than 0 (variable `ESC>0`).
4. Those classified as “Primary” and with `educ=99` are moved to “Secondary” if their years of schooling are more than 8 (variable `ESC>8`).
5. Those classified as “Secondary” and with `educ=99` are moved to “Higher” if their years of schooling are more than 12 (variable `ESC>12`).
6. Those classified as “Secondary” or “Higher”, and declared that they do not know how to read and/or write (variable `e1` is 2, 3, or 4), are moved to “Primary”.
7. Those classified as “Not educated” and with `educ=99` (value for ‘do not know/do not answer’) are removed from the sample (28 cases, 0.37%).

Income Quintile: Raw variable `QAUTR.MN` is used.

Health Status: Variable `s16` is used. Values from 1 to 7 are considered only and those observations with value 99 (value for ‘do not know’) are removed from the sample (36 cases, 0.48%).

Overcrowding: Variable `hacinamiento` is used. Values from 1 to 4 are considered only and those observations with value 9 (value for ‘NA’) are removed from the sample (21 cases, 0.28%).

Children: This is a special case, we use the whole data set and a dummy variable to identify those people under 15 years old. Then grouping by the house identifier called `folio` we get the number of children by house. We incorporate this information back in the sample of householders living in the capital and reported the quality of life self assessment. Finally, we grouped the number of children when it was greater than or equal to 4.

2.2 Non-Ordinal predictors

None of the non-ordinal predictors was transformed. The name of variables “Age”, “Activity”, and “Gender” are `edad`, `activ`, and `sexo`, correspondingly.