0.1 chopit: Compound Hierarchical Ordered Probit for Survey Vignettes

The Compound Hierarchical Ordered Probit (CHOPIT) model corrects for "differential item functioning" or "interpersonal comparability" in ordinal survey responses. Given a self-assessment question (such as, "How healthy are you? Excellent, good, fair, or poor."), different respondents may interpret the response categories in different ways, such that excellent health to one individual may be fair health to a hypochondriac. For each ordinal self-assessment to be corrected, the CHOPIT model requires one or more vignette question (such as a description of a hypothetical person's health, followed by the same response categories as the self-assessment), and a set of associated explanatory variables for the respondent. The key assumption of the approach is that the thresholds (which determine how respondents translate their views into the response categories) have the same effect for different questions asked of the same respondent, but may differ across respondents; the model uses a parametric specification to predict the thresholds associated with an individual. The self-assessment and vignette questions may be taken from different surveys, so long as both surveys include the same explanatory variable questions to predict the thresholds. For ordinal data (without vignettes), see Section **??**, Section **??**, and Section **??**.

Syntax

```
> fml <- list(self = Y ~ X1 + X2,
              vign = cbind(Z1, Z2, Z3) ~ 1,
              tau  = ~ X1 + X2)
> z.out <- zelig(fml, data = list(self = data1, vign = data2),
                 model = "chopit")
> x.out <- setx(z.out)
> s.out <- sim(z.out, x = x.out, x1 = NULL)
```

Inputs

In this hierarchical model, the `formula` and `data` inputs to `zelig()` are lists with the following structure:

- The `formula` is a list with three `formula` objects corresponding to:
  - `self`: Specifies the self-response question (`Y`) as a function of a set of explanatory variables.
  - `vign`: Specifies the vignette questions on the left-hand side as a matrix in the form `cbind(Z1, Z2, Z3)`.
  - `tau`: Specifies explanatory variables that constrain the cut points across both the vignette and self-response questions. These explanatory variables do not necessarily need to overlap with the set of explanatory variables specified in the `self`

formula, but must be observed in both the vign and self data frames, described below.

- The data argument is a list of two data frames with
  - self: A data frame containing the self-response question(s) specified in the self formula and associated explanatory variables listed in the self and tau formulas.
  - vign: A data frame containing the vignette questions specified in the vign formula and associated explanatory variables listed in the tau formula.

Additional Inputs

In addition to the standard inputs, zelig() takes many additional options for compound hierarchical ordered probit regression, see help(chopit) and Wand et al. (2007, forthcoming) for details.

Examples

1. Basic Example

   Setting up the formula as a list for the self-response, vignettes, and the cut points (drawn from both the self-response and vignette data sets).

   ```
   > formula <- list(self = y ~ sex + age + educ + factor(country),
   +     vign = cbind(v1, v2, v3, v4, v5) ~ 1, tau = ~sex + age +
   +         educ + factor(country))
   ```

   Attaching the sample data sets. The free1 data correspond to the self-response data, and the free2 data correspond to the vignette subset. Note that the variables specified in the tau formula must be in both data sets.

   ```
   > data(free1, free2)
   ```

   ```
   > data <- list(self = free1, vign = free2)
   ```

   Estimating parameter values for the CHOPIT regression:

   ```
   > z.out <- zelig(formula, data = data, model = "chopit")
   ```

   Setting values for the explanatory variables to their default values:

   ```
   > x.out1 <- setx(z.out)
   ```

   Simulating quantities of interest from the sampling distribution.

   ```
   > s.out1 <- sim(z.out, x = x.out1)
   ```

```
> summary(s.out1)
```

2. Simulating First Differences

   Estimate the first difference in expected values between the average age (about 40 years old) and a 25 year old individual, with the other explanatory variables held at their default values:

   ```
   > x.out2 <- setx(z.out, age = 25)

   > s.out2 <- sim(z.out, x = x.out1, x1 = x.out2)

   > summary(s.out2)
   ```

3. Conditional prediction

   Conditional prediction generates expected values that are conditional on the observed self-response.

   ```
   > x.out3 <- setx(z.out, cond = TRUE)
   ```

   Since conditional prediction involves numeric integration, the procedure takes approximately one second per observation in `x.out3` on 64-bit R.

   ```
   > s.out3 <- sim(z.out, x = x.out3)

   > summary(s.out3)
   ```

Model

This model has two sets of response variables, one for the self-assessment and one for the vignettes. Let Y_i be the observed ordinal self-assessment for respondents $i = 1, \ldots, n$, and Z_{lj} be the ordinal vignette responses for individuals $l = 1, \ldots, L$ in the vignette subset for $j = 1, \ldots, J$ vignette questions, such that both $\{Y_i, Z_{lj}\}$ take integer values $k = 1, \ldots, K$ corresponding to the same ordinal assessment response categories.

- The *stochastic components* are described by unobserved continuous variables, Y_i^* and Z_{lj}^*, which follows normal distributions with mean μ_i and variance σ^2 in the case of Y_i^*, and mean θ_j and variance σ_j^2 in the case of each Z_{lj}^*. Using the default identification mechanism, the variance σ^2 for the self-assessment is fixed to 1. Thus,

  $$\begin{aligned} Y_i^* &\sim N(\mu_i, 1) \\ Z_{lj}^* &\sim N(\theta_j, \sigma_j^2) \end{aligned}$$

  such that each vignette response j has a scalar mean θ_j and variance σ_j^2 that does not vary over observations l. In cases where more than one self-response was administered to the same subject, an additional random effect may be included in the distribution of the latent Y_i^* in the form

$$Y_i^* \quad \sim \quad N(\mu_i, 1 + \omega^2)$$

where the variance term is obtained via the proof described in Appendix A of King et al. (2004).

The observation mechanisms that divide the continuous $\{Y_i^*, Z_{lj}^*\}$ into the discrete $\{Y_i, Z_{lj}\}$ are

$$\begin{aligned} Y_i &= k \quad \text{if} \quad \tau_i^{k-1} \leq Y_i^* \leq \tau_i^k \text{ for } k = 1, \ldots, K \\ Z_{lj} &= k \quad \text{if} \quad \tau_l^{k-1} \leq Z_{lj}^* \leq \tau_l^k \text{ for } k = 1, \ldots, K \end{aligned}$$

where the threshold parameters τ vary over individuals $\{i, l\}$, but are subject to the following constraints within each individual: $\tau^p < \tau^q$ for all $p < q$ and $\tau_0 = -\infty$ and $\tau_K = \infty$.

- There are three *systematic components* in the model.
  - For the self-assessment component, let

    $$\mu_i \quad = \quad x_i\beta$$

    where x_i is the vector of q explanatory variables for observation i, and β is the associated vector of coefficients.
  - In addition, the threshold parameters also vary over individuals in the self-assessment component as follows

    $$\begin{aligned} \tau_i^1 &= v_i\gamma^1 \\ \tau_i^k &= \tau_i^{k-1} + \exp(v_i\gamma^k) \text{ for } k = 2, \ldots, K \end{aligned}$$

    where v_i is the vector of p explanatory variables for observation i, and γ^k for $k = 1, \ldots, K$ are the vectors of coefficients associated with each categorical response. Thus, the threshold parameters vary over individuals since v_i vary, and over response categories since the γ^k vary over the threshold parameters.
  - Similarly, the threshold parameters vary over individuals in the vignette component as follows

    $$\begin{aligned} \tau_l^1 &= v_l\gamma^1 \\ \tau_l^k &= \tau_l^{k-1} + \exp(v_l\gamma^k) \text{ for } k = 2, \ldots, K \end{aligned}$$

    where v_l is a vector of p explanatory variables for observation l in the vignette subset, and γ^k are restricted to be the same γ^k used to parameterize the threshold parameters for the self-assessment component.

As King et al. (2004) note, the interpersonal comparability of responses (or response consistency) is achieved by constraining γ^k to be the same in both the self-assessment and vignette components of the model. Note that the variables included in v_i and v_l are the same, but the observed values of those variables differ across the vignette and self-response samples.

Quantities of Interest

- The expected value (`qi$ev`) for the CHOPIT model is the expected value of the posterior density for the systematic component μ_i,

  $$\text{EV} \;=\; E(\mu_i \mid x_i) \;=\; x_i\beta$$

  given draws of β from its sampling distribution.

- The first difference is the difference in the expected value of the posterior density for the systematic component μ_i given x_1 and x_0:

  $$\text{FD} = E(\mu_i \mid x_1) - E(\mu_i \mid x_0).$$

- In conditional prediction models, the conditional expected values (`qi$cev`) are the expected value of the distribution of μ_i conditional on the observed self-assessment response Y_i, where

  $$P(\mu_i \mid \tau_i, \beta, x_i, Y_i) \;=\; \prod_{k=1}^{K}[\Phi(\tau_i^k - \mu_i) - \Phi(\tau_i^{k-1} - \mu_i)] \times N(x_i\beta, x_i\widehat{V}(\widehat{\beta})x_i' + \widehat{\omega}^2)$$

  given the simulations of the threshold parameters calculated above, draws of β from its sampling distribution, and the estimated variance-covariance matrix for $\widehat{\beta}$.

Output Values

The output of each Zelig command contains useful information which you may view. For example, if you run `z.out <- zelig(..., model = "chopit")`, then you may examine the available information in `z.out` by using `names(z.out)`, see the estimated parameters by using `z.out$par`, and a default summary of information through `summary(z.out)`. Other elements available through the `$` operator are listed below.

- From the `zelig()` output object `z.out`, you may extract:
  - `par`: the maximum likelihood parameter estimates for $\widehat{\gamma}^k$ for $k = 1, \ldots, K$ response categories, $\log(\widehat{\omega})$ (if estimated), $\log(\widehat{\sigma})$ (if estimated), $\log(\widehat{\sigma_j})$ for $j = 1, \ldots, J$ vignette questions, $\widehat{\theta_j}$, and $\widehat{\beta}$.
  - `chopit.hessian`: the estimated Hessian matrix, with rows and columns corresponding to the elements in `par`.

- `value`: the value of the log-likelihood at its maximum
  - `counts`: the number of function and gradient calls to reach the maximum.
  - `formula`: the formula for `self`, `vign`, and `tau` selected by the user.
  - `call`: the call to `zelig()`.
  - `...`: additional outputs described in `help(chopit)`.

- Typing `summary(z.out)` provides a useful summary of the output from `zelig()`, but no items can be extracted.

- From the `sim()` output object `s.out`, you may extract quantities of interest arranged as matrices indexed by simulation × `x`-observation (for more than one `x`-observation). Available quantities are:

  - `qi$ev`: the simulated expected values for the specified values of `x`.
  - `qi$fd`: the simulated first difference in the expected values for the values specified in `x` and `x1`.
  - `qi$cev`: the simulated conditional expected value given `x`.

How to Cite

To cite the *chopit* Zelig model:

> Kosuke Imai, Gary King, and Oliva Lau. 2007. ”chopit: Compound Hierarchical Ordinal Probit Regression for Survey Vignettes” in Kosuke Imai, Gary King, and Olivia Lau, ”Zelig: Everyone’s Statistical Software,”`http://gking.harvard.edu/zelig`

To cite Zelig as a whole, please reference these two sources:

> Kosuke Imai, Gary King, and Olivia Lau. 2007. “Zelig: Everyone’s Statistical Software,” `http://GKing.harvard.edu/zelig`.

> Imai, Kosuke, Gary King, and Olivia Lau. (2008). “Toward A Common Framework for Statistical Analysis and Development.” Journal of Computational and Graphical Statistics, Vol. 17, No. 4 (December), pp. 892-913.

See also

The CHOPIT model is part of the anchors package by Jonathan Wand, Gary King, and Olivia Lau (Wand et al. 2007, forthcoming). Advanced users may wish to refer to `help(chopit)`, as well as King et al. (2004) and King and Wand (2007).

Bibliography

King, G., Murray, C. J., Salomon, J. A., and Tandon, A. (2004), "Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research," *American Political Science Review*, 98, 191–207, http://gking.harvard.edu/files/abs/vign-abs.shtml.

King, G. and Wand, J. (2007), "Comparing Incomparable Survey Responses: New Tools for Anchoring Vignettes," *Political Analysis*, 15, 46–66, http://gking.harvard.edu/files/abs/c-abs.shtml.

Wand, J., King, G., and Lau, O. (2007, forthcoming), "Anchors: Software for Anchoring Vignettes Data," *Journal of Statistical Software*.