

## 0.1 `probit.mixed`: Mixed effects probit Regression

Use generalized multi-level linear regression if you have covariates that are grouped according to one or more classification factors. The probit model is appropriate when the dependent variable is dichotomous.

While generally called multi-level models in the social sciences, this class of models is often referred to as mixed-effects models in the statistics literature and as hierarchical models in a Bayesian setting. This general class of models consists of linear models that are expressed as a function of both *fixed effects*, parameters corresponding to an entire population or certain repeatable levels of experimental factors, and *random effects*, parameters corresponding to individual experimental units drawn at random from a population.

### Syntax

```
z.out <- zelig(formula= y ~ x1 + x2 + tag(z1 + z2 | g),
               data=mydata, model="probit.mixed")

z.out <- zelig(formula= list(mu=y ~ x1 + x2 + tag(z1, gamma | g),
                           gamma= ~ tag(w1 + w2 | g)), data=mydata, model="probit.mixed")
```

### Inputs

`zelig()` takes the following arguments for mixed:

- **formula**: a two-sided linear formula object describing the systematic component of the model, with the response on the left of a `~` operator and the fixed effects terms, separated by `+` operators, on the right. Any random effects terms are included with the notation `tag(z1 + ... + zn | g)` with `z1 + ... + zn` specifying the model for the random effects and `g` the grouping structure. Random intercept terms are included with the notation `tag(1 | g)`.

Alternatively, **formula** may be a list where the first entry, **mu**, is a two-sided linear formula object describing the systematic component of the model, with the response on the left of a `~` operator and the fixed effects terms, separated by `+` operators, on the right. Any random effects terms are included with the notation `tag(z1, gamma | g)` with `z1` specifying the individual level model for the random effects, `g` the grouping structure and **gamma** references the second equation in the list. The **gamma** equation is one-sided linear formula object with the group level model for the random effects on the right side of a `~` operator. The model is specified with the notation `tag(w1 + ... + wn | g)` with `w1 + ... + wn` specifying the group level model and `g` the grouping structure.

### Additional Inputs

In addition, `zelig()` accepts the following additional arguments for model specification:

- **data**: An optional data frame containing the variables named in **formula**. By default, the variables are taken from the environment from which **zelig()** is called.
- **na.action**: A function that indicates what should happen when the data contain **NA**s. The default action (**na.fail**) causes **zelig()** to print an error message and terminate if there are any incomplete observations.

Additionally, users may wish to refer to **lmer** in the package **lme4** for more information, including control parameters for the estimation algorithm and their defaults.

## Examples

### 1. Basic Example with First Differences

Attach sample data:

```
> data(voteincome)
```

Estimate model:

```
> z.out1 <- zelig(vote ~ education + age + female + tag(1 | state),
+               data = voteincome, model = "probit.mixed")
```

Summarize regression coefficients and estimated variance of random effects:

```
> summary(z.out1)
```

Set explanatory variables to their default values, with high (80th percentile) and low (20th percentile) values for education:

```
> x.high <- setx(z.out1, education = quantile(voteincome$education,
+      0.8))
> x.low <- setx(z.out1, education = quantile(voteincome$education,
+      0.2))
```

Generate first differences for the effect of high versus low education on voting:

```
> s.out1 <- sim(z.out1, x = x.high, x1 = x.low)
> summary(s.out1)
```

## Mixed effects probit Regression Model

Let  $Y_{ij}$  be the binary dependent variable, realized for observation  $j$  in group  $i$  as  $y_{ij}$  which takes the value of either 0 or 1, for  $i = 1, \dots, M$ ,  $j = 1, \dots, n_i$ .

- The *stochastic component* is described by a Bernoulli distribution with mean vector  $\pi_{ij}$ .

$$Y_{ij} \sim \text{Bernoulli}(y_{ij}|\pi_{ij}) = \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}}$$

where

$$\pi_{ij} = \Pr(Y_{ij} = 1)$$

- The  $q$ -dimensional vector of *random effects*,  $b_i$ , is restricted to be mean zero, and therefore is completely characterized by the variance covariance matrix  $\Psi$ , a  $(q \times q)$  symmetric positive semi-definite matrix.

$$b_i \sim \text{Normal}(0, \Psi)$$

- The *systematic component* is

$$\pi_{ij} \equiv \Phi(X_{ij}\beta + Z_{ij}b_i)$$

where  $\Phi(\mu)$  is the cumulative distribution function of the Normal distribution with mean 0 and unit variance, and

where  $X_{ij}$  is the  $(n_i \times p \times M)$  array of known fixed effects explanatory variables,  $\beta$  is the  $p$ -dimensional vector of fixed effects coefficients,  $Z_{ij}$  is the  $(n_i \times q \times M)$  array of known random effects explanatory variables and  $b_i$  is the  $q$ -dimensional vector of random effects.

## Quantities of Interest

- The predicted values (`qi$pr`) are draws from the Binomial distribution with mean equal to the simulated expected value,  $\pi_{ij}$  for

$$\pi_{ij} = \Phi(X_{ij}\beta + Z_{ij}b_i)$$

given  $X_{ij}$  and  $Z_{ij}$  and simulations of  $\beta$  and  $b_i$  from their posterior distributions. The estimated variance covariance matrices are taken as correct and are themselves not simulated.

- The expected values (`qi$ev`) are simulations of the predicted probability of a success given draws of  $\beta$  from its posterior:

$$E(Y_{ij}|X_{ij}) = \pi_{ij} = \Phi(X_{ij}\beta).$$

- The first difference (`qi$fd`) is given by the difference in predicted probabilities, conditional on  $X_{ij}$  and  $X'_{ij}$ , representing different values of the explanatory variables.

$$FD(Y_{ij}|X_{ij}, X'_{ij}) = Pr(Y_{ij} = 1|X_{ij}) - Pr(Y_{ij} = 1|X'_{ij})$$

- The risk ratio (`qi$rr`) is defined as

$$RR(Y_{ij}|X_{ij}, X'_{ij}) = \frac{Pr(Y_{ij} = 1|X_{ij})}{Pr(Y_{ij} = 1|X'_{ij})}$$

- In conditional prediction models, the average predicted treatment effect (`qi$att.pr`) for the treatment group is given by

$$\frac{1}{\sum_{i=1}^M \sum_{j=1}^{n_i} t_{ij}} \sum_{i=1}^M \sum_{j:t_{ij}=1}^{n_i} \{Y_{ij}(t_{ij} = 1) - Y_{ij}(\widehat{t_{ij} = 0})\},$$

where  $t_{ij}$  is a binary explanatory variable defining the treatment ( $t_{ij} = 1$ ) and control ( $t_{ij} = 0$ ) groups. Variation in the simulations is due to uncertainty in simulating  $Y_{ij}(t_{ij} = 0)$ , the counterfactual predicted value of  $Y_{ij}$  for observations in the treatment group, under the assumption that everything stays the same except that the treatment indicator is switched to  $t_{ij} = 0$ .

- In conditional prediction models, the average expected treatment effect (`qi$att.ev`) for the treatment group is given by

$$\frac{1}{\sum_{i=1}^M \sum_{j=1}^{n_i} t_{ij}} \sum_{i=1}^M \sum_{j:t_{ij}=1}^{n_i} \{Y_{ij}(t_{ij} = 1) - E[Y_{ij}(t_{ij} = 0)]\},$$

where  $t_{ij}$  is a binary explanatory variable defining the treatment ( $t_{ij} = 1$ ) and control ( $t_{ij} = 0$ ) groups. Variation in the simulations is due to uncertainty in simulating  $E[Y_{ij}(t_{ij} = 0)]$ , the counterfactual expected value of  $Y_{ij}$  for observations in the treatment group, under the assumption that everything stays the same except that the treatment indicator is switched to  $t_{ij} = 0$ .

## Output Values

The output of each Zelig command contains useful information which you may view. You may examine the available information in `z.out` by using `slotNames(z.out)`, see the fixed effect coefficients by using `summary(z.out)$coefs`, and a default summary of information through `summary(z.out)`. Other elements available through the operator are listed below.

- From the `zelig()` output stored in `summary(z.out)`, you may extract:
  - `fixef`: numeric vector containing the conditional estimates of the fixed effects.

- `ranef`: numeric vector containing the conditional modes of the random effects.
- `frame`: the model frame for the model.
- From the `sim()` output stored in `s.out`, you may extract quantities of interest stored in a data frame:
  - `qi$pr`: the simulated predicted values drawn from the distributions defined by the expected values.
  - `qi$ev`: the simulated expected values for the specified values of `x`.
  - `qi$fd`: the simulated first differences in the expected values for the values specified in `x` and `x1`.
  - `qi$ate.pr`: the simulated average predicted treatment effect for the treated from conditional prediction models.
  - `qi$ate.ev`: the simulated average expected treatment effect for the treated from conditional prediction models.

## How to Cite

To cite the *probit.mixed* Zelig model:

Delia Bailey and Ferdinand Alimadhi. 2007. "probit.mixed: Mixed effects probit model" 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

Mixed effects probit regression is part of `lme4` package by Douglas M. Bates (Bates 2007). For a detailed discussion of mixed-effects models, please see ?

# Bibliography

Bates, D. (2007), *lme4: Fit linear and generalized linear mixed-effects models*.