0.1 `relogit`: Rare Events Logistic Regression for Dichotomous Dependent Variables

The `relogit` procedure estimates the same model as standard logistic regression (appropriate when you have a dichotomous dependent variable and a set of explanatory variables; see Section ??), but the estimates are corrected for the bias that occurs when the sample is small or the observed events are rare (i.e., if the dependent variable has many more 1s than 0s or the reverse). The `relogit` procedure also optionally uses prior correction for case-control sampling designs.

Syntax

```
> z.out <- zelig(Y ~ X1 + X2, model = "relogit", tau = NULL,
                 case.correct = c("prior", "weighting"),
                 bias.correct = TRUE, robust = FALSE,
                 data = mydata, ...)
> x.out <- setx(z.out)
> s.out <- sim(z.out, x = x.out)
```

Arguments

The `relogit` procedure supports four optional arguments in addition to the standard arguments for `zelig()`. You may additionally use:

- `tau`: a vector containing either one or two values for τ, the true population fraction of ones. Use, for example, `tau = c(0.05, 0.1)` to specify that the lower bound on `tau` is 0.05 and the upper bound is 0.1. If left unspecified, only finite-sample bias correction is performed, not case-control correction.

- `case.correct`: if `tau` is specified, choose a method to correct for case-control sampling design: `"prior"` (default) or `"weighting"`.

- `bias.correct`: a logical value of `TRUE` (default) or `FALSE` indicating whether the intercept should be corrected for finite sample (rare events) bias.

- `robust`: defaults to `FALSE` (except when `case.control = "weighting"`; the default in this case becomes `robust = TRUE`). If `TRUE` is selected, `zelig()` computes robust standard errors via the `sandwich` package (see Zeileis (2004)). The default type of robust standard error is heteroskedastic and autocorrelation consistent (HAC), and assumes that observations are ordered by time index.

  In addition, `robust` may be a list with the following options:

  - `method`: Choose from
    - `"vcovHAC"`: (default if `robust = TRUE`) HAC standard errors.

        * "kernHAC": HAC standard errors using the weights given in Andrews (1991).
        * "weave": HAC standard errors using the weights given in Lumley and Heagerty (1999).
    - order.by: defaults to NULL (the observations are chronologically ordered as in the original data). Optionally, you may specify a vector of weights (either as order.by = z, where z exists outside the data frame; or as order.by = ~z, where z is a variable in the data frame) The observations are chronologically ordered by the size of z.
    - ...: additional options passed to the functions specified in method. See the sandwich library and Zeileis (2004) for more options.

Note that if tau = NULL, bias.correct = FALSE, robust = FALSE, the relogit procedure performs a standard logistic regression without any correction.

Example 1: One Tau with Prior Correction and Bias Correction

Due to memory and space considerations, the data used here are a sample drawn from the full data set used in King and Zeng, 2001, The proportion of militarized interstate conflicts to the absence of disputes is $\tau = 1,042/303,772 \approx 0.00343$. To estimate the model,

```
> data(mid)

> z.out1 <- zelig(conflict ~ major + contig + power + maxdem +
+     mindem + years, data = mid, model = "relogit", tau = 1042/303772)
```

Summarize the model output:

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

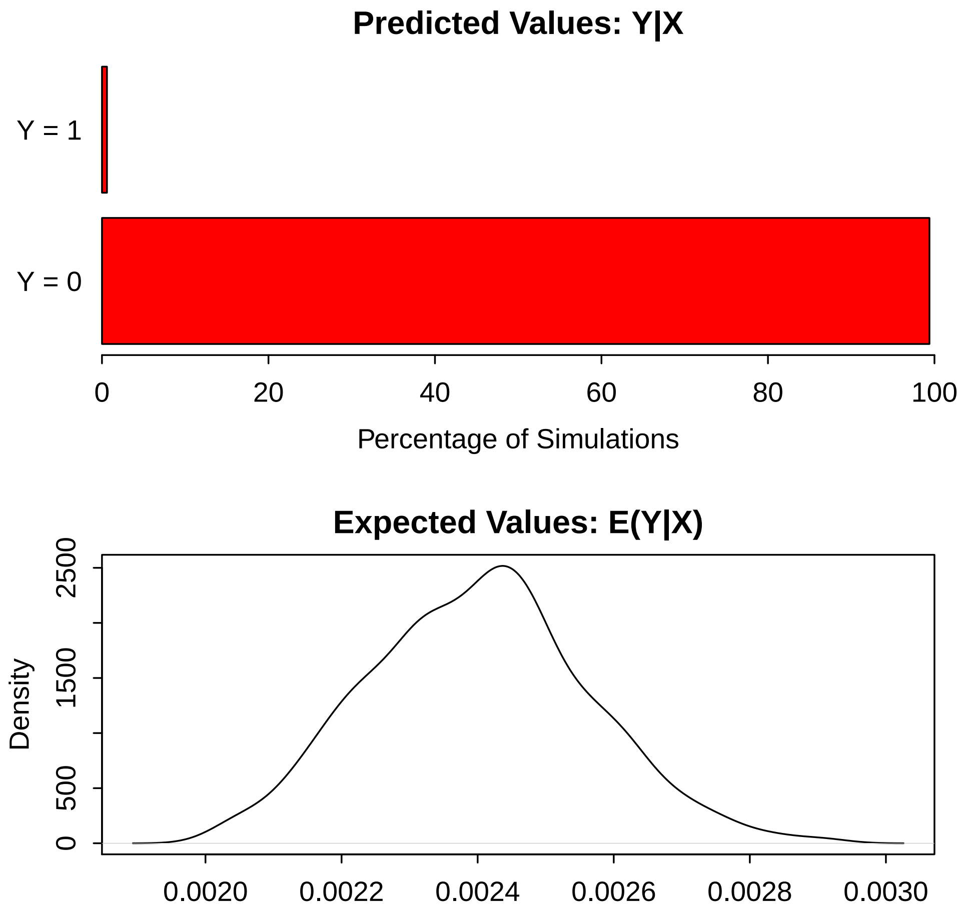
Set the explanatory variables to their means:

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

Simulate quantities of interest:

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

> plot(s.out1)
```



Example 2: One Tau with Weighting, Robust Standard Errors, and Bias Correction

Suppose that we wish to perform case control correction using weighting (rather than the default prior correction). To estimate the model:

```
> z.out2 <- zelig(conflict ~ major + contig + power + maxdem +
+     mindem + years, data = mid, model = "relogit", tau = 1042/303772,
+     case.control = "weighting", robust = TRUE)
```

Summarize the model output:

```
> summary(z.out2)
```

Set the explanatory variables to their means:

```
> x.out2 <- setx(z.out2)
```

Simulate quantities of interest:

```
> s.out2 <- sim(z.out2, x = x.out2)
> summary(s.out2)
```

Example 3: Two Taus with Bias Correction and Prior Correction

Suppose that we did not know that $\tau \approx 0.00343$, but only that it was somewhere between $(0.002, 0.005)$. To estimate a model with a range of feasible estimates for τ (using the default prior correction method for case control correction):

```
> z.out2 <- zelig(conflict ~ major + contig + power + maxdem +
+     mindem + years, data = mid, model = "relogit", tau = c(0.002,
+     0.005))
```

Summarize the model output:

```
> summary(z.out2)
```

Set the explanatory variables to their means:

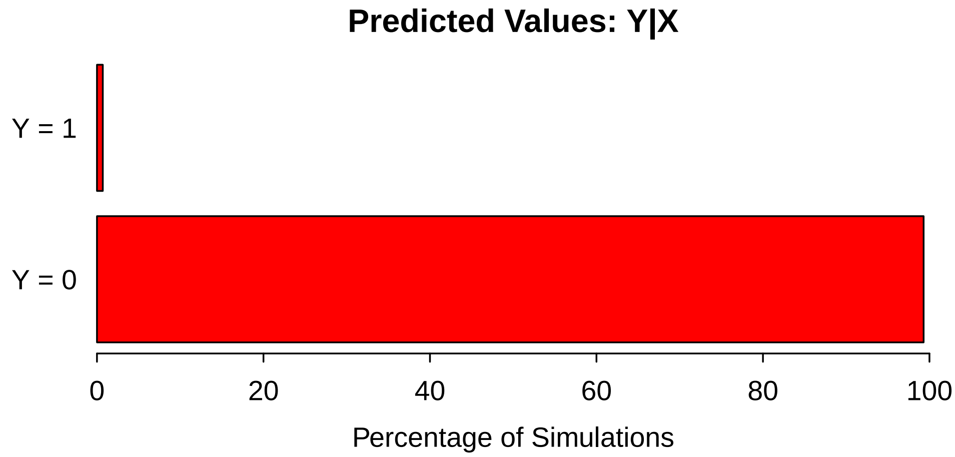
```
> x.out2 <- setx(z.out2)
```

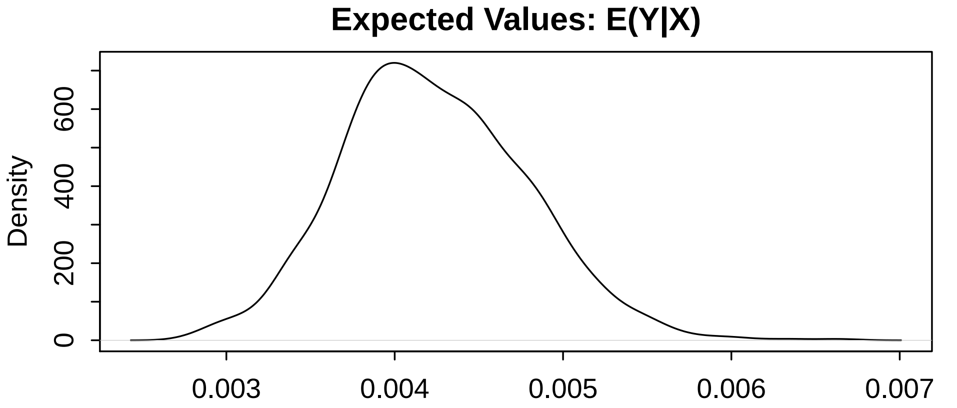
Simulate quantities of interest:

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

> summary(s.out2)

> plot(s.out2)
```





The cost of giving a range of values for τ is that point estimates are not available for quantities of interest. Instead, quantities are presented as confidence intervals with significance less than or equal to a specified level (e.g., at least 95% of the simulations are contained in the nominal 95% confidence interval).

Model

- Like the standard logistic regression, the *stochastic component* for the rare events logistic regression is:

  $$Y_i \; \sim \; \text{Bernoulli}(\pi_i),$$

  where Y_i is the binary dependent variable, and takes a value of either 0 or 1.

- The *systematic component* is:

  $$\pi_i \; = \; \frac{1}{1 + \exp(-x_i \beta)}.$$

- If the sample is generated via a case-control (or choice-based) design, such as when drawing all events (or "cases") and a sample from the non-events (or "controls") and going backwards to collect the explanatory variables, you must correct for selecting on the dependent variable. While the slope coefficients are approximately unbiased, the constant term may be significantly biased. Zelig has two methods for case control correction:

  1. The "prior correction" method adjusts the intercept term. Let τ be the true population fraction of events, $\bar{y}$ the fraction of events in the sample, and $\hat{\beta}_0$ the uncorrected intercept term. The corrected intercept β_0 is:

  $$\beta = \hat{\beta}_0 - \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right].$$

  2. The "weighting" method performs a weighted logistic regression to correct for a case-control sampling design. Let the 1 subscript denote observations for which the dependent variable is observed as a 1, and the 0 subscript denote observations for which the dependent variable is observed as a 0. Then the vector of weights w_i

  $$\begin{aligned} w_1 &= \frac{\tau}{\bar{y}} \\ w_0 &= \frac{(1-\tau)}{(1-\bar{y})} \\ w_i &= w_1 Y_i + w_0(1 - Y_i) \end{aligned}$$

  If τ is unknown, you may alternatively specify an upper and lower bound for the possible range of τ. In this case, the `relogit` procedure uses "robust Bayesian" methods to generate a confidence interval (rather than a point estimate) for each quantity of interest. The nominal coverage of the confidence interval is at least as great as the actual coverage.

- By default, estimates of the the coefficients β are bias-corrected to account for finite sample or rare events bias. In addition, quantities of interest, such as predicted probabilities, are also corrected of rare-events bias. If $\widehat{\beta}$ are the uncorrected logit coefficients and bias($\widehat{\beta}$) is the bias term, the corrected coefficients $\tilde{\beta}$ are

  $$\widehat{\beta} - \text{bias}(\widehat{\beta}) = \tilde{\beta}$$

  The bias term is

  $$\text{bias}(\widehat{\beta}) = (X'WX)^{-1}X'W\xi$$

  where

  $$\begin{aligned} \xi_i &= 0.5Q_{ii}\Big((1 + w - 1)\widehat{\pi}_i - w_1\Big) \\ Q &= X(X'WX)^{-1}X' \end{aligned}$$

  $$W = \text{diag}\{\widehat{\pi}_i(1 - \widehat{\pi}_i)w_i\}$$

where w_i and w_1 are given in the “weighting” section above.

Quantities of Interest

- For either one or no τ:
  - The expected values (qi$ev) for the rare events logit are simulations of the predicted probability
    $$E(Y) = \pi_i = \frac{1}{1+\exp(-x_i\beta)},$$
    given draws of β from its posterior.
  - The predicted value (qi$pr) is a draw from a binomial distribution with mean equal to the simulated π_i.
  - The first difference (qi$fd) is defined as
    $$\mathrm{FD} = \Pr(Y = 1 \mid x_1, \tau) - \Pr(Y = 1 \mid x, \tau).$$
  - The risk ratio (qi$rr) is defined as
    $$\mathrm{RR} = \Pr(Y = 1 \mid x_1, \tau) \ / \ \Pr(Y = 1 \mid x, \tau).$$

- For a range of τ defined by $[\tau_1, \tau_2]$, each of the quantities of interest are $n \times 2$ matrices, which report the lower and upper bounds, respectively, for a confidence interval with nominal coverage at least as great as the actual coverage. At worst, these bounds are conservative estimates for the likely range for each quantity of interest. Please refer to King and Zeng (2002) for the specific method of calculating bounded quantities of interest.

- In conditional prediction models, the average expected treatment effect (att.ev) for the treatment group is
  $$\frac{1}{\sum_{i=1}^n t_i} \sum_{i:t_i=1}^n \left\{ Y_i(t_i = 1) - E[Y_i(t_i = 0)] \right\},$$
  where t_i is a binary explanatory variable defining the treatment ($t_i = 1$) and control ($t_i = 0$) groups. Variation in the simulations are due to uncertainty in simulating $E[Y_i(t_i = 0)]$, the counterfactual expected value of Y_i for observations in the treatment group, under the assumption that everything stays the same except that the treatment indicator is switched to $t_i = 0$.

- In conditional prediction models, the average predicted treatment effect (att.pr) for the treatment group is
  $$\frac{1}{\sum_{i=1}^n t_i} \sum_{i:t_i=1}^n \left\{ Y_i(t_i = 1) - \widehat{Y_i(t_i = 0)} \right\},$$

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

Output Values

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

- From the `zelig()` output object `z.out`, you may extract:
  - `coefficients`: parameter estimates for the explanatory variables.
  - `bias.correct`: `TRUE` if bias correction was selected, else `FALSE`.
  - `prior.correct`: `TRUE` if prior correction was selected, else `FALSE`.
  - `weighting`: `TRUE` if weighting was selected, else `FALSE`.
  - `tau`: the value of `tau` for which case control correction was implemented.
  - `residuals`: the working residuals in the final iteration of the IWLS fit.
  - `fitted.values`: the vector of fitted values for the systemic component, π_i.
  - `linear.predictors`: the vector of $x_i\beta$
  - `aic`: Akaike's Information Criterion (minus twice the maximized log-likelihood plus twice the number of coefficients).
  - `df.residual`: the residual degrees of freedom.
  - `df.null`: the residual degrees of freedom for the null model.
  - `zelig.data`: the input data frame if `save.data = TRUE`.

  Note that for a range of τ, each of the above items may be extracted from the `"lower.estimate"` and `"upper.estimate"` objects in your `zelig` output. Use `lower <- z.out$lower.estimate`, and then `lower$coefficients` to extract the coefficients for the empirical estimate generated for the smaller of the two τ.

- From `summary(z.out)`, you may extract:
  - `coefficients`: the parameter estimates with their associated standard errors, p-values, and t-statistics.
  - `cov.scaled`: a $k \times k$ matrix of scaled covariances.

  - `cov.unscaled`: a $k \times k$ matrix of unscaled covariances.

- 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$ev`: the simulated expected values, or predicted probabilities, for the specified values of `x`.
  - `qi$pr`: the simulated predicted values drawn from Binomial distributions given the predicted probabilities.
  - `qi$fd`: the simulated first difference in the predicted probabilities for the values specified in `x` and `x1`.
  - `qi$rr`: the simulated risk ratio for the predicted probabilities simulated from `x` and `x1`.
  - `qi$att.ev`: the simulated average expected treatment effect for the treated from conditional prediction models.
  - `qi$att.pr`: the simulated average predicted treatment effect for the treated from conditional prediction models.

Differences with Stata Version

The Stata version of ReLogit and the R implementation differ slightly in their coefficient estimates due to differences in the matrix inversion routines implemented in R and Stata. Zelig uses orthogonal-triangular decomposition (through `lm.influence()`) to compute the bias term, which is more numerically stable than standard matrix calculations.

How to Cite

To cite the *relogit* Zelig model:

> Kosuke Imai, Gary King, and Oliva Lau. 2007. ”relogit: Rare Events Logistic Regression for Dichotomous Dependent Variables” 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

For more information see King and Zeng (2001a),King and Zeng (2001b),King and Zeng (2002). Sample data are from King and Zeng (2001a).

Bibliography

Andrews, D. W. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

King, G. and Zeng, L. (2001a), "Explaining Rare Events in International Relations," *International Organization*, 55, 693–715, http://gking.harvard.edu/files/abs/baby0s-abs.shtml.

— (2001b), "Logistic Regression in Rare Events Data," *Political Analysis*, 9, 137–163, http://gking.harvard.edu/files/abs/0s-abs.shtml.

— (2002), "Estimating Risk and Rate Levels, Ratios, and Differences in Case-Control Studies," *Statistics in Medicine*, 21, 1409–1427, http://gking.harvard.edu/files/abs/1s-abs.shtml.

Lumley, T. and Heagerty, P. (1999), "Weighted Empirical Adaptive Variance Estimators for Correlated Data Regression," *jrssb*, 61, 459–477.

Zeileis, A. (2004), "Econometric Computing with HC and HAC Covariance Matrix Estimators," *Journal of Statistical Software*, 11, 1–17.