

# 1 logit: Logistic Regression for Dichotomous Dependent Variables

Logistic regression specifies a dichotomous dependent variable as a function of a set of explanatory variables.

## 1.0.1 Syntax

```
> z.out <- zelig(Y ~ X1 + X2, model = "logit", data = mydata)
> x.out <- setx(z.out)
> s.out <- sim(z.out, x = x.out, x1 = NULL)
```

## 1.0.2 Additional Inputs

In addition to the standard inputs, `zelig()` takes the following additional options for logistic regression:

- **robust**: defaults to `FALSE`. If `TRUE` is selected, `zelig()` computes robust standard errors via the `sandwich` package (see [8]). 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 [1].
  - \* `"weave"`: HAC standard errors using the weights given in [5].
- **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 [8] for more options.

## 1.0.3 Examples

### 1. Basic Example

Attaching the sample turnout dataset:

```
> data(turnout)
```

Estimating parameter values for the logistic regression:

```
> z.out1 <- zelig(vote ~ age + race, model = "logit", data = turnout)
```

Setting values for the explanatory variables:

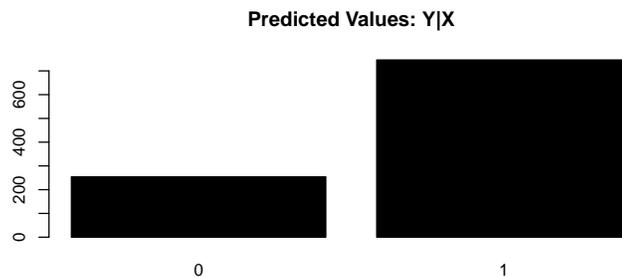
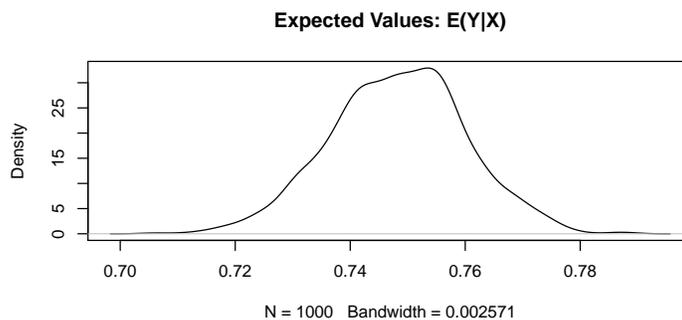
```
> x.out1 <- setx(z.out1, age = 36, race = "white")
```

Simulating quantities of interest from the posterior distribution.

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

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

```
> plot(s.out1)
```



## 2. Simulating First Differences

Estimating the risk difference (and risk ratio) between low education (25th percentile) and high education (75th percentile) while all the other variables held at their default values.

```
> z.out2 <- zelig(vote ~ race + educate, model = "logit", data = turnout)
```

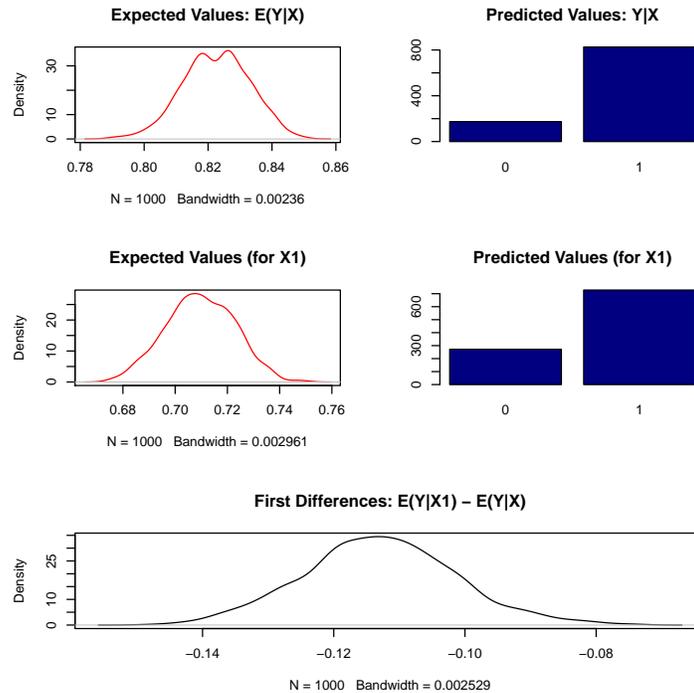
```
> x.high <- setx(z.out2, educate = quantile(turnout$educate, prob = 0.75))
```

```
> x.low <- setx(z.out2, educate = quantile(turnout$educate, prob = 0.25))
```

```
> s.out2 <- sim(z.out2, x = x.high, x1 = x.low)
```

```
> summary(s.out2)
```

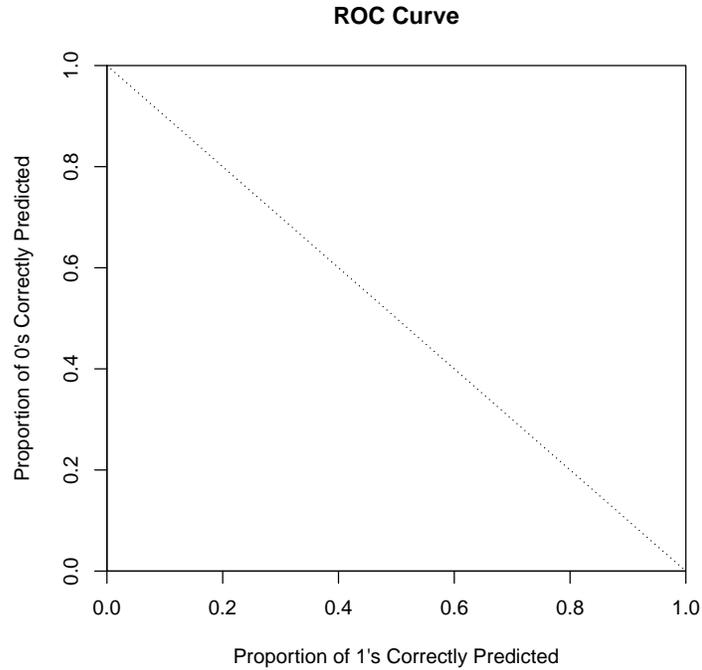
```
> plot(s.out2)
```



### 3. Presenting Results: An ROC Plot

One can use an ROC plot to evaluate the fit of alternative model specifications. (Use `demo(roc)` to view this example, or see King and Zeng (2002).)

```
> z.out1 <- zelig(vote ~ race + educate + age, model = "logit",  
+ data = turnout)  
> z.out2 <- zelig(vote ~ race + educate, model = "logit", data = turnout)  
  
> rocplot(z.out1$y, z.out2$y, fitted(z.out1), fitted(z.out2))
```



#### 1.0.4 Model

Let  $Y_i$  be the binary dependent variable for observation  $i$  which takes the value of either 0 or 1.

- The *stochastic component* is given by

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(y_i \mid \pi_i) \\ &= \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \end{aligned}$$

where  $\pi_i = \Pr(Y_i = 1)$ .

- The *systematic component* is given by:

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

where  $x_i$  is the vector of  $k$  explanatory variables for observation  $i$  and  $\beta$  is the vector of coefficients.

#### 1.0.5 Quantities of Interest

- The expected values (q̄) for the logit model are simulations of the predicted probability of a success:

$$E(Y) = \pi_i = \frac{1}{1 + \exp(-x_i\beta)}$$

given draws of  $\beta$  from its sampling distribution.

- The predicted values (`qi$pr`) are draws from the Binomial distribution with mean equal to the simulated expected value  $\pi_i$ .
- The first difference (`qi$fd`) for the logit model is defined as

$$\text{FD} = \Pr(Y = 1 \mid x_1) - \Pr(Y = 1 \mid x).$$

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

$$\text{RR} = \Pr(Y = 1 \mid x_1) / \Pr(Y = 1 \mid x).$$

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

$$\frac{1}{\sum_{i=1}^n t_i} \sum_{i:t_i=1}^n \{Y_i(t_i = 1) - E[Y_i(t_i = 0)]\},$$

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

$$\frac{1}{\sum_{i=1}^n t_i} \sum_{i:t_i=1}^n \{Y_i(t_i = 1) - Y_i(\widehat{t_i = 0})\},$$

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  $Y_i(\widehat{t_i = 0})$ , 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  $t_i = 0$ .

### 1.0.6 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 = "logit", 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.
  - `residuals`: the working residuals in the final iteration of the IWLS fit.
  - `fitted.values`: the vector of fitted values for the systemic component,  $\pi_i$ .
  - `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.
  - `data`: the name of the input data frame.
- 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  $\times$  `x`-observation (for more than one `x`-observation). Available quantities are:
  - `qi$ev`: the simulated expected probabilities for the specified values of `x`.
  - `qi$pr`: the simulated predicted values for the specified values of `x`.
  - `qi$fd`: the simulated first difference in the expected probabilities for the values specified in `x` and `x1`.
  - `qi$rr`: the simulated risk ratio for the expected 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.

## How to Cite the Logit Model

Kosuke Imai, Olivia Lau, and Gary King. *logit: Logistic Regression for Dichotomous Dependent*, 2011

## How to Cite the Zelig Software Package

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 logit model is part of the stats package by **(author?)** [7]. Advanced users may wish to refer to `help(glm)` and `help(family)`, as well as [6]. Robust standard errors are implemented via the sandwich package by **(author?)** [8]. Sample data are from [3].

## References

- [1] Donald W.K. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858, May 1991.
- [2] Kosuke Imai, Olivia Lau, and Gary King. *logit: Logistic Regression for Dichotomous Dependent*, 2011.
- [3] Gary King, Michael Tomz, and Jason Wittenberg. Making the most of statistical analyses: Improving interpretation and presentation. *American Journal of Political Science*, 44(2):341–355, April 2000. <http://gking.harvard.edu/files/abs/making-abs.shtml>.
- [4] Gary King and Langche Zeng. Improving forecasts of state failure. *World Politics*, 53(4):623–658, July 2002. <http://gking.harvard.edu/files/abs/civil-abs.shtml>.
- [5] Thomas Lumley and Patrick Heagerty. Weighted empirical adaptive variance estimators for correlated data regression. *jrssb*, 61(2):459–477, 1999.
- [6] Peter McCullagh and James A. Nelder. *Generalized Linear Models*. Number 37 in Monograph on Statistics and Applied Probability. Chapman & Hall, 2nd edition, 1989.
- [7] William N. Venables and Brian D. Ripley. *Modern Applied Statistics with S*. Springer-Verlag, 4th edition, 2002.
- [8] Achim Zeileis. Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, 11(10):1–17, 2004.