

## 0.1 lognorm: Log-Normal Regression for Duration Dependent Variables

The log-normal model describes an event's duration, the dependent variable, as a function of a set of explanatory variables. The log-normal model may take time censored dependent variables, and allows the hazard rate to increase and decrease.

### Syntax

```
> z.out <- zelig(Surv(Y, C) ~ X, model = "lognorm", data = mydata)
> x.out <- setx(z.out)
> s.out <- sim(z.out, x = x.out)
```

Log-normal models require that the dependent variable be in the form `Surv(Y, C)`, where `Y` and `C` are vectors of length  $n$ . For each observation  $i$  in  $1, \dots, n$ , the value  $y_i$  is the duration (lifetime, for example) of each subject, and the associated  $c_i$  is a binary variable such that  $c_i = 1$  if the duration is not censored (*e.g.*, the subject dies during the study) or  $c_i = 0$  if the duration is censored (*e.g.*, the subject is still alive at the end of the study). If  $c_i$  is omitted, all `Y` are assumed to be completed; that is, time defaults to 1 for all observations.

### Input Values

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

- **robust**: defaults to `FALSE`. If `TRUE`, `zelig()` computes robust standard errors based on sandwich estimators (see Huber (1981) and White (1980)) based on the options in `cluster`.
- **cluster**: if `robust = TRUE`, you may select a variable to define groups of correlated observations. Let `x3` be a variable that consists of either discrete numeric values, character strings, or factors that define strata. Then

```
> z.out <- zelig(y ~ x1 + x2, robust = TRUE, cluster = "x3",
               model = "exp", data = mydata)
```

means that the observations can be correlated within the strata defined by the variable `x3`, and that robust standard errors should be calculated according to those clusters. If `robust = TRUE` but `cluster` is not specified, `zelig()` assumes that each observation falls into its own cluster.

## Example

Attach the sample data:

```
> data(coalition)
```

Estimate the model:

```
> z.out <- zelig(Surv(duration, ciepl2) ~ fract + numst2, model = "lognorm",  
+               data = coalition)
```

View the regression output:

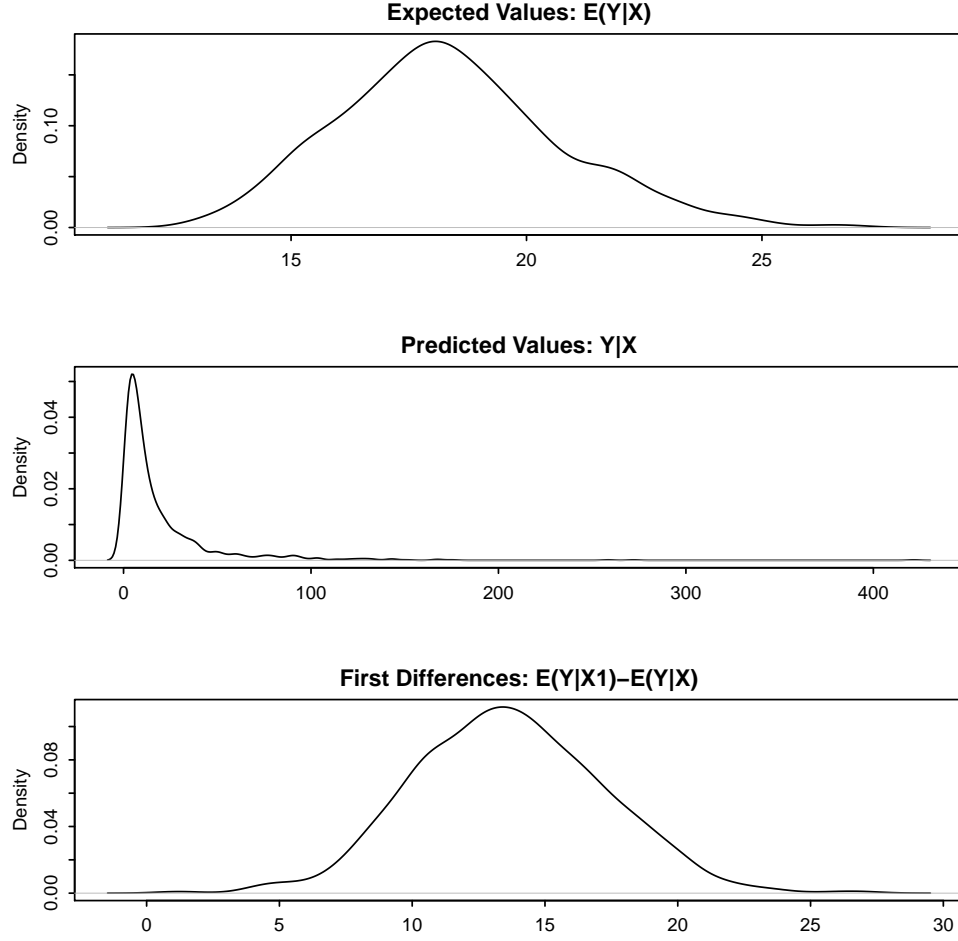
```
> summary(z.out)
```

Set the baseline values (with the ruling coalition in the minority) and the alternative values (with the ruling coalition in the majority) for X:

```
> x.low <- setx(z.out, numst2 = 0)  
> x.high <- setx(z.out, numst2 = 1)
```

Simulate expected values (`qi$ev`) and first differences (`qi$fd`):

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



## Model

Let  $Y_i^*$  be the survival time for observation  $i$  with the density function  $f(y)$  and the corresponding distribution function  $F(t) = \int_0^t f(y)dy$ . This variable might be censored for some observations at a fixed time  $y_c$  such that the fully observed dependent variable,  $Y_i$ , is defined as

$$Y_i = \begin{cases} Y_i^* & \text{if } Y_i^* \leq y_c \\ y_c & \text{if } Y_i^* > y_c \end{cases}$$

- The *stochastic component* is described by the distribution of the partially observed variable,  $Y^*$ . For the lognormal model, there are two equivalent representations:

$$Y_i^* \sim \text{LogNormal}(\mu_i, \sigma^2) \quad \text{or} \quad \log(Y_i^*) \sim \text{Normal}(\mu_i, \sigma^2)$$

where the parameters  $\mu_i$  and  $\sigma^2$  are the mean and variance of the Normal distribution. (Note that the output from `zelig()` parameterizes `scale=σ`.)

In addition, survival models like the lognormal have three additional properties. The hazard function  $h(t)$  measures the probability of not surviving past time  $t$  given survival up to  $t$ . In general, the hazard function is equal to  $f(t)/S(t)$  where the survival function  $S(t) = 1 - \int_0^t f(s)ds$  represents the fraction still surviving at time  $t$ . The cumulative hazard function  $H(t)$  describes the probability of dying before time  $t$ . In general,  $H(t) = \int_0^t h(s)ds = -\log S(t)$ . In the case of the lognormal model,

$$\begin{aligned} h(t) &= \frac{1}{\sqrt{2\pi}\sigma t S(t)} \exp\left\{-\frac{1}{2\sigma^2}(\log \lambda t)^2\right\} \\ S(t) &= 1 - \Phi\left(\frac{1}{\sigma} \log \lambda t\right) \\ H(t) &= -\log\left\{1 - \Phi\left(\frac{1}{\sigma} \log \lambda t\right)\right\} \end{aligned}$$

where  $\Phi(\cdot)$  is the cumulative density function for the Normal distribution.

- The *systematic component* is described as:

$$\mu_i = x_i\beta.$$

## Quantities of Interest

- The expected values (`qi$ev`) for the lognormal model are simulations of the expected duration:

$$E(Y) = \exp\left(\mu_i + \frac{1}{2}\sigma^2\right),$$

given draws of  $\beta$  and  $\sigma$  from their sampling distributions.

- The predicted value is a draw from the log-normal distribution given simulations of the parameters  $(\lambda_i, \sigma)$ .
- The first difference (`qi$fd`) is

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

- 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. When  $Y_i(t_i = 1)$  is censored rather than observed, we replace it with a simulation from the model given available knowledge of the censoring process. Variation in the simulations is due to two factors: uncertainty in the imputation process

for censored  $y_i^*$  and 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

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

where  $t_i$  is a binary explanatory variable defining the treatment ( $t_i = 1$ ) and control ( $t_i = 0$ ) groups. When  $Y_i(t_i = 1)$  is censored rather than observed, we replace it with a simulation from the model given available knowledge of the censoring process. Variation in the simulations are due to two factors: uncertainty in the imputation process for censored  $y_i^*$  and 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(Surv(Y, C) ~ X, model = "lognorm", 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.
  - **icoef**: parameter estimates for the intercept and  $\sigma$ .
  - **var**: Variance-covariance matrix.
  - **loglik**: Vector containing the log-likelihood for the model and intercept only (respectively).
  - **linear.predictors**: the vector of  $x_i\beta$ .
  - **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`.
- Most of this may be conveniently summarized using `summary(z.out)`. From `summary(z.out)`, you may additionally extract:
  - **table**: the parameter estimates with their associated standard errors,  $p$ -values, and  $t$ -statistics.

- 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$ev`: the simulated expected values for the specified values of `x`.
  - `qi$pr`: the simulated predicted values drawn from the distribution defined by  $(\lambda_i, \sigma)$ .
  - `qi$fd`: the simulated first differences between the simulated expected values for `x` and `x1`.
  - `qi$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

To cite the *lognorm* Zelig model:

Kosuke Imai, Gary King, and Oliva Lau. 2007. "lognorm: Log-Normal Regression for Duration 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

The exponential function is part of the survival library by by Terry Therneau, ported to R by Thomas Lumley. Advanced users may wish to refer to `help(survfit)` in the survival library, and Venables and Ripley (2002). Sample data are from King et al. (1990).

# Bibliography

Huber, P. J. (1981), *Robust Statistics*, Wiley.

King, G., Alt, J., Burns, N., and Laver, M. (1990), “A Unified Model of Cabinet Dissolution in Parliamentary Democracies,” *American Journal of Political Science*, 34, 846–871, <http://gking.harvard.edu/files/abs/coal-abs.shtml>.

Venables, W. N. and Ripley, B. D. (2002), *Modern Applied Statistics with S*, Springer-Verlag, 4th ed.

White, H. (1980), “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity,” *Econometrica*, 48, 817–838.