0.1 exp: Exponential Regression for Duration Dependent Variables

Use the exponential duration regression model if you have a dependent variable representing a duration (time until an event). The model assumes a constant hazard rate for all events. The dependent variable may be censored (for observations have not yet been completed when data were collected).

Syntax

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

Exponential 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, ..., n, the value y_i is the duration (lifetime, for example), 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 and is know to live at least as long as y_i). 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 exponential regression:

- `robust`: defaults to `FALSE`. If `TRUE`, `zelig()` computes robust standard errors based on sandwich estimators (see Huber (1981) and White (1980)) and the options selected 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, ciep12) ~ fract + numst2, model = "exp",
+     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)
```

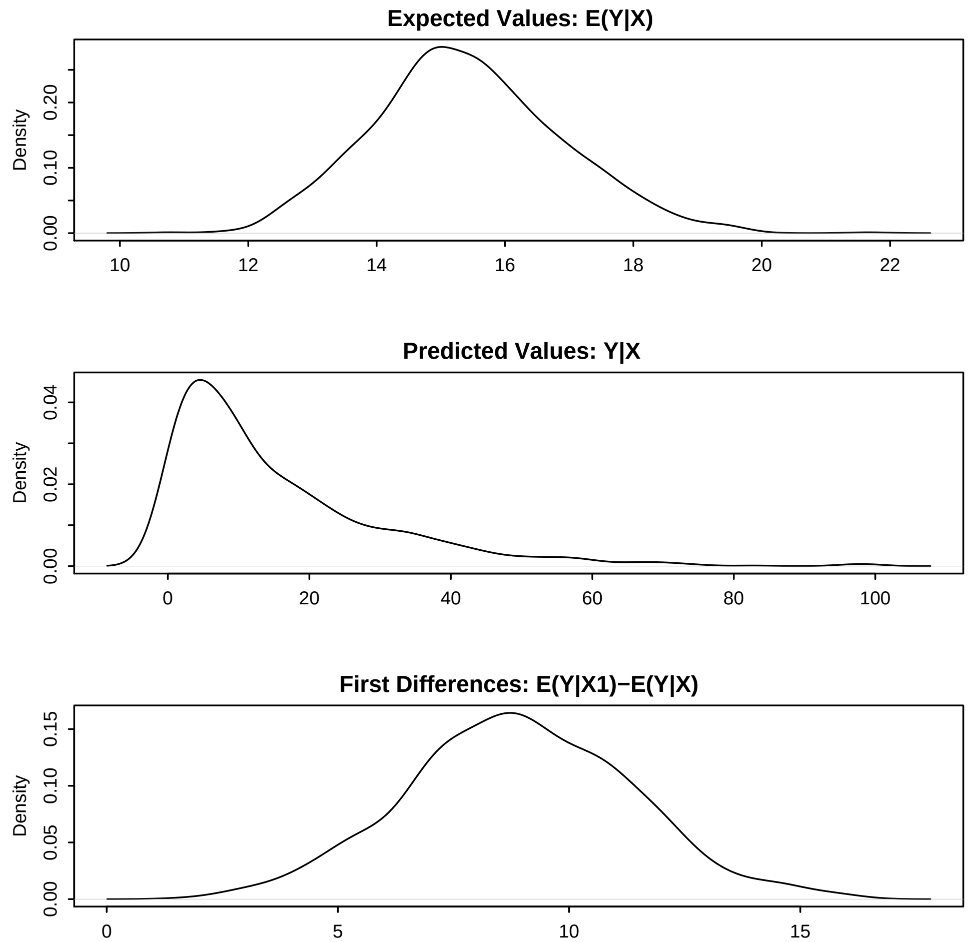
Simulate expected values (`qi$ev`) and first differences (`qi$fd`):

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

Summarize quantities of interest and produce some plots:

```
> summary(s.out)
```

```
> plot(s.out)
```



Model

Let Y_i^* be the survival time for observation i. 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^*. We assume Y_i^* follows the exponential distribution whose density function is given by

  $$f(y_i^* \mid \lambda_i) = \frac{1}{\lambda_i} \exp\left(-\frac{y_i^*}{\lambda_i}\right)$$

  for $y_i^* \geq 0$ and $\lambda_i > 0$. The mean of this distribution is λ_i.

  In addition, survival models like the exponential 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 exponential model,

$$\begin{aligned} h(t) &= \frac{1}{\lambda_i} \\ S(t) &= \exp\left(-\frac{t}{\lambda_i}\right) \\ H(t) &= \frac{t}{\lambda_i} \end{aligned}$$

For the exponential model, the hazard function $h(t)$ is constant over time. The Weibull model and lognormal models allow the hazard function to vary as a function of elapsed time (see Section **??** and Section **??** respectively).

- The *systematic component* λ_i is modeled as

$$\lambda_i = \exp(x_i \beta),$$

where x_i is the vector of explanatory variables, and β is the vector of coefficients.

Quantities of Interest

- The expected values (`qi$ev`) for the exponential model are simulations of the expected duration given x_i and draws of β from its posterior,

$$E(Y) = \lambda_i = \exp(x_i \beta).$$

- The predicted values (`qi$pr`) are draws from the exponential distribution with rate equal to the expected value.

- The first difference (or difference in expected values, `qi$ev.diff`), is

$$\text{FD} \; = \; E(Y \mid x_1) - E(Y \mid x), \tag{1}$$

where x and x_1 are different vectors of values for the explanatory variables.

- 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. 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

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

  where t_i is a binary explanatory variable defining the treatment ($t_i = 1$) and control ($t_i = 0$) groups. 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 $\widehat{Y_i(t_i = 0)}$, 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 = "exp", 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 scale parameter. While the scale parameter varies for the Weibull distribution, it is fixed to 1 for the exponential distribution (which is modeled as a special case of the Weibull).
  - `var`: the variance-covariance matrix for the estimates of β.
  - `loglik`: a 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. For example, `summary(z.out)$table`
- From the `sim()` output stored in `s.out`:
- 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 a distribution defined by the expected values.
  - `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 *exp* Zelig model:

> Kosuke Imai, Gary King, and Olivia Lau. 2007. "exp: Exponential 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`.

> Kosuke Imai, Gary King, and Olivia Lau. 2008. "Toward A Common Framework for Statistical Analysis and Development," *Journal of Computational and Graphical Statistics*, forthcoming, `http://gking.harvard.edu/files/abs/z-abs.shtml`.

See also

The exponential function is part of the survival library 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.