0.1 `aov`: Analysis of Variance for Continuous Dependent Variables

Model "aov" uses least squares regression to estimate the residual sum of squares and degrees of freedom for each explanatory variable around the best linear predictor for the specified dependent variables. Model "aov" is particularly useful for the analysis of randomized experiments with more than one strata or group (e.g., balanced incomplete block design). For the model with only one strata, "aov" reduces to "ls".

Syntax

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

where the `Error()` term is optional and takes strata formula.

Examples

1. Basic Example of aov.

   Attach sample data and set orthogonal contrasts:

   ```
   > data(npk, package = "MASS")
   > op <- options(contrasts = c("contr.helmert", "contr.poly"))
   ```

   Estimate the model (Venables and Ripley 2002, p.165):

   ```
   > z.out1 <- zelig(yield ~ block + N * P + K, model = "aov", data = npk)
   ```

   Summarize the fitted model:

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

   Set explanatory variables to their default (mean/mode) values

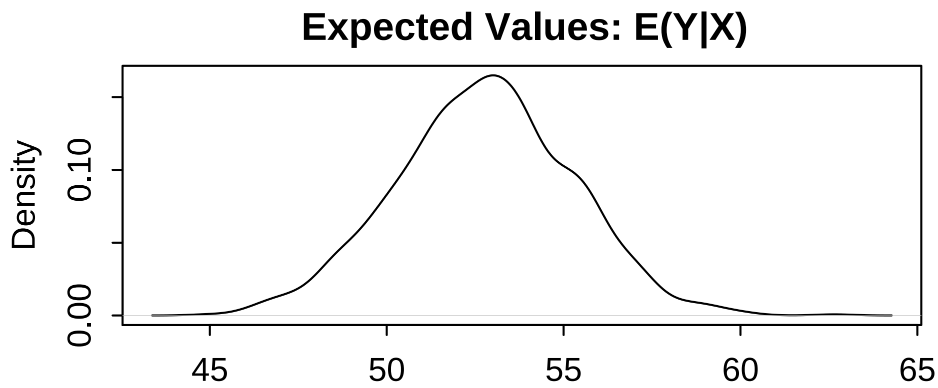
   ```
   > x <- setx(z.out1)
   ```

   Simulate model at values explanatory variables as in x

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

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

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



2. Example with `Error()` term allowing for more than one source of random variation in an experiment (multistratum model).

   Estimate the model:

   ```
   > z.out2 <- zelig(yield ~ N * P * K + Error(block), model = "aov",
   +      data = npk)
   ```

   Summarize regression coefficients:

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

   Set explanatory variables to their default (mean/mode) values

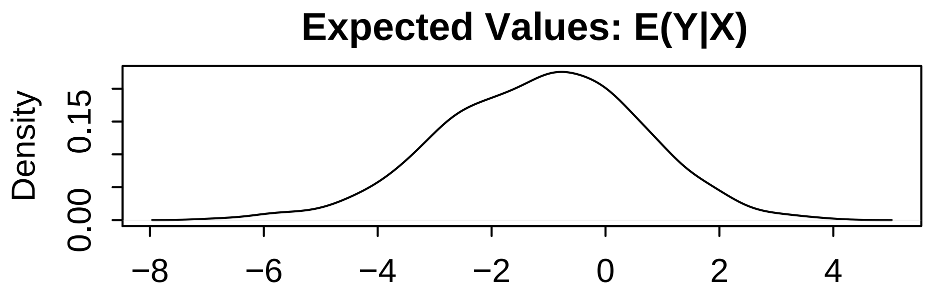
   ```
   > x <- setx(z.out2)
   ```

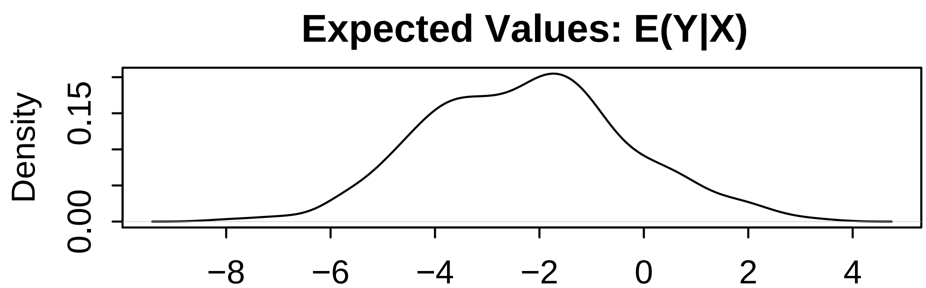
   Simulate model at values explanatory variables as in x

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

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

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





3. Example with *Error()* term (multistratum model) and first differences.

   Reset to previous contrasts

```
> options(op)
```

   Estimate model (Venables and Ripley 2002, p.283):

```
> z.out3 <- zelig(Y ~ N * V + Error(B/V), model = "aov", data = oats)
```

   Summarize regression coefficients:

```
> summary(z.out3)
```

   Set explanatory variables using mode

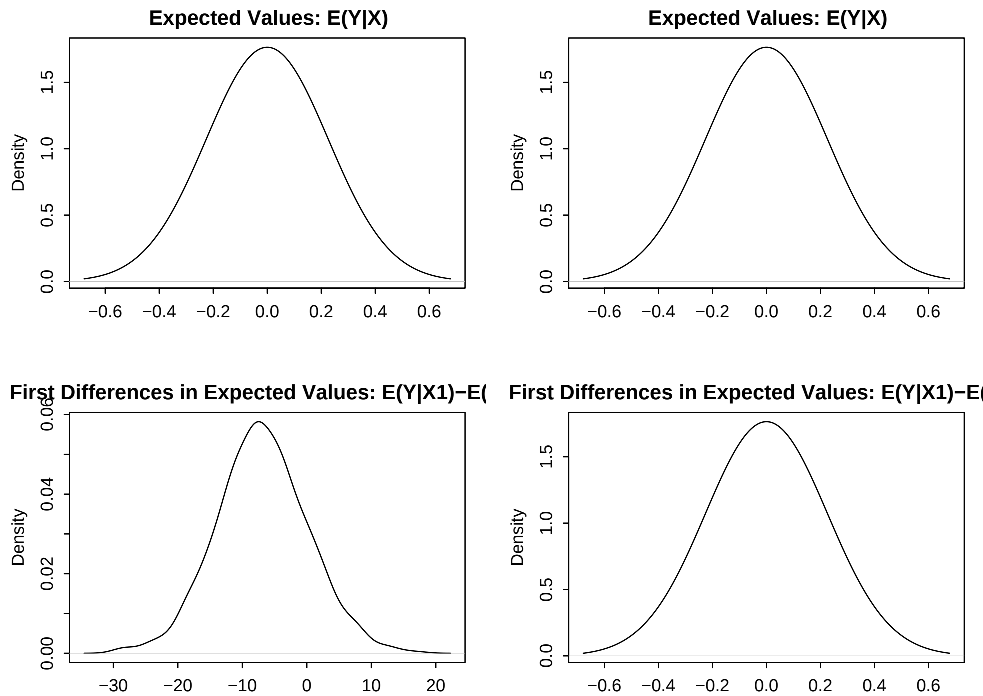
```
> x.out <- setx(z.out3, N = "0.0cwt", V = "Golden.rain")
> x.out1 <- setx(z.out3, N = "0.0cwt", V = "Victory")
```

   Simulate model at values explanatory variables as in x

```
> s.out3 <- sim(z.out3, x = x.out, x1 = x.out1)
```

```
> summary(s.out3)
```

```
> plot(s.out3)
```



Model

- The *stochastic component* is described by a normal density with mean μ_i and the common variance σ^2

$$Y_i \sim f(y_i \mid \mu_i, \sigma^2).$$

- The *systematic component* models the conditional mean as

$$\mu_i = x_i \beta$$

where x_i is the vector of covariates, and β is the vector of coefficients.

The least squares estimator is the best linear predictor of a dependent variable given x_i, and minimizes the sum of squared residuals, $\sum_{i=1}^{n}(Y_i - x_i\beta)^2$. The output of `aov` model is the sum of squared residuals. Note that `aov` is the same model as `ls` but the output values of function call `zelig` are different. You may check that `name(z.out)` returns the same arguments for the two models.

Quantities of Interest

- The expected value (`qi$ev`) is the mean of simulations from the stochastic component,

$$E(Y) = x_i \beta,$$

given a draw of β from its sampling distribution.

- 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 $t_i = 0$.

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 = "aov", 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`: fitted values.
  - `df.residual`: the residual degrees of freedom.
  - `zelig.data`: the input data frame if `save.data = TRUE`.
- From `summary(z.out)`, you may extract:
  - `coefficients`: the residuals sum of squares estimated with their associated degree of freedom, their mean squares, F-values, and F-statistics for all explanatory variables.
  - `residuals`: the sum of square, mean, degree of freedom, F-values, and F-statistics for the vector of residuals or standard errors that check the adequecy of the fit for the dependent variable versus the true values or data points.
- 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$fd`: the simulated first differences (or differences in expected values) for the specified values of `x` and `x1`.
- `qi$att.ev`: the simulated average expected treatment effect for the treated from conditional prediction models.

How to Cite

To cite the *aov* Zelig model:

> Kosuke Imai, Gary King, and Olivia Lau. 2007. "aov: Analysis of Variance for Continuous Dependent Variables," in Kosuke Imai, Gary King, and Olivia Lau, "Zelig: Everyone's Statistical Software," `http://gking.harvard.edu/zelig`. Elena Villalon implemented the software.

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 analysis of variance model is part of the stats package by William N. Venables and Brian D. Ripley (Venables and Ripley 2002). In addition, advanced users may wish to refer to `help(aov)` and `help(lm)`.

Bibliography

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