Evian package, EVIdential ANalysis

Lisa J. Strug, Ted Chiang

May 6, 2010

The `evian` package implements the EVidential ANalysis approach to interpret and analyze genetic association data. Evidential and standard frequentist analyses differ in significant thresholds, sample size estimation, and methods of adjustments for multiple hypothesis testing among other characteristics. `Evian` provides a graphical display incorporating all the evidence of association in a given dataset avoiding data reduction to a single measure like the p-value.

This documentation provides an introduction to the fundamental concepts underlying the `evian` package. It contains tools to analyze dichotomous and quantitative outcome data. Covariates can be accommodated.

1 A genetic association dataset

Human SNP genotyping provides a measurement of the genetic variation between individuals. A SNP is a single DNA base pair at a specific locus, consists of two alleles, and is found to be the cause of certain human diseases. A genetic association study may aim to test whether SNPs or genotype frequencies are responsible for disease.

A sample dichotomous dataset included in this package is `eviandata`. It consists of 250 individuals genotyped for 30 snps. A genotype call for a given SNP is coded as `0,1,2,NA` for the number of minor alleles. Individuals affected with disease are coded as (`1`), and unaffected as (`0`). Three covariates are provided, `age, weight, city`.

Similarly, a quantitative dataset included is `eviandata_linear`. It consists of 1444 individuals genotyped for 10 snps. The `Y` outcome is a continous variable called `Y_norma`. Three covariates are provided, `Fev, BMI_group, Age_group`.

The map for the SNP data is in `evianmap` and `evianmap_linear`. It consists of chromosome, snp, and position.

Loading the data for this example analysis is done by the following R statements:

```
> library(evian)
> data(eviandata)
> data(evianmap)
> data(eviandata_linear)
> data(evianmap_linear)
```

2 Using a logistic regression model for evidential analysis

First, we analyze the dichotomous dataset. The evian function for logistic regresssion is `evian_logit`. The dependent variable is `0,1` for presence or absence of disease. The five options for genetic models (inheritance pattern) are; `additive`, `dominant`, `recessive`, `overdominance`, and `2df` genotypic.

2.1 Additive model without covariates

The `evian_logit` function will compute four sets of likelihood intervals for the odds ratio: 1/8, 1/32, 1/100, 1/1000, and plot the segments on the y axis, and the chromosomal position on the x axis. In the following, formula `f` contains a string where `Y` is the phenotype, and `x` is the SNP parameter coded automatically for the model.

```
> f <- "Y ~ x"
> evian_logit(data = eviandata, map = evianmap,
+     ycol = 6, xcols = 10:39, formula = f, model = "additive")

[1] "Computing likelihood intervals..."
[1] "Plotting intervals..."
```

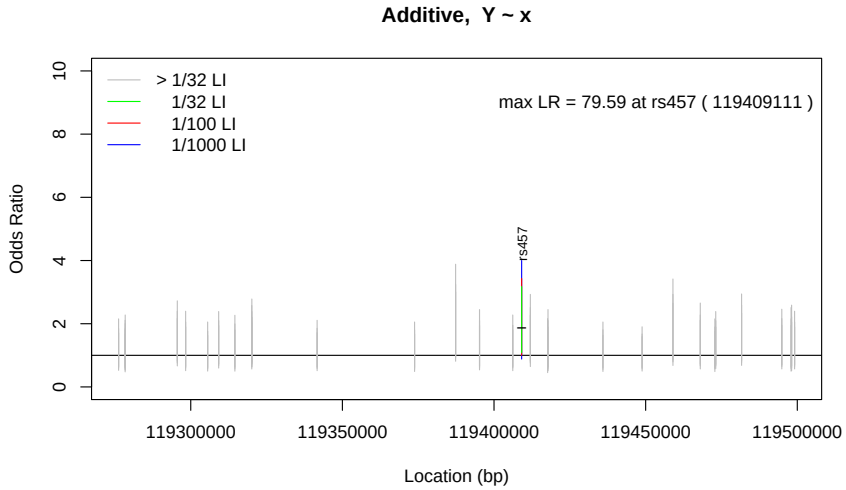

Figure 1: Likelihood intervals plot using logistic regression.

2.2 Additive model with one covariate (weight)

We now add a covariate into our logistic regression model such as `weight` (continuous variable). One must modify `f`, the glm formula for Y, e.g., `f <- Y ~ x + Z$weight` If the variable is categorical, the formula would be `f <- Y ~ x + as.factor(Z$weight)` .

By convention, `f` must follow a strict character syntax of using `Y` for the outcome, `x` for the snp data, `Z` for the covariate data. Each element is separated by single whitespaces. Furthermore, for logistic genetic models where the snp covariate is a 2df parameterization (e.g., `overdominance` or `2df`) genotypic, the formula would include a `x1` term, such as, `f <- Y ~ x + x1 + Z$age` . In sum, for analysis without covariates, the formula must be at least `Y ~ x` for 1df parameterization, and `Y ~ x + x1` for 2df.

```
> f <- "Y ~ x + Z$weight"
> evian_logit(data = eviandata, map = evianmap,
+     ycol = 6, xcols = 10:39, formula = f, model = "additive")

[1] "Computing likelihood intervals..."
[1] "Plotting intervals..."
```

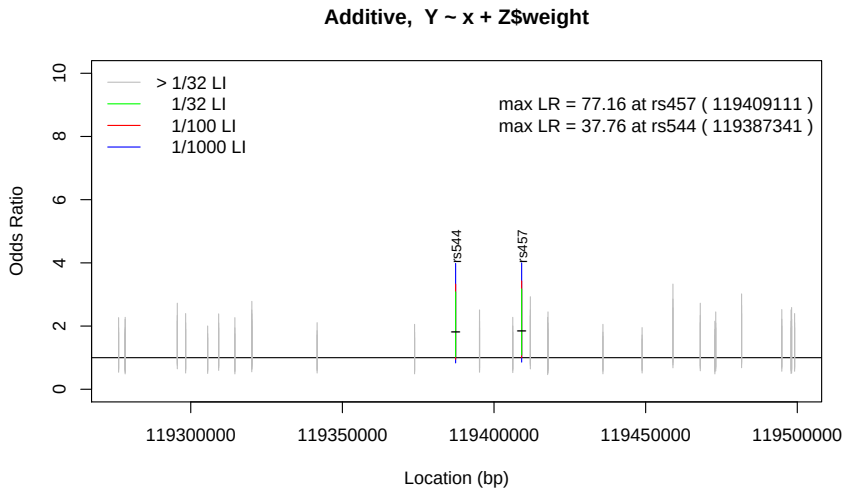

Figure 2: Likelihood intervals plot for the Additive genetic model with one covariate, weight.

2.3 Plotting likelihood curves for single SNPs

In Figure 1 and 2, rs457 and rs455 stand out as SNPs of interest when k=100 is used to demarcate evidence for association. Using `evian_logit_plotsnp`, we plot the standardized likelihood curve for snp rs457 to get a handle on the evidence for association.

```
> f <- "Y ~ x"
> evian_logit_plotsnp(snp = "rs457", data = eviandata,
+     map = evianmap, ycol = 6, xcols = 10:39, formula = f,
+     model = "additive")
```

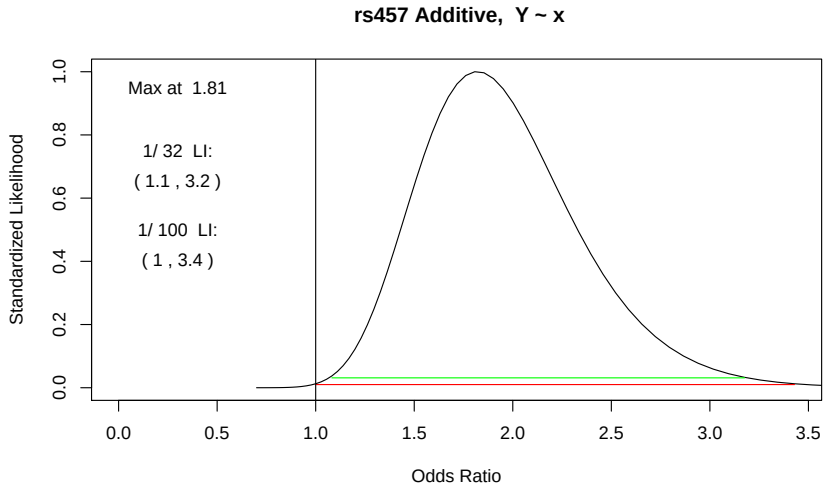

Figure 3: A Standardized Likelihood curve for rs457.

2.4 Likelihood curves with robust adjustment

We can apply a robust adjustment to the likelihood function to account for the cluster nature in the data, e.g. family id, FID. Using rs461, we generate two standardized likelihood curves with and without the robust feature.

```
> par(mfrow = c(2, 1))
> evian_logit_plotsnp(snp = "rs461", eviandata,
+     evianmap, 6, 10:39, model = "recessive", formula = f,
+     robust = TRUE)
> evian_logit_plotsnp(snp = "rs461", eviandata,
+     evianmap, 6, 10:39, model = "recessive", formula = f,
+     robust = FALSE)
```

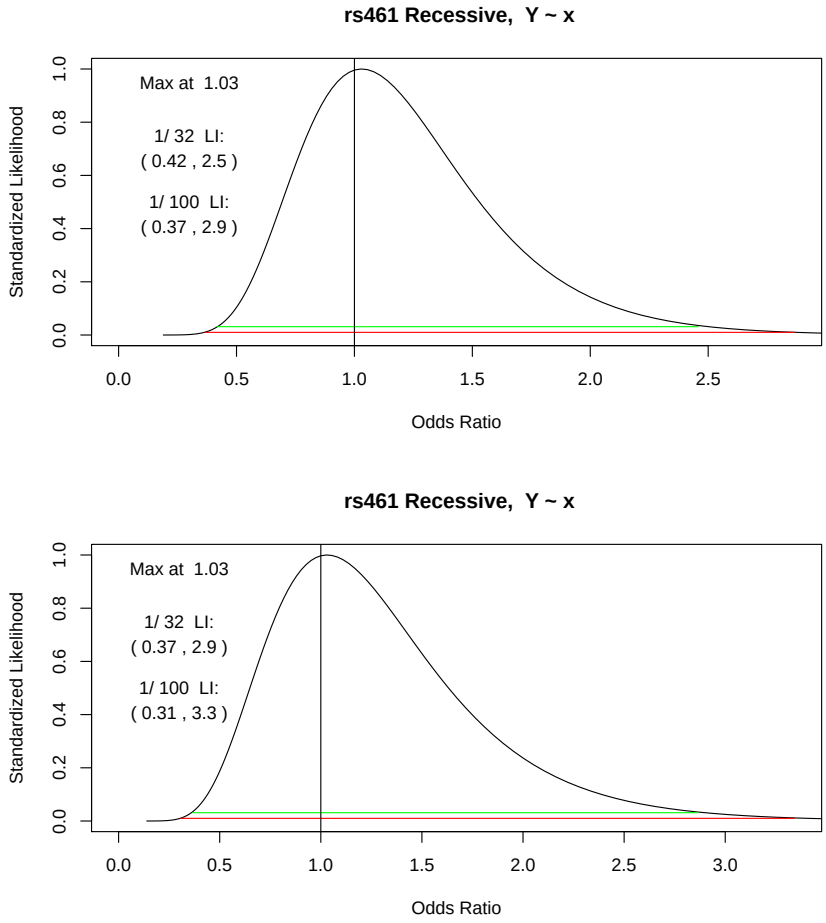

Figure 4: Robust adjustment for rs461.

3 Using a linear regression model for evidential analysis

We will apply the same analysis to a quantitative outcome dataset where the dependent variable is a continous `Y_norma`. The evian function for linear regression is `evian_linear`.

3.1 Dominant model without covariates

`evian_linear` will compute the same four sets of likelihood intervals for the beta: 1/8, 1/32, 1/100, 1/1000, and it plot the segments on the y axis, with the chromosomal position on the x axis. In the following R statements, formula `f` contains a string where `Y` is the quantitative trait, and `x` is the SNP parameter coded depending on the genetic model chosen. We choose the dominant model.

```
> f <- "Y ~ x"
> evian_linear(data = eviandata_linear, map = evianmap_linear,
+     ycol = 6, xcols = 10:19, formula = f, model = "dominant")

[1] "Computing Linear regression likelihood intervals..."
[1] "Your glm formulas:"
[1] "Y ~ -1 + Z$intercept"
[1] "X ~ Z$intercept"
[1] "dominant model - linear likelihood calculations done..."
[1] "Plotting intervals..."
```

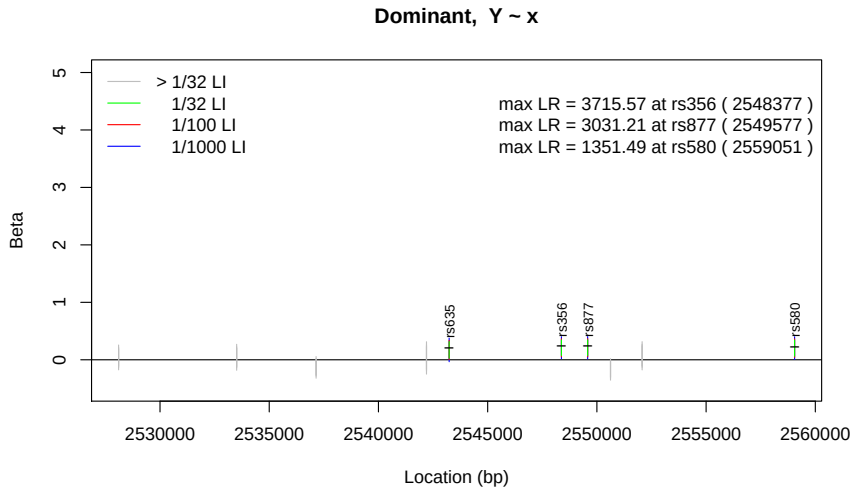

Figure 5: Likelihood intervals plot using linear regression.

3.2 Dominant model with covariates (Fev, BMI_group)

We can add any number of covariates into our linear regression model, such as a continuous variable called `Fev`, and a categorical variable called `BMI_group`. Thus, one must modify `f`, the glm formula for Y. Here for example, `f <-`

`Y ~ x + Z$Fev + as.factor(Z$BMI_group)` adjusts the model with two snp covariates which are the individual's Fev and BMI_group. If the variable is categorical, we would simply include the `as.factor()` to the variable.

By convention, `f` must follow a strict character syntax of using `Y` for the outcome, `x` for the snp data, and `Z` for the covariate data. As before, each element is to be separated by a single whitespace. For linear genetic models where the snp covariate is a 2df parameterization (e.g., `overdominance` or `2df` genotypic), the formula should remain the same as in 1df, e.g. `f <- Y ~ x + Z$Fev` . In other words, for analysis without covariates the formula will be `Y ~ x` for both 1df and 2df parameterizations. Thus, the `x1` term is omitted completely. This is an important difference between logistic and linear models that should be carefully noted to avoid errors.

```
> f <- "Y ~ x + Z$Fev + as.factor(Z$BMI_group)"
> evian_linear(data = eviandata_linear, map = evianmap_linear,
+     ycol = 6, xcols = 10:19, formula = f, model = "dominant")

[1] "Computing Linear regression likelihood intervals..."
[1] "Your glm formulas:"
[1] "Y ~ -1 + Z$intercept + Z$Fev + as.factor(Z$BMI_group)"
[1] "X ~ Z$intercept + Z$Fev + as.factor(Z$BMI_group)"
[1] "dominant model - linear likelihood calculations done..."
[1] "Plotting intervals..."
```

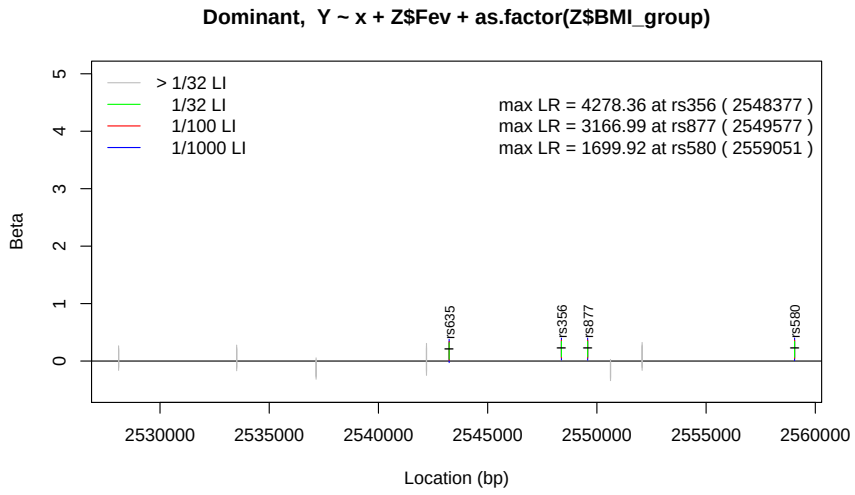

Figure 6: Likelihood intervals plot for the Dominant genetic model with two covariates, Fev and BMI_group.

3.3 Plotting likelihood curves for single SNPs

We observe in Figure 5 and 6 that rs356 and rs877 would be SNPs of interest when k=1000 is used to demarcate evidence for association. We now consider plotting the standardized likelihood curve for rs356 to get a handle on the evidence for association at this snp. The `evian_linear_plotsnp` will accomplish this with the following R statements:

```
> f <- "Y ~ x"
> evian_linear_plotsnp(snp = "rs356", data = eviandata_linear,
+     map = evianmap_linear, ycol = 6, xcols = 10:19,
+     formula = f, model = "dominant")
```

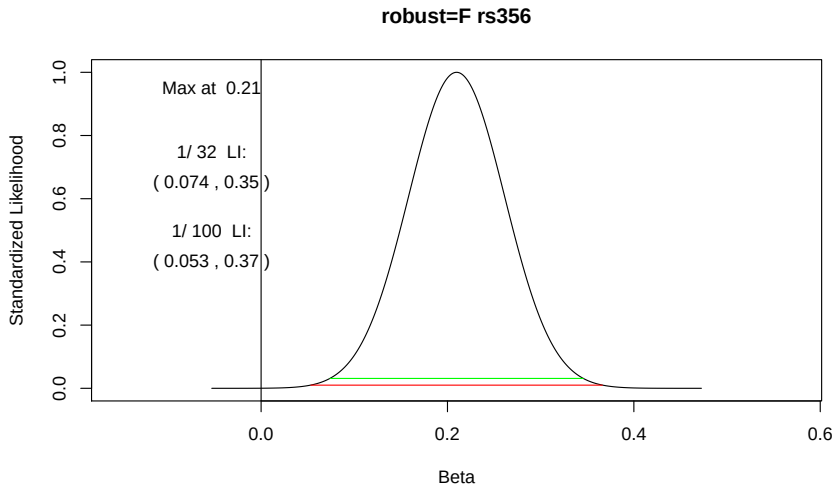

Figure 7: A Standardized Likelihood curve for rs356.

3.4 Likelihood curves with robust adjustment

As shown previously with the logistic regression, we can also apply a robust adjustment to the likelihood function to account for the cluster nature in the data, e.g. family id, FID. We demonstrate this feature using rs356 below.

```
> par(mfrow = c(2, 1))
> evian_linear_plotsnp(snp = "rs356", eviandata_linear,
+     evianmap_linear, 6, 10:19, model = "dominant",
+     formula = f, robust = TRUE)
> evian_linear_plotsnp(snp = "rs356", eviandata_linear,
```

```
+      evianmap_linear, 6, 10:19, model = "dominant",
+      formula = f, robust = FALSE)
```

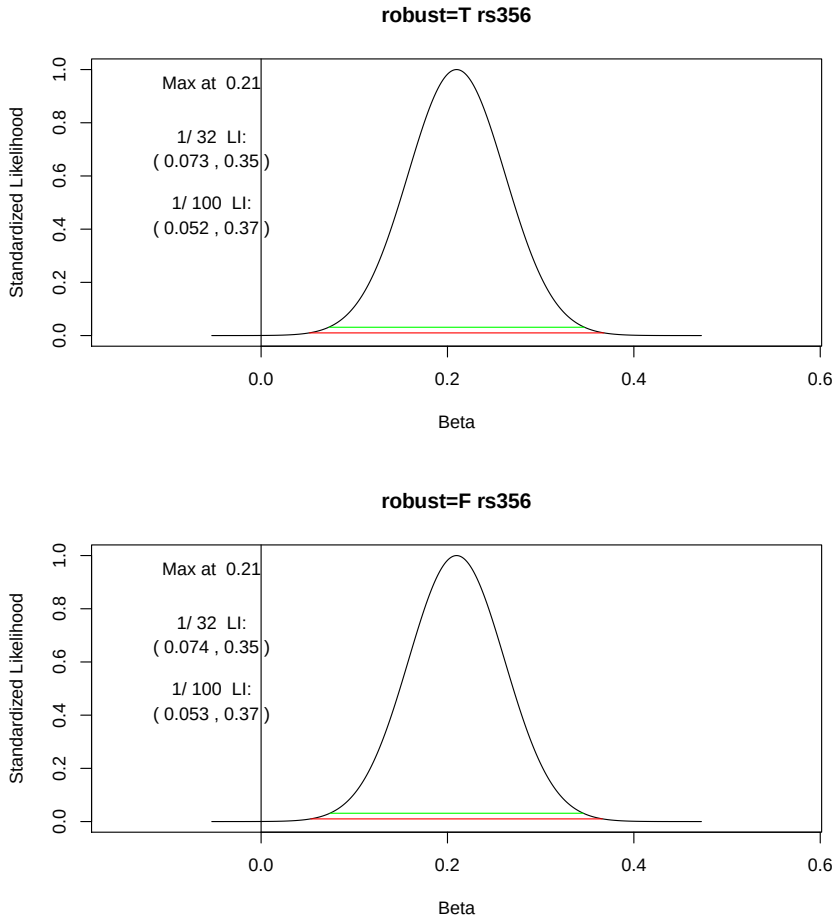

Figure 8: Robust adjustment for rs356.

4 Zooming feature

In Figure 2, one may wish to visualize the likelihood intervals plot of a defined chromosomal region of interest, such as the area surrounding rs457 and rs455. We can utilize `li_plot` in the following R statement to zoom:

```
> li_plot(bpstart = 119382000, bpend = 119414400,
+     dframe = data.additive, title = "Zoomed plot")

[1] "Plotting intervals..."
```

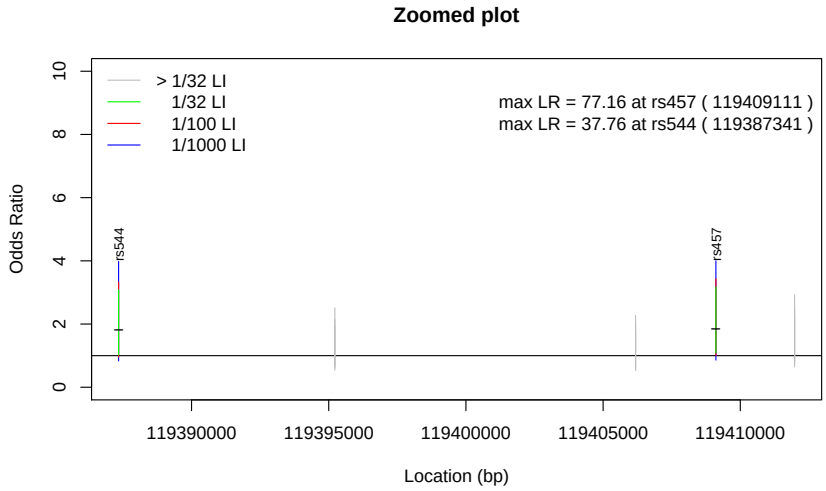

Figure 9: Likelihood intervals plot of a user-defined chromosomal region.

5 Notes on β grid and options

A β grid is defined as the grid of values for the snp parameter at which to evaluate the likelihood function. The density of the grid can be defined by the user in the `m` option. To constrain the β grid, one can define the number of β standard errors in the `bse` option. In other words for example, β grid is evaluated at $\hat{\beta}$ +/- 5 s.e. Alternatively, instead of defining the number of s.e., one can specify a lower limit for the grid, or the minimum value of the log(OR) (logistic) at which to calculate the likelihood function in the `lolim` option. And likewise the upper limit can be specified in the `hilim` option.

For logistic regression, evian will utilize the lolim/hilim values by default: `lolim` = log(0.025) and `hilim` = log(4), if the user doesn't define anything (ie. `bse` value is not defined). If `bse` is defined, then `lolim` and `hilim` will be ignored. Thus, if the user accidentally defines all three; `bse, lolim, hilim`, then evian will only use `bse` as it takes precedence.

However, for linear regression the opposite is true. Evian will utilize the bse values by default: `bse = 5` if the user doesn't define anything. If `lolim` and

`hilim` are defined, then evian will ignore the `bse` value. And if all three are defined, then `lolim, hilim` takes precedence.

In some cases the beta grid (using `bse` or `lolim/hilim`), may need to be increased substantially (`bse` as large as 15) if covariates are present in the formula.

When plotting likelihoods for single snps, the β grid limits may need to be adjusted or broadened in plots when the calculated likelihood intervals are not available, e.g., NAs provided.

Finally, estimation may become inaccurate with large number of correlated covariates, similar to known limitation of profile likelihoods.

6 References

- Strug, L.J., Rohde, C.A. and Corey, P.N. (2007). An introduction to evidential sample size calculations. American Statistician, 61, 207-212.
- Strug, L.J. and Hodge, S.E. (2006). An alternative foundation for the planning and evaluation of linkage analysis. I. Decoupling Error Probabilities from Measures of Evidence. Human Heredity, 61, 166-188.
- Strug, L.J. and Hodge, S.E. (2006). An alternative foundation for the planning and evaluation of linkage analysis. II. Implications for multiple test adjustments. Human Heredity, 61, 200-209.
- Strug, L.J., Clarke, T., Chiang, T., Chien, M., Baskurt, Z., Li, W., Dorfman, R., Bali, B., Wirrell, E., Kugler, S.L., Mandelbaum, D.E., Wolf, S.M., McGoldrick, P., Hardison, H., Novotny, E.J., Ju, J., Greenberg, D.A., Russo, J.J., Pal, D.K. (2009). Centrotemporal sharp wave EEG trait in Rolandic epilepsy maps to Elongator Protein Complex 4. European Journal of Human Genetics. 17:1171-1181, PMID:19172991.
- Royall, R.M,. Statistical evidence: a likelihood paradigm. ISBN:9780412044113.
- Edwards, A.W.F., Likelihood, Johns Hopkins University Press, 1992. ISBN:0801844436.
- Blume J.D. (2002). Tutorial in Biostatistics: Likelihood methods for measuring statistical evidence. Statistics in Medicine 21:2563-2599.
- Strug, L.J., Hodge, S.E., Chiang, T., Pal, DK., Corey, P.N., Rohde, C. (2010). A pure likelihood approach to the analysis of genetic association data: an alternative to Bayesian and frequentist analysis. European Journal of Human Genetics. Epub 2010, Apr 28; doi: 10.1038/ejhg.2010.47.
- Blume J.D., Su L., Olveda R.M., McGarvey S.T. (2007). Statistical evidence for GLM regression parameters: a robust likelihood approach. Stat Med. 26(15):2919-36.