Linear regression model: classical approach

The observed data

(i). Any single observation is given by $(Y, \boldsymbol{z})$:

— Y is the observed variable;
— $\boldsymbol{z} = (z_0, \ldots, z_k)$ is the covariate;
— $z_0 \in \{0, 1, 2\}$ is assumed to be a genotype.

The distribution of Y is assumed to be a normal; any element z_i of the covariate $\boldsymbol{z}$ can be

— nominal (*alias* categorical or factor);
— ordinal;
— quantitative (*alias* numeric).

Covariates:
— the covariate $z_0 \in \{0, 1, 2\}$ (genotype) is assumed to be of special ordinal/nominal type (object "genotype");
— type of each other covariate should be specified by user or selected by default;

By default
— a binary categorical covariate (2 levels) is assumed to be a nominal, and should be converted to 0 and 1 levels;
— a categorical covariate having 3-8 levels is assumed to be an ordinal;
— other covariates are assumed to be numeric.

(ii). The observed data is a sample $(\boldsymbol{Y}, \mathbf{z})$:
— $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ is $(n \times 1)$-vector;
— $\mathbf{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$, $\boldsymbol{z}_i = (z_{i,0}, \ldots, z_{i,k})$ are the covariates;
— $Y_1, \ldots, Y_n$ are conditionally independent on given $\mathbf{z}$;
— $\mathbb{V}\mathbf{ar}(Y \mid \boldsymbol{z}) = \sigma^2$ for any $\boldsymbol{z}$ ($\mathbb{V}\mathbf{ar}(Y_1) = \ldots = \mathbb{V}\mathbf{ar}(Y_n)$).

Linear regression model

(i). Regressor $\boldsymbol{X} : \mathbb{R}^k \to \mathbb{R}^m$; $\boldsymbol{X}(\boldsymbol{z})$ is defined for any observation.
(ii). Linear regression model for single observation

$$\mathbb{E}_\theta(Y \mid \boldsymbol{z}) = \boldsymbol{X}(\boldsymbol{z})'\boldsymbol{\beta},$$

— $\boldsymbol{\beta} = (\beta_1, \ldots \beta_m)'$ is $m \times 1$-vector of parameters;
— $Y \sim \mathcal{N}(\boldsymbol{X}(\boldsymbol{z})'\boldsymbol{\beta}, \sigma)$, where σ is the nuisance variance parameter ;
— $\boldsymbol{\theta} = (\beta_1, \ldots, \ldots \beta_m, \sigma)$ is the full parameter of the model.

The linear regression model is determined by the formula

$$Y \sim g(z_0) + f(\boldsymbol{z}^*), \quad \boldsymbol{z}^* = (z_1, \ldots, z_k),$$

where $g \in \{CD, D, R, A\}$ is the genetic model and f is the covariate term
— $f(\boldsymbol{z}^*) = f_1(\boldsymbol{z}^*) + \ldots + f_r(\boldsymbol{z}^*)$,
— $f_i(\boldsymbol{z}^*) = z_1^{r_{1i}} * \ldots * z_k^{r_{ki}}$, $r_{1i}, \ldots, r_{ki} \in \{0, 1, \ldots\}$ is of multiple regression type;
— *warning:* values $r_{ki} \geq 2$ are not available for factors.

The parameter.
— The intercept parameter is included into the model by default.

- The covariate z_0 generates
  - single parameter if $g \in \{\text{“}D, R, A\text{”}\}$ (difference with the baseline level $z_0 = 0$);
  - vector of length two if $g = \text{“}CD\text{”}$ (main effect with respect to the baseline level $z_0 = 0$).
- Term $f(\boldsymbol{z}^*)$ defines the (ordered) set K of subsets of $\{1, \ldots, k\}$, every $\kappa \in K$ will be equipped with the map $\mathrm{Deg}_\kappa \colon \kappa \to \mathbb{N}$ defined below:
  - consider the term $f_i(\boldsymbol{z}^*) = z_1^{r_{1i}} * \ldots * z_k^{r_{ki}}$ and the set $\kappa_i = \{\ell \mid r_{\ell i} > 0\}$;
  - then $K = \cup_{i=1}^{r} \mathbf{2}^{\kappa_i}$, where $\mathbf{2}^{\kappa_i}$ is a powerset of a set κ_i;
  - for each $\kappa \in K$ by definition $\mathrm{Deg}_\kappa(\ell) = \max\{r_{\ell i} \mid \kappa_i \supset \kappa, i = 1, \ldots, k\}$;
  - the elements of K should be sorted (*e.g.* κ's of smaller length may be sorted before κ's of larger length).

  We associate a number d_κ to each index $\kappa \in K$. For each $i \in \kappa$

  - define $d_i = \mathrm{Deg}_\kappa(i)$ if z_i is numeric or ordinal;
  - define d_i, so that $d_i + 1$ is a number of levels of the covariate z_i, if z_i is a factor. Note, that $\mathrm{Deg}_\kappa(i) = 1$. Also note, that in that case d_i does not actually depend on κ, which will be used below;
  - in general case $d_\kappa = \prod_{i \in \kappa} d_i$.

  The total length of the nuisance (part corresponding to f) parameter is $d = \sum_{\kappa \in K} d_\kappa$.

The regressor.

- The regressor is a vector (x_0, x_K) such that
  - $x_0 = (1, z_0)$ if $g \in \{\text{“}D, R, A\text{”}\}$;
  - $x_0 = (1, \mathbb{1}_{\{z_0=1\}}, \mathbb{1}_{\{z_0=2\}})$ if $g = \text{“}CD\text{”}$ (“Model for TEST”);
  - $x_0 = (1, z_0)$ if $g = \text{“}CD\text{”}$ (“Model for QAS”).
  - $x_0 = 1$ (“Baseline model”).

  We use the ordered set K to create regressor x_K for nuisance parameter by merging x_κ over $\kappa \in K$ in a proper order, where x_κ is defined as follows. For each $i \in \kappa$

  - define $x_i = (z_i, z_i^2, \ldots, z_i^{\mathrm{Deg}_\kappa(i)})$ if z_i is numeric or ordinal;
  - define x_i as the vector of length d_i of $\{0, 1\}$ with 1 on j-th position and 0 elsewhere if $z_i = a_{j+1}$, where a_j is the j-th level of z_i (for example $(0, 0, 1, 0, 0)$);
  - in general case $x_\kappa = \text{as.vector}(\otimes_{i \in \kappa} x_i)$, where $\otimes$ is the outer product (vector should be sorted in a proper order).

(iii). Linear regression model for the observed data

$$\mathbb{E}(\boldsymbol{Y} \mid \mathbf{z}) = \mathbf{X}'\boldsymbol{\beta},$$

- $\mathbf{X} = (\boldsymbol{X}(\boldsymbol{z}_1), \ldots, \boldsymbol{X}(\boldsymbol{z}_n))$ is the $m \times n$-matrix of regressors;
- $\boldsymbol{\beta}$ is the $1 \times m$-vector of parameters;
- assumption $\boldsymbol{Y} \sim \mathcal{N}(\mathbf{X}'\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, $\mathbf{I}_n$ is the identity matrix;
- the matrix $\mathbf{X} = (\mathbf{X}_g', \mathbf{X}^{*\prime})'$, where $\mathbf{X}_g$ contains the regressors related to g and $\mathbf{X}^*$ are the last d rows of the matrix $\mathbf{X}$.

Nuisance parameter reduction

- the parameter related to $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k)$ is not a subject of our interests;
- If the matrix $\mathbf{X}^*$, which contains the last d rows of the matrix $\mathbf{X}$, is not of the full rank d, we reduce the matrix $\mathbf{X}^*$ to $\mathbf{X}^\circ$ by choosing only basis vectors of the linear space generated by rows of the matrix $\mathbf{X}^*$;
- The reduced matrix $\mathbf{X}$ is determined by changing block $\mathbf{X}^*$ to $\mathbf{X}^\circ$ in the matrix $\mathbf{X}$.

QAS and P-value

Under $\mathbf{X}\mathbf{X}'$ is of full rank (positively definite), the least square estimator (LSE) is given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\,\boldsymbol{Y}.$$

The sum of squares

$$S(\boldsymbol{\beta}) = \|\,\boldsymbol{Y} - \mathbf{X}'\boldsymbol{\beta}\|^2 = (\,\boldsymbol{Y} - \mathbf{X}'\boldsymbol{\beta})'(\,\boldsymbol{Y} - \mathbf{X}'\boldsymbol{\beta}).$$

Notations

Note, that "Model for QAS" is equal to "Model for TEST" if $g =$ *"A,D,R"*.

- $\mathbf{X}$, $\mathbf{X}_q$, $\mathbf{X}_0$ are the regressors in the "Model for TEST", "Model for QAS", "Baseline model" respectively;
- $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\beta}}_q$, $\widehat{\boldsymbol{\beta}}_0$ is the LSE from the "Model for TEST", "Model for QAS", "Baseline model" respectively;
- $S(\widehat{\boldsymbol{\beta}})$, $S_q(\widehat{\boldsymbol{\beta}}_q)$, $S_0(\widehat{\boldsymbol{\beta}}_0)$ is the sums of squares from the "Model for TEST", "Model for QAS", "Baseline model" respectively.

The algorithm

The input data contain: "phenotype"(numeric vector), "genotype","covariates"(matrix),"covariate types"(vector) and "formula".

(i) Create regressors related to covariates $\mathbf{X}^*$.

(ii) Check that the rank of the matrix $\mathbf{X}^*$ is full.

  - if the rank of matrix $\mathbf{X}^*$ is not equal to d, apply nuisance parameter reduction $\mathbf{X}^* := \mathbf{X}^\circ$.

(iii) Loop by genotypes.

- Create regressors $\mathbf{X}_g$ for "Model for TEST", for "Model for QAS" and for "Baseline model".
- Check that the rank of the matrix $\mathbf{X}_g$ related to g is maximal (2 if $g =$ *"A,D,R"* or 3 if $g =$ *"CD"*):
  - if the rank of the matrix $\mathbf{X}_g$ in the "Model for TEST" is not maximal, report `qas`$[i]$ = `NA` and `pv`$[i]$ = `NA`;
  - otherwise, continue.
- Create matrix $\mathbf{X}$ by merging $\mathbf{X}_g$ and $\mathbf{X}^*$.
- If the rank of matrix $\mathbf{X}$ is not equal to its number of rows use nuisance parameter reduction to whole $\mathbf{X}$ ($\mathbf{X}_g$ should not be changed).
- Fit the "Model for QAS":
  - assign `qas`$[i]$ the coefficient, related to $\boldsymbol{z}_0$ (second element of the vector $\widehat{\boldsymbol{\beta}}$).
- Fit the "Model for TESTS" and the "Baseline model":
  - $SS_e = S(\widehat{\boldsymbol{\beta}})$; $\overline{SS}_e = SS_e/\mathrm{df}$, where $\mathrm{df} = n - \mathbf{rk}(\mathbf{X})$;
  - $SS_h = S_0(\widehat{\boldsymbol{\beta}}_0) - SS_e$; $\overline{SS}_h = SS_h/\mathrm{df}_h$, where $\mathrm{df}_h = 2$ if $g =$ *"CD"* and $\mathrm{df}_h = 1$ if $g =$ *"A,D,R"*;
  - $\mathbb{F} = \overline{SS}_h/\overline{SS}_e$ is the F-statistic, which has under null hypothesis $F_{\mathrm{df}_h,\mathrm{df}}$-distribution;
  - assign `pv`$[i] = 1 - F_{\mathrm{df}_h,\mathrm{df}}(\mathbb{F})$, where $F_{\mathrm{df}_h,\mathrm{df}}$ is the distribution function of the Fisher–Snedecor distribution $F_{\mathrm{df}_h,\mathrm{df}}$.

(iv) Return the vectors `pv` and `qas`.