STATMODEL DOCUMENTATION

ANDREA ZORRO-ARANDA

November 2021

Based on the Statmodel method (Hernandez, 2018b), we proposed a novel method for GRN inference. This method is an alternative tool for statistical modeling and analysis of experiments than ANOVA. Since this methodology allows us to determine the influence of independent factors (TF expression) over a response variable (TG expression), we found it as a proper tool for GRN inference. For this purpose, we proposed a modification of the methodology initially presented and a suggested score for the interaction reliability. Statmodel presents some advantages concerning ANOVA, such as no assumption of the data distributions and the minimization of the variance of the residual error probability model, reducing the chances of over-fitting. Moreover, this methodology reduces the spurious effects minimizing the number of predictor variables, which is mainly what we look for in GRN inference.

The method is based on the following model where a response variable (Y) is represented as a linear combination of the predictor variables (X_i)

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \varepsilon \tag{1}$$

where β_i is the coefficient for each one of the predictors, β_0 is the independent coefficient and ε is the random error of the model. To find the best model representing the response variable, we minimize the error variance maintaining the model unbiased and parsimonious. This is achieved through the following optimization problem:

$$\begin{aligned} &\min_{\beta} \{Var(\varepsilon), n\} \\ &\text{s. to } E(\varepsilon) = 0 \end{aligned} \tag{2}$$

where ε is obtain form Equation (1) as

$$\varepsilon = Y - \beta_0 - \sum_{i=1}^{n} \beta_i (X_i) \tag{3}$$

To accomplish the restriction $E(\varepsilon) = 0$ and reduce the effect caused by different orders of magnitudes among the variable, we can rewrite the Equation (1) as

$$\frac{Y - \langle Y \rangle}{\max(Y) - \min(Y)} = \sum_{i=1}^{n} \beta_i^* \frac{X_i - \langle X_i \rangle}{\max(X_i) - \min(X_i)} + \varepsilon^* \tag{4}$$

which is obtained by replacing

$$\beta_0 = \langle Y \rangle - \sum_{i=1}^{n} \beta_i \langle X_i \rangle \tag{5a}$$

$$\beta_i = \beta_i^* \frac{\max(Y) - \min(Y)}{\max(X_i) - \min(X_i)} \tag{5b}$$

$$\varepsilon = \varepsilon^* (\max(Y) - \min(Y)) \tag{5c}$$

Considering that the expected value of the average value of a random variable X is

$$E(\langle X \rangle) = E(X) \tag{6}$$

From Equations (4), (5c) and (6)

$$\begin{aligned} E(\varepsilon) &= (\max(Y) - \min(Y)) E(\varepsilon^*) \\ &= E(Y - \langle Y \rangle) - \sum_{i=1}^{n} \beta_i^* \frac{\max(Y) - \min(Y)}{\max(X_i) - \min(X_i)} \\ &\quad E(X_i - \langle X_i \rangle) \\ &= 0 \end{aligned} \tag{7}$$

Therefore, the restriction in Equation (2) is fulfilled and the predictions of the model can be considered unbiased.

Moreover, to obtain a parsimonious model, we should reduce the number of parameters of the model. This can be done through a hypothesis test to find the coefficients β_i^* that are significantly different from zero, as following

$$\begin{aligned} H_0 &: \beta_i^* = 0 \\ H_a &: \beta_i^* \neq 0 \end{aligned} \tag{8}$$

To perform this test without making assumptions of the distributions of the variables, we can rewrite Equation (4) as

$$\begin{aligned} \frac{Y - \langle Y \rangle}{\max(Y) - \min(Y)} &= \beta_i^* \frac{X_i - \langle X_i \rangle}{\max(X_i) - \min(X_i)} \\ &\quad + \sum_{i \neq j} \beta_j^* \frac{X_i - \langle X_i \rangle}{\max(X_i) - \min(X_i)} + \varepsilon^* \\ &= \beta_i^* \frac{X_i - \langle X_i \rangle}{\max(X_i) - \min(X_i)} + \varepsilon_i^* \end{aligned} \tag{9}$$

where

$$\varepsilon_i^* = \sum_{i \neq j} \beta_j^* \frac{X_i - \langle X_i \rangle}{\max(X_i) - \min(X_i)} + \varepsilon^* \tag{10}$$

is a random variable consolidating all the effects different from X_i, maintaining the restriction $E(\varepsilon_i^*) = 0$.

Therefore, we can evaluate each predictor variable independently. Considering two subgroups from the data, one positive $X_i \geq \langle X_i \rangle$ and one negative $X_i \leq \langle X_i \rangle$, we can perform the hypothesis test. If the expected value of the standardized response variable $\left(\frac{Y - \langle Y \rangle}{\max(Y) - \min(Y)} \right)$ for the positive group is different form the one for the negative group (Equation (11), the coefficient β_j^* is considered to be significantly different from zero.

$$\begin{aligned} H_0 :& E(Y|X_i \geq \langle X_i \rangle) = E(Y|X_i \leq \langle X_i \rangle) \\ H_a :& E(Y|X_i \geq \langle X_i \rangle) \neq E(Y|X_i \leq \langle X_i \rangle) \end{aligned} \tag{11}$$

To perform the hypothesis test, first, we find the probability distribution that better adjusts to Y. Then, we evaluate if the data for the smallest subgroup is adjusted to this probability distribution. To evaluate this, we applied a χ^2 test in Matlab. The coefficient β_j^* is considered to be significantly different from zero if the alternative hypothesis cannot be rejected. This means that the null hypothesis is rejected by the χ^2 test. We selected the smallest group since it has less statistical power for the hypothesis rejection (Hernandez, 2018a). Thus, a hypothesis rejected with the smallest group will be rejected with the other one. Then, the non-zero β_i^* can be computed as

$$\beta_i^* = \frac{Cov(Y, X_i)}{Var(X_i)} \cdot \frac{\max(Y) - \min(Y)}{\max(X_i) - \min(X_i)} \tag{12}$$

This β_j^* gives us the minimum $Var(\varepsilon_j^*)$.

For the GRN inference, we proposed as a score of reliability of the interaction the $-\log(\text{p-value})$ of the χ^2 test. This, since we look for the TF that most affect the expression of the evaluated TG, and according to the method should be the ones that produce the most different distributions of Y between both subgroups. Thus, they are the ones that have the smaller p-values of the χ^2 test.

REFERENCES

Hernandez, Hugo (2018a). "Parameter Identification using Standard Transformations: An Alternative Hypothesis Testing Method". In: *ForsChem Research Reports* 4.

Hernandez, Hugo (2018b). "Statistical Modeling and Analysis of Experiments without ANOVA". In: *ForsChem Research Reports* 5.