From SNPs to genes

David Lamparter in collaboration with Stefan Milosavljevic

May 27, 2014

1 Example: simple linear regression

Assuming that a measurement of a SNP in 1000 people was done and a measurement about their phenotypes (e.g. the height) was done too. The question of immediate interest is: Does the SNP influence the height phenotype? Or more specifically: Does the people who carry the allele with the SNP differ significantly in height compared to the people who carry the other allele? The aim of the answer to this question is to make a statement that holds in general and not about a particular group that was chosen to measure. The idea is to see if the SNPs can have a significant impact on the phenotype. To formalize the question, a model is needed and it should contain a phenotype measure (Y_i) and a genotype value (X_i). The starting model is:

$$Y_i = X_i b + \varepsilon_i \tag{1}$$

Where b represents the association between genotype and phenotype and ε_i is a random variable normally distributed with mean 0 and variance σ^2 and can be written $\epsilon \approx N(0, \sigma^2)$. To be more clear about the meaning of b in the formula, we can have an example. If we assume that $b = 0$, so that there is no association between genotype and phenotype the formula (1) becomes:

$$Y_i = \varepsilon_i \tag{2}$$

Which means that the phenotype value will be random, so the phenotype will be random and independent from the genotype. Now the question of interest is:
Is $b = 0$ or not?
To answer this question, a first step must be done and it consists in finding a value for b when the effect of ε_i is minimal:

$$\frac{\partial}{\partial x}(\sum_{i=1}^{n} \varepsilon_i^2) = 0 \tag{3}$$

If the formula is developed:

$$\begin{aligned}\frac{\partial}{\partial x}(\sum_{i=1}^{n} \varepsilon_i^2) &= \frac{\partial}{\partial b}(\sum_{i=1}^{n} Y_i - X_i b)^2 = \sum_{i=1}^{n} \frac{\partial}{\partial b}(Y_i - X_i b)^2 \\ &= \sum_{i=1}^{n} 2(X_i b - Y_i)X_i = \sum_{i=1}^{n} 2X_i^2 b - \sum_{i=1}^{n} 2X_i Y_i \end{aligned} \tag{4}$$

With the last equality, a new b can be found:

$$\sum_{i=1}^{n} 2X_i^2 b - \sum_{i=1}^{n} 2X_i Y_i = 0 \Rightarrow \widehat{b} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \tag{5}$$

Where $\widehat{b}$ is a good estimator of b. The next step is finding the expectation and the variance of $\widehat{b}$:

$$E[\,\widehat{b}\,] = E[\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}] = \frac{1}{\sum_{i=1}^{n} X_i^2} \sum_{i=1}^{n} X_i E[Y_i] = 0 \tag{6}$$

$$\begin{aligned} Var(\widehat{b}) &= E[(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2})^2] = \frac{1}{(\sum_{i=1}^{n} X_i^2)^2} E[(\sum_{i=1}^{n} Y_i X_i)^2] \\ &= \frac{1}{(\sum_{i=1}^{n} X_i^2)^2} E[\sum_{i=1}^{n} \sum_{j=1}^{n} X_i Y_i X_j Y_j] \\ &= \frac{1}{(\sum_{i=1}^{n} X_i^2)^2} \sum_{i=1}^{n} X_i^2 = \frac{1}{\sum_{i=1}^{n} X_i^2} \end{aligned} \tag{7}$$

If $\widehat{b}$ is scaled following the variance:

$$b^* = \widehat{b} \sqrt{\sum_{i=1}^{n} X_i^2} \tag{8}$$

Where b^* has $E[b^*] = 0$ and $Var(b^*) = 1$, so it's under a **normal law**. Now that the distribution of b^* is known, many simulations can be done to have a vector that contains many of these values.