note0202

Daniel Sorensen

February 13, 2025

Test of Association Using the Likelihood Ratio Test

This note illustrates how to test for association between a single genetic marker (X, the covariate, a column vector of length n) and a continuous trait (the phenotype, y, a column vector of responses of length n) using the likelihood ratio test. The context could be a genomewide association study, where a large number of genetic markers are studied, one at a time. Here, the focus is on one such marker; the test of hypothesis using the ratio of likelihoods requires first defining the so-called *full or unrestricted* model and the *restricted or null* model.

The Likelihood Function and the ML Estimators

The *full or unrestricted* statistical model is defined as follows:

1. Inferences are conditional on the observed vector of covariates X that can be treated as fixed or random

2. For an observed value $X_i = x_i$, y_i is linked to x_i via the relationship

$$y_i = \beta_0 + \beta_1 x_i + e_i \tag{1}$$

   where β_0 and β_1 are unobserved parameters to be estimated

3. The residual term e_i is normally distributed with mean zero and variance σ^2 and is independent of X_i. This holds for all i, $i = 1, \ldots n$. The terms e_i are independent across all responses

4. It follows that the elements of y are conditionally independent, given the covariate x

With these definitions, the loglikelihood can be written

$$\begin{aligned}
\ell\left(y|x,\beta_0,\beta_1,\sigma^2\right) &= \ln\left[\prod_{i=1}^{n} p\left(y_i|x_i,\beta_0,\beta_1,\sigma^2\right)\right] \\
&= \sum_{i=1}^{n} \ln\left[p\left(y_i|x_i,\beta_0,\beta_1,\sigma^2\right)\right] \\
&= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \beta_1 x_i\right)\right)^2 . \qquad (2)
\end{aligned}$$

The loglikelihood (2) is maximised with respect to β_1, β_0 and σ^2 for

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)\left(x_i-\overline{x}\right)}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}, \qquad \overline{y}=\frac{1}{n}\sum\nolimits_{i=1}^{n} y_i, \quad \overline{x}=\frac{1}{n}\sum\nolimits_{i=1}^{n} x_i, & \text{(3a)}\\ \hat{\beta}_0 &= \overline{y}-\hat{\beta}_1\overline{x}, & \text{(3b)}\\ \hat{\sigma}^2 &= \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(y_i-\left(\hat{\beta}_0+\hat{\beta}_1 x_i\right)\right)^2. & \text{(3c)}\end{aligned}$$

We shall also need to define the *restricted or null* statistical model that assigns the value of zero to β_1. The model takes the form

$$y_i = \mu + \epsilon_i,$$

where $\epsilon_i \sim N\left(0, \sigma_0^2\right)$ for all i, with the $\epsilon_i' s$ independent across observations. The loglikelihood is

$$\ell\left(y|\mu, \sigma_0^2\right) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma_0^2 - \frac{1}{2\sigma_0^2}\sum\nolimits_{i=1}^{n}\left(y_i-\mu\right)^2.$$

Maximisation with respect to μ and σ_0^2 yields the ML estimators of the restricted model

$$\begin{aligned}\hat{\mu} &= \frac{1}{n}\sum\nolimits_{i=1}^{n} y_i, & \text{(4a)}\\ \hat{\sigma}_0^2 &= \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(y_i-\hat{\mu}\right)^2, & \text{(4b)}\end{aligned}$$

where the second line is the sampling variance of the observations.

The Likelihood Ratio Test

In the present example the parameters of the full model $\left(\beta_0, \beta_1, \sigma^2\right)$ take values in some region Ω of the parameter space and those of the restricted model $\left(\mu, \sigma_0^2\right)$ take values in a subspace of Ω that is labelled ω. This defines a nested setup: the null or restricted model is nested within the full model. The full model can always explain the data at least as well as the null model; one is interested to know whether this difference is significant.

A well established test statistic to compare nested models that can be used provided that certain regularity conditions are met, is based on the approximation (the likelihood ratio test or loglikelihood ratio test)

$$\lambda = -2\ln\left[\frac{L_{\max}(\omega)}{L_{\max}(\Omega)}\right] \tag{5}$$

(Wilks, 1938), where $L_{\max}(\omega)$ is the maximum of the likelihood of the data based on the restricted model and $L_{\max}(\Omega)$ is the maximum of the likelihood of the full model. If the null hypothesis is true (under the restricted model), λ has an asymptotic chi-square distribution with degrees of freedom equal to the difference in the number of parameters of the full and restricted models and non-centrality parameter equal to zero. The important requirements for the approximation to hold is that the ML estimator does not lie on a boundary point, that the parameters are real numbers taking values in some interval and that the models are nested. Cox and Hinkley (1974) and Stuart and Ord (1991) give an account of the classical theory and an accessible proof of the asymptotic distribution of λ can be found in Sorensen and Gianola (2002).

Returning to the example of the simple linear regression, the loglikelihood of the full model when parameters are replaced by the ML estimates is

$$\begin{aligned}\ell\left(y|x,\hat{\beta}_0,\hat{\beta}_1,\hat{\sigma}^2\right) &= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2}\sum\nolimits_{i=1}^{n}\left(y_i-\left(\hat{\beta}_0+\hat{\beta}_1x_i\right)\right)^2 \\ &= -\frac{n}{2}\left(1+\ln 2\pi\right) - \frac{n}{2}\ln\hat{\sigma}^2,\end{aligned}$$

where in the first line $\sum_{i=1}^{n}\left(y_i-\left(\hat{\beta}_0+\hat{\beta}_1x_i\right)\right)^2$ is replaced by $n\hat{\sigma}^2$. For the restricted model we have

$$\begin{aligned}\ell\left(y|\hat{\mu},\hat{\sigma}_0^2\right) &= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\hat{\sigma}_0^2 - \frac{1}{2\hat{\sigma}_0^2}\sum\nolimits_{i=1}^{n}\left(y_i-\hat{\mu}\right)^2 \\ &= -\frac{n}{2}\left(1+\ln 2\pi\right) - \frac{n}{2}\ln\hat{\sigma}_0^2,\end{aligned}$$

where $\sum_{i=1}^{n}\left(y_i-\hat{\mu}\right)^2 = n\hat{\sigma}_0^2$. Substituting in (5)

$$\begin{aligned}\lambda &= 2\left(\ell\left(y|x,\hat{\beta}_0,\hat{\beta}_1,\hat{\sigma}^2\right) - \ell\left(y|\hat{\mu},\hat{\sigma}_0^2\right)\right) \\ &= n\ln\left[\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right] \sim \chi^2\left(1\right).\end{aligned} \tag{6}$$

Therefore the $p-$value can be obtained by comparing λ to a central chi-square distribution with 1 degree of freedom.

Write the equation for the data (1) in the alternative form

$$y_i = \hat{y}_i + \hat{e}_i, \tag{7}$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1x_i$ and $\hat{e}_i = y_i - \hat{\beta}_0 + \hat{\beta}_1x_i$. Then, since $\hat{y}_i$ and $\hat{e}_i$ are orthogonal, $\sum_{i=1}^{n}y_i = \sum_{i=1}^{n}\hat{y}_i$ and $\sum_{i=1}^{n}\hat{e}_i = 0$, a little algebra yields (on both sides of the equal sign of (7), subtract $\overline{y}$, raise to the power 2, sum over i and divide by n)

$$\begin{aligned}\hat{\sigma}_0^2 &= \hat{\sigma}_{\hat{y}_i}^2 + \hat{\sigma}^2 \\ &= \frac{\hat{\sigma}_{\hat{y}_i}^2}{\hat{\sigma}_0^2}\hat{\sigma}_0^2 + \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\hat{\sigma}_0^2 \\ &= \hat{r}^2\hat{\sigma}_0^2 + \left(1-\hat{r}^2\right)\hat{\sigma}_0^2,\end{aligned}$$

where $\hat{\sigma}_0^2$ is the sampling variance of the y_i as defined in (4b), $\hat{\sigma}_{\hat{y}_i}^2$ is the sampling variance of predictions $\hat{y}_i$ and

$$\hat{r}^2 = \frac{\hat{\sigma}_{\hat{y}_i}^2}{\hat{\sigma}_0^2}, \tag{8}$$

the ratio of the sampling variance of predictions to the sampling variance of the observations, or the estimated proportion of variance explained by the covariate. The argument in the natural logarithm of (6) can be written

$$\begin{aligned}\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} &= \frac{\hat{\sigma}_0^2}{\left(1-\hat{r}^2\right)\hat{\sigma}_0^2} \\ &= \frac{1}{1-\hat{r}^2}.\end{aligned}$$

Substituting in (6),

$$\begin{aligned}\lambda &= n\ln\left[\frac{1}{1-\hat{r}^2}\right] \\ &= -n\ln\left(1-\hat{r}^2\right) \\ &\approx n\hat{r}^2. \qquad (9)\end{aligned}$$

The estimated proportion of variance explained by the covariate/SNP depends on the estimated effect and the estimated variance of the SNP. Indeed (assume without loss of generality that the covariate in question has been column-centered, so that $\sum_{i=1}^{n} x_i = 0$),

$$\begin{aligned}\hat{\sigma}_{\hat{y}_i}^2 &= \frac{1}{n}\left[\sum_{i=1}^{n}\left(\hat{\beta}_0+\hat{\beta}_1 x_i\right)^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\left(\hat{\beta}_0+\hat{\beta}_1 x_i\right)\right)^2\right] \\ &= \hat{\beta}_1^2\frac{1}{n}\left[\sum_{i=1}^{n} x_i^2\right] \\ &= \hat{\beta}_1^2\widehat{\sigma}_{x_i}^2 \qquad (10)\end{aligned}$$

where $\widehat{\sigma}_{x_i}^2$ is the sampling variance among the n elements of column vector x. Therefore

$$\hat{r}^2 = \hat{\beta}_1^2\frac{\hat{\sigma}_{x_i}^2}{\hat{\sigma}_0^2}$$

and the test statistic (9) takes the form

$$\lambda \approx n\,\hat{\beta}_1^2\,\frac{\hat{\sigma}_{x_i}^2}{\hat{\sigma}_0^2}.$$

Sample size n, estimated SNP effect $\hat{\beta}_1$ and the frequency of the SNP in the sample $\widehat{\sigma}_{x_i}^2$, are factors affecting the test statistic. SNP's at extreme frequencies display relatively small $\hat{\sigma}_{x_i}^2$ and power of detection is compromised.

References

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. Chapman and Hall.

Sorensen, D. and D. Gianola (2002). *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer-Verlag. 740 pp., Reprinted with corrections, 2006.

Stuart, A. and J. K. Ord (1991). *Kendall's Advanced Theory of Statistics. Classical Inference and Relationship*. Edward Arnold.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics 9*, 60–62.