Homework 5

Name: Chengzhi Dong Student ID: 19331027
Email address: dongchzh3@mail2.sysu.edu.cn
GitHub link: https://github.com/Dongchengzhi/SYSU-Biotechnology/blob/main/Biostatistics/Homework/Homework5.pdf

Method of moments

In statistics, the method of moments is a method of estimation of population parameters. It starts by expressing the population moments (i.e., the expected values of powers of the random variable under consideration) as functions of the parameters of interest. Those expressions are then set equal to the sample moments. The number of such equations is the same as the number of parameters to be estimated. Those equations are then solved for the parameters of interest. The solutions are estimates of those parameters. The method of moments was introduced by Pafnuty Chebyshev in 1887 in the proof of the central limit theorem. The idea of matching empirical moments of a distribution to the population moments dates back at least to Pearson.

Suppose that the problem is to estimate k unknown parameters $\theta_1, \theta_2, \dots, \theta_k$ characterizing the distribution $f_W(w;\theta)$ of the random variable W. Suppose the first k moments of the true distribution (the "population moments") can be expressed as functions of the θ_s:

$$\begin{aligned}
\mu_1 &\equiv \mathrm{E}[W] = g_1(\theta_1, \theta_2, \dots, \theta_k), \\
\mu_2 &\equiv \mathrm{E}[W^2] = g_2(\theta_1, \theta_2, \dots, \theta_k), \\
&\vdots \\
\mu_k &\equiv \mathrm{E}[W^k] = g_k(\theta_1, \theta_2, \dots, \theta_k).
\end{aligned}$$

Suppose a sample of size n is drawn, resulting in the values$w_1, \dots, w_n$. For $j = 1, \dots, k$, let $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} w_i^j$ be the j-th sample moment, an estimate of μ_j. The method of moments estimator for $\theta_1, \theta_2, \dots, \theta_k$ denoted by $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ is defined as the solution (if there is one) to the equations:

$$\begin{aligned}
\hat{\mu}_1 &= g_1(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k), \\
\hat{\mu}_2 &= g_2(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k), \\
&\vdots \\
\hat{\mu}_k &= g_k(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k).
\end{aligned}$$

The method of moments is fairly simple and yields consistent estimators (under very weak assumptions), though these estimators are often biased.It is an alternative to the method of maximum likelihood.

However, in some cases the likelihood equations may be intractable without computers, whereas the method-of-moments estimators can be computed much more

quickly and easily. Due to easy computability, method-of-moments estimates may be used as the first approximation to the solutions of the likelihood equations, and successive improved approximations may then be found by the Newton–Raphson method. In this way the method of moments can assist in finding maximum likelihood estimates.

In some cases, infrequent with large samples but not so infrequent with small samples, the estimates given by the method of moments are outside of the parameter space (as shown in the example below); it does not make sense to rely on them then. That problem never arises in the method of maximum likelihood[citation needed]. Also, estimates by the method of moments are not necessarily sufficient statistics, i.e., they sometimes fail to take into account all relevant information in the sample.

When estimating other structural parameters (e.g., parameters of a utility function, instead of parameters of a known probability distribution), appropriate probability distributions may not be known, and moment-based estimates may be preferred to maximum likelihood estimation.

An experiment took a certain forest as the research object, which was divided into different types according to stand density, averag stand diameter, and high stand advantage. Then, the researchers used Weibull distribution to indicate its diameter structure law[1].

Ordered statistics estimation

Ordered statistic estimation refers to the estimation constructed with order statistic or its function. Given any random variables $x_1, x_2, \ldots, x_n$, the order statistics $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are also random variables, defined by sorting the values (realizations) of $x_1, x_2, \ldots, x_n$ increasing order.

When the random variables $x_1, x_2, \ldots, x_n$ form a sample they are independent and identically distributed. This is the case treated below. In general, the random variables $x_1, x_2, \ldots, x_n$ can arise by sampling from more than one population. Then they are independent, but not necessarily identically distributed, and their joint probability distribution is given by the Bapat–Beg theorem.

For a random sample as above, with cumulative distribution $F_X(x)$, the order statistics for that sample have cumulative distributions as follows (where r specifies which order statistic):

$$F_{X_{(r)}}(x) = \sum_{j=r}^{n} \binom{n}{j} [F_X(x)]^j [1 - F_X(x)]^{nj}$$

The corresponding probability density function may be derived from this result, and is found to be:

$$f_{X_{(r)}}(x) = \frac{n!}{(r-1)!\,(n-r)!} f_X(x) [F_X(x)]^{r-1} [1 - F_X(x)]^{n-r}$$

Moreover, there are two special cases, which have CDFs which are easy to compute.

$$F_{X_{(n)}}(x) = \text{Prob}(\max\{X_1, \ldots, X_n\} \leq x) = [F_X(x)]^n$$

$$F_{X_{(1)}}(x) = \text{Prob}(\min\{X_1, \ldots, X_n\} \leq x) = 1 - [1 - F_X(x)]^n$$

Which can be derived by careful consideration of probabilities.

Usually, ordered statistics and their related moments are used for analyzing data from some known lifetime distribution[2]. If the type of life distribution is known, for timed or fixed-number censored life data, statistical inference methods based on order statistics can estimate or test the relevant distribution parameters or life characteristics.

Maximum likelihood estimation

In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of an assumed probability distribution, given some observed data. This is achieved by maximizing a likelihood function so that, under the assumed statistical model, the observed data is most probable. The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimate. The logic of maximum likelihood is both intuitive and flexible, and as such the method has become a dominant means of statistical inference.

If the likelihood function is differentiable, the derivative test for determining maxima can be applied. In some cases, the first-order conditions of the likelihood function can be solved explicitly; for instance, the ordinary least squares estimator maximizes the likelihood of the linear regression model. Under most circumstances, however, numerical methods will be necessary to find the maximum of the likelihood function.

Suppose we have a random sample $x_1, x_2, \ldots, x_n$ whose assumed probability distribution depends on some unknown parameter θ. Our primary goal here will be to find a point estimator $\mathrm{u}(x_1, x_2, \ldots, x_n)$, such that $\mathrm{u}(x_1, x_2, \ldots, x_n)$ is a "good" point estimate of θ, where $x_1, x_2, \ldots, x_n$ are the observed values of the random sample. For example, if we plan to take a random sample $x_1, x_2, \ldots, x_n$ for which the x_i are assumed to be normally distributed with mean μ and variance σ2, then our goal will be to find a good estimate of μ, say, using the data $x_1, x_2, \ldots, x_n$ that we obtained from our specific random sample.

When regarded as a function of $\theta_1, \theta_2, \ldots, \theta_m$, the joint probability density (or mass) function of $X_1, X_2, \ldots, X_n$:

$$L(\theta_1, \theta_2, \ldots, \theta_m) = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2, \ldots, \theta_m)$$

is called the likelihood function.

The idea of maximum likelihood estimation is that for a given observation data x, we hope to find the parameter θ that can generate the observation data with the greatest probability from all the parameters $\theta_1, \theta_2, \ldots, \theta_m$ as the estimation result.

$$L(\theta^*|x) = p(x|\theta^*) \geq p(x|\theta)$$

The statistical principles associated with maximum likelihood underlie the logarithm of the odds (LOD) score approach in linkage analysis. Consider estimating the recombination fraction θ between two bi-allelic markers using parent–offspring trios, in which one parent is doubly heterozygous with known phase and the other is homozygous for both loci.

Maximum likelihood as a method of analysis is popular among many molecular biologists, particularly those who are more interested in models of evolution than in the actual phylogenetic pattern of taxon relationships. ML requires an explicit model of character transformation, with associated probabilities for each possible transformation from one state to another. Trees are searched in a manner similar to methods used in parsimony (e.g., branch swapping) but each tree is evaluated not by overall length, but instead by a measure of compound probability. Those trees with the highest compound probabilities (maximum likelihood) of character distribution are selected as best. Within certain theoretical frameworks, parsimony can be viewed as a particular form of maximum likelihood with reduced assumptions and infinite parameters. Under such circumstances, parsimony analyses and maximum likelihood analyses will give the same results. Under most circumstances, maximum likelihood and parsimony analyses of the same data sets have provided very similar results. However, at the present time maximum likelihood is not feasible for larger data sets due to massive computation times (at least with today's hardware and software). The computational problems are unlikely to be resolved by the mere improvement of hardware and will require advances in software that are probably not possible with current engineering capabilities in this field. However, it is possible that recent advances in algorithms for parsimony searches can be incorporated into maximum likelihood programs with similar relative levels of improvement. Because of the additional complexities of ML, even with such improvements, the ability to perform maximum likelihood analyses on the ever-larger data sets being produced with molecular techniques will lag behind parsimony for the foreseeable future[3].

Least squares method

The method of least squares is a standard approach in regression analysis to approximate the solution of overdetermined systems (sets of equations in which there are more equations than unknowns) by minimizing the sum of the squares of the residuals made in the results of every single equation.

The most important application is in data fitting. The best fit in the least-squares sense minimizes the sum of squared residuals (a residual being: the difference between an observed value, and the fitted value provided by a model). When the problem has substantial uncertainties in the independent variable (the x variable), then simple regression and least-squares methods have problems; in such cases, the methodology required for fitting errors-in-variables models may be considered instead of that for least squares.

Least-squares problems fall into two categories: linear or ordinary least squares and nonlinear least squares, depending on whether or not the residuals are

linear in all unknowns. The linear least-squares problem occurs in statistical regression analysis; it has a closed-form solution. The nonlinear problem is usually solved by iterative refinement; at each iteration the system is approximated by a linear one, and thus the core calculation is similar in both cases.

The equation for the regression line is usually expressed as $Y = a+bX$, where a is the Y intercept and b is the slope. Once we know a and b *by ordinary least squares*, we can use this equation to predict the value of Y for a given value of X.

For example, the equation for the heart rate-speed experiment is $rate=63.357+3.749\times speed$. I could use this to predict that for a speed of 10 kph, my heart rate would be 100.8 bpm. We should do this kind of prediction within the range of X values found in the original data set (interpolation). Predicting Y values outside the range of observed values (extrapolation) is sometimes interesting, but it can easily yield ridiculous results if we go far outside the observed range of X. In the frog example below, we could mathematically predict that the inter-call interval would be about 16 seconds at −40°C. The inter-calling interval would be infinity at that temperature, because all the frogs would be frozen solid.

Sometimes we want to predict X from Y. The most common use of this is constructing a standard curve. For example, we might weigh some dry protein and dissolve it in water to make solutions containing 0, 100, 200 … 1000 µg protein per ml, add some reagents that turn color in the presence of protein, then measure the light absorbance of each solution using a spectrophotometer. Then when we have a solution with an unknown concentration of protein, we add the reagents, measure the light absorbance, and estimate the concentration of protein in the solution.

Sometimes we want to predict X from Y. The most common use of this is constructing a standard curve. For example, we might weigh some dry protein and dissolve it in water to make solutions containing 0, 100, 200 … 1000 µg protein per ml, add some reagents that turn color in the presence of protein, then measure the light absorbance of each solution using a spectrophotometer. Then when we have a solution with an unknown concentration of protein, we add the reagents, measure the light absorbance, and estimate the concentration of protein in the solution.

There are two common methods to estimate X from Y. One way is to do the usual regression with X as the independent variable and Y as the dependent variable; for the protein example, we'd have protein as the independent variable and absorbance as the dependent variable. We get the usual equation, $Y = a+bX$, then rearrange it to solve for X, giving we $X=(Y-a)/b$. This is called "classical estimation."

The other method is to do linear regression with Y as the independent variable and X as the dependent variable, also known as regressing X on Y. For the protein standard curve, we would do a regression with absorbance as the X variable and protein concentration as the Y variable. We then use this regression equation to predict unknown values of X from Y. This is known as "inverse estimation."

Several simulation studies have suggested that inverse estimation gives a more accurate estimate of X than classical estimation, so that is what I recommend. However, some statisticians prefer classical estimation. If the $r2$ is high (the points are close to the regression line), the difference between classical estimation and inverse estimation is

pretty small. When we're construction a standard curve for something like protein concentration, the $r2$ is usually so high that the difference between classical and inverse estimation will be trivial. But the two methods can give quite different estimates of x when the original points were scattered around the regression line[4,5,6,7].

Reference:

[1] 孙帅超. 林分结构, 竞争与生长动态预测方法研究[D]. 西北农林科技大学, 2019.
[2] Abd-Elrahman A M. Utilizing ordered statistics in lifetime distributions production: a new lifetime distribution and applications[J]. Journal of Probability and Statistical Science, 2013, 11: 153-164.
[3] Levin S A. Encyclopedia of biodiversity[M]. 2001.
[4] Kannan, N., J.P. Keating, and R.L. Mason. 2007. A comparison of classical and inverse estimators in the calibration problem. Communications in Statistics: Theory and Methods 36: 83-95.
[5] Krutchkoff, R.G. 1967. Classical and inverse regression methods of calibration. Technometrics 9: 425-439.
[6] Krutchkoff, R.G. 1969. Classical and inverse regression methods of calibration in extrapolation. Technometrics 11: 605-608.
[7] Lwin, T., and J.S. Maritz. 1982. An analysis of the linear-calibration controversy from the perspective of compound estimation. Technometrics 24: 235-242.