Estimating Treatment Effects in the Presence of Regression to the Mean

Kazemnejad A[1], Sanagoo M.[2,], Mohebbi M.[3]

[1]*Assistant Professor, Department of Biostatistics, Tarbiat Modares University, Tehran, I.R.Iran.*

[2,3] *M.Sc. Student, Department of Biostatistics, Tarbiat Modares University, Tehran, I.R.Iran.*

ABSTRACT

The usual paired-sample t-test for data from test-retest experiment where only a subset of the population retakes the exam may not be completely adequate when the correlation between the test and retest subpopulation is weak .We propose a two-step procedure to estimate an additive regression to the mean model. First estimate the model parameters by least square method where we should assume that the population mean is known. This is not reasonable in many situations. Second use this parameter estimates as initial values to fit into maximum likelihood procedure for refinement. This method does not have the restriction of knowing the exact value of the population mean.

Key Words: Bivariate normal; Test-retest, Wald test; Likelihood ratio test;

1. INTRODUCTION

The regression to the mean problem occurs often in medical, social and other studies. Often, subjects are selected for treatment because they have high (or low) values of some baseline measurement of interest. A review of previous approaches and a discussion of techniques of parameters estimation from a model which is assumed to have a bivariat normal distribution was provided by James (1973) that proposed a model with a multiplicative treatment effect and obtained moment estimators. Senn and Brown (1985) considered the same model and used maximum likelihood estimation of parameters. Chen and Cox (1992) studied this model in large screening program and propose an approach that is asymptotically equivalent to the maximum likelihood method. Chen, Cox and Cui (1998) again in screening in program used a stratified regression to the mean model including both additive and multiplicative treatment effects. As an alternative to the multiplicative model, Mee and Chua (1991) considered an additive model for the regression to the mean problem.

2.THE MODEL AND INFERENCE

The approach of using maximum likelihood does not require knowledge of the population mean. It could also be taken to a generalization of Mee and Chua's problem because deals with the bivariate normal.

Assume that the value x (the first measurement) and y (the second measurement) follow a bivariate normal distribution. When there is no treatment effect, the two variates have the same mean (μ), standard deviation (σ), and correlation coefficient (ρ). We often expect $\rho > 0$, although this is not necessary. In the presence of a treatment effect, the marginal distribution of x remains unchanged and, for given x, the conditional distribution of y is specified by regression model:

$$y = \mu + \tau + \rho(x - \mu) + \varepsilon \quad (2.1)$$

Where ε is distributed normally and independently of x with mean 0 and standard deviation $\sigma\left(1-\rho^2\right)^{\frac{1}{2}}$, and $\tau > 0$ is treatment parameter.

Cohen (1955) and Senn and Brown (1985) described four kinds of samples from the population of baseline values: truncated, censored, selected, and complete sample. For a truncated sample, the sampling procedure is continued until n specimens for certain rang of x (e. g, $x \le k$ or $k_1 \le x \le k_2$) are recorded. It is not possible to observe the eliminated x values. Corresponded y measurements are made on the n chosen individuals. In the case of a censored sample, both x and y are recorded if x is in the interval. However, account is kept of censored specimens for which x is not in the interval, although neither the x nor the y values of such specimens are recorded. For selected samples, full measurement is made and recorded for all the values of x whether it is in the interval or not. However, only the y values corresponded to the x in the interval are recorded. Finally a complete sample is one in which all the values of x and y is recorded whether they are in the interval or not. This definition is more general than the Senn and Brown's expression.

For convenience of notation, we assume that the first n subjects are chosen to receive treatment among a total of n+m specimens in the sample. Therefore we will have such a data set:

$$(x_1, y_2), \ldots (x_n, y_n), x_{n+1}, \ldots x_{n+m}$$

Our interest is in τ and ρ, and we treat μ and σ as nuisance parameters in (2.1). Following Cohen (1955), the likelihood function is

$$L(\tau,\rho,\mu,\sigma) = G \times \frac{\exp\left[-\frac{1}{2}\sum_{1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}\right]}{(2\pi\sigma^2)^{\frac{n}{2}}} \times \frac{\exp\left[\frac{-\frac{1}{2}\sum_{1}^{n}[(y_i-\mu)-\tau-\rho(x_i-\mu)]^2}{\sigma^2(1-\rho^2)}\right]}{(2\pi\sigma^2(1-\rho^2))^{\frac{n}{2}}} \qquad (2.2)$$

Where G, is a restriction function; depending on the type of sample.

The following restriction functions are appropriate

Truncated sample

$$G = \{\frac{1}{[1-\phi(z)]}\}^n \qquad (2.3)$$

Censored sample

$$G = [\phi(z)]^m \qquad (2.4)$$

Selected sample

$$G = \left[\frac{1}{2\pi\sigma^2}\right]^{\frac{m}{2}} \exp\left[\frac{-\frac{1}{2}\sum_{n+1}^{n+m}(x_i-\mu)^2}{\sigma^2}\right] \qquad (2.5)$$

And complete sample

$$G = \left[\frac{1}{2\pi\sigma^2}\right]^{\frac{m}{2}} \exp\left[\frac{-\frac{1}{2}\sum_{n+1}^{n+m}(x_i-\mu)^2}{\sigma^2}\right] \times \frac{1}{(2\pi\sigma^2(1-\rho^2))^{\frac{m}{2}}} \exp\{\frac{-\frac{1}{2}\sum_{n+1}^{n+m}[y_i-\rho x_i+(1-\rho)\mu]^2}{\sigma^2(1-\rho^2)}\} \qquad (2.6)$$

Where $z = \frac{k-\mu}{\sigma}$ and $\phi(z)$ is the standard normal distribution function. The score vector and empirical information matrix can be derived by differentiating natural log of (2.2).

The maximum likelihood estimates are the solutions for the score equation system of any given type of problem by standard numerical methods. The estimated variance-covariance matrix of $\hat{\beta} = (\hat{\mu}, \hat{\sigma}^2, \hat{\rho}, \hat{\tau})$ (2.7) can be obtained by substituting the estimated parameters into the inverse of the empirical information matrix.

3. DISCUSSION

We first checked whether the distribution of x and y values were approximately normal. In fact the distributions were slightly skewed to the left and the distribution of natural logarithmically transformed values was more nearly normal. Therefore we based the analysis on the log-transformed data. Since $\beta_1 = \rho \leq 1$, in the null model departures from null hypothesis or presence of any positive intervention cause $\beta_1 > 1$. There are two possible actions. First impose the constraint $\rho \leq 1$. Least squares estimation with this constraint is not difficult, Mee and Chau's (1991) maximum likelihood estimation can be obtained by this constrain. But it clearly reduces the power of test. Second the alternative hypothesis could be $H_a : y_2 = \mu + \tau + \beta_1 x + \varepsilon$ with $\tau > 0$ and p.

A model which includes both the additive and multiplacative treatment effect without considering assumption in nuisance parameters (e .g, μ and σ) seems to be more flexible. Any method based on maximum likelihood is likely to be sensitive to departure from assumptions.

REFRENCES

Chen, S. and Cox, C. (1992). "Use of baseline data for estimation of treatment effects in the presence of regression to the mean," Biometrics, 48, 593 – 598 .

Chen, S. and Cox, C. and Cui, L. (1998) "A more flexible regression-to- the mean model with possible stratification," Biometrics, 54, 939 – 947.

Cohen, A. C. (1955). "Restriction and selection in samples from bivariate normal distributions," Journal of the American Statistical Association, 50, 884 – 893.

James, K. E. (1973). "Regression Toward the mean in uncontrolled clinical studies," Biometrics, 29 , 121 – 130.

Mee, R. W., Chua, T. C. (1991), "Regression towards the mean and the paired sample t-test", The American Statistician, 45, 39-42.

Seen, S. J. and Brown, R. A. (1985). "Estimating treatment effects in clinical trials subject to regression to the mean," Biometrics, 41, 555 – 560 .