Semiparametric Analysis of Covariance

Michael G. Schimek
Karl-Franzens-University Graz, Institute for Medical Informatics, Statistics & Documentation
Engelgasse 13, A-8010 Graz, Austria, Europe
michael.schimek@uni-graz.at

1. Introduction

Analysis of covariance (ANCOVA) is a common technique for comparing the values of a response variable across several factor levels in the presence of a covariate effect. The idea is to contrast factor levels while adjusting for the association between response and covariate. This is different from analysis of variance where mean responses in the groups (factor levels) are assumed constant but possibly different. ANCOVA is very useful for improving power when testing factor effects. Classical applications are error variance reduction in observational studies or field trials. The general ANCOVA model in k factor levels can be written

$$y_{ij} = m_i(x_{ij}) + \epsilon_{ij}; \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_j.$$

For a simple parametric (e.g. linear) relationship between response and covariate, $m_i(x) = \alpha_i + \beta_i x$ under the basic assumptions of additivity, random errors $\epsilon \sim N(0, \sigma^2 \mathbf{I})$, and a balanced design ($n_1 = n_2 = \ldots = n_k$). (Note that a random assignment of treatments is not necessary.) For the purpose of more flexibility several non- and semiparametric alternatives have been recently proposed.

2. A semiparametric approach

Our approach replaces the parametric relationship by an arbitrary smooth one, fitted by cubic smoothing splines. Moreover we drop the normality assumption. Potential factor level effects are tested in a parametric manner. The latter two features make the procedure semiparametric. The emphasis is on curve estimation and improved smoothing parameter choice for $\hat{\lambda}$, most critical when testing is involved. The covariate values (design points) can be different for each level. The model equation is $m_i(x_{ij}) = u_{ij}\gamma_i + f(x_{ij})$ with 0/1-elements u_{ij} forming a $n \times k$ design matrix $\mathbf{U}$ ($n = n_1 + n_2 + \ldots + n_k$), an unknown parameter γ_i, an unknown smooth curve f, and homoscedastic independent zero mean errors ϵ. The design matrix characterizes k factor levels (treatment conditions).

We apply asymptotically unbiased partial spline estimation. This semiparametric approach is related to that of Speckman (1988). Because we fit an overall spline curve it is different from Kulasekera (1995) who calculates kernel estimates for each factor level and then

compares the curves via nonparametric tests. Other related work is due to Young and Bowman (1995). We apply their idea of significance trace to control the relationship between λ and the p-value in our test situation.

For the choice of the smoothing parameter λ an unbiased risk (UBR) estimator is used. It requires the consistent estimation of the error variance which is calculated based on pseudo-residuals (Eubank et al., 1998). It can be shown that the coefficient vector γ is asymptotically NID without distributional assumption on the covariate respectively errors (Speckman, 1988). We evaluate the variance-covariance matrix for γ to obtain standard errors SE of the estimated coefficients $\hat{\gamma}_i$ (factor levels). The easiest way to perform inference is an approximate standard normal test $z \approx \hat{\gamma}_i / SE(\hat{\gamma}_i)$. This test is used in our simulations to study the power of our semiparametric approach.

3. Simulations

A simulation study is carried out for $n = 100$ in a one-way ANCOVA model. For the relationship between covariate x and response y we examine several nonparametric functions, such as $y = \phi(x^2 - x + 0.15)$ and $y = \phi \cos(\pi x)$, where ϕ controls the signal-to-noise ratio, with standard error one. The covariate is constructed in two ways, (i) $x_j = (j - 0.5)/n$ on $[0, 1]$ (equally spaced) and (ii) x_j ordered values from Un(0,1) (not equally spaced) for $j = 1, 2, \ldots, n$. The design vector u consists of rounded Un(0,1)-generated data (balanced design $n_1 = n_2$ in expectation). γ-values are 0.5, 1.5 and 2.5. A range of significance levels is studied and the power is calculated over 500 replications.

The main outcome is that the power of the semiparametric approach is satisfying in almost all settings as long as λ is properly specified (significance trace as reference). The data-driven ("automatic") UBR criterion turns out to yield quite adequate $\hat{\lambda}$ for testing.

REFERENCES

Eubank, R.L., Kambour, E.L., Kim, J.T., Klipple, K., Reese, C.S. and Schimek, M. (1998) Estimation in partially linear models. *CSDA*, **29**, 27-34.

Kulasekera, K.B. (1995). Comparison of regression curves using quasi-residuals. *JASA*, **90**, 1085-1093.

Speckman, P. (1988). Kernel smoothing in partial linear models. *J. R. Statist. Soc. B*, **50**, 413-436.

Young, S.G. and Bowman, A.W. (1995). Non-parametric analysis of covariance. *Biometrics*, **51**, 920-931.

RESUME

Un modèle d'analyse de covariance semi-paramétrique basé sur des splines cubiques de lissage est présenté. Une nouvelle méthode pour le choix du paramètre de lissage entraine une amélioration du test des effets du facteur. Ceci est confirmé par les simulations.