Non-parametric analysis of covariance

Holger Dette
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum, Germany
holger.dette@ruhr-uni-bochum.de

Natalie Neumeyer
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum, Germany
natalie.neumeyer@ruhr-uni-bochum.de

1. Introduction

An important problem in applied regression analysis is the comparison of a response Y across several groups in the presence of a covariate effect. In general this model can be written as

$$Y_{ij} = g_i(t_{ij}) + \sigma_i(t_{ij})\varepsilon_{ij} \tag{1.1}$$

$(i = 1, \ldots, k; j = 1, \ldots, n_i)$, where ε_{ij} are independently identically distributed errors, g_i, σ_i are the regression and variance function in the ith group $(i = 1, \ldots, k)$ and the covariate t_{ij} varies in the interval $[0, 1]$. In this paper we are interested in the problem of testing the equality of the mean functions, i.e.

$$H_0 : g_1 = g_2 = \ldots = g_k \quad \text{versus} \quad H_1 : g_i \neq g_j \quad (\exists\, i, j \in \{1, \ldots, k\}). \tag{1.2}$$

Much effort has been devoted to this problem in the recent literature [see e.g. Hall and Hart (1990), Delgado (1993), Young and Bowman (1995), Kulasekera (1995)]. In this paper we introduce a new test for the hypothesis (**??**) which is directly applicable in the general model (**??**), does not require any additional assumptions (as homoscedasticity or equal design points).

2. The test statistic and its asymptotic distribution

Let $N = \sum_{i=1}^{k} n_i$ denote the total sample size and assume $\frac{n_i}{N} = \kappa_i + O(\frac{1}{N})$, $i = 1, \ldots, k$ for given constants $\kappa_1, \ldots \kappa_k \in (0, 1)$. Let $r_1, \ldots, r_k$ denote positive densities on the interval $[0, 1]$ such that the design poins t_{ij} satisfy $\int_0^{t_{ij}} r_i(t)dt = \frac{j}{n_i}$; $j = 1, \ldots, n_i$, $i = 1, \ldots, k$. Throughout this paper we will assume that the design densities and the regression functions are sufficiently smooth, i.e. $g_j, r_j \in C^{(r)}[0, 1] \qquad j = 1, \ldots, k$ where $r \geq 2$ and $C^{(r)}[0, 1]$ denotes the space

of r-times continuously differentiable functions on the interval $[0,1]$. The integrated variance function $\int_0^1 \sigma_i^2(t) r_i(t) dt$ of each sample can be estimated by [see Hall and Marron (1990)]

$$\hat{\sigma}_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (Y_{ij} - \hat{g}_i(t_{ij}))^2 \quad (i = 1, \ldots, k) \tag{2.1}$$

where $\hat{g}_i$ is the Nadaraya-Watson estimator of the ith regression function g_i from the ith sample and h the corresponding bandwidth [for the sake of brevity we assume equal banwidths in all samples]. We assume that the kernel is supported on a compact interval, say $[-1,1]$, and of order $r \geq 2$. where $\int_{-1}^1 K^2(u) du < \infty$ and let $K * K$ denote the convolution of the kernel with itself. The bandwidth should satisfy

$$n_i h^2 \to \infty; \quad h_i = O(n_i^{-2/(4r+1)}) \quad (i = 1, \ldots, k). \tag{2.2}$$

If the hypothesis of equal variances is valid, the total sample could be used to estimate the common regression and we consider the analogue of (**??**)

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \hat{g}(t_{ij}))^2 \,, \tag{2.3}$$

where $\hat{g}$ is the Nadaraya-Watson estimator of the total sample. It can be shown [see Dette and Neumeyer (1999)] that under the hypothesis of equal regression curves this is essentially an estimator of an integrated convex combination of the individual variance functions. The test statistic for the hypothesis (**??**) is defined by

$$T_N = \hat{\sigma}^2 - \frac{1}{N} \sum_{i=1}^{k} n_i \hat{\sigma}_i^2.$$

and its asymptotic properties are listed in the following theorem [for a proof see Dette and Neumeyer (1999)]

Theorem *If the hypothesis of equal regression functions is valid, then the statistic T_N satisfies*

$$N\sqrt{h}(T_N - B_k h^{2r} - \frac{1}{Nh} D_k) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \beta_k^2)$$

where

$$B_k = k_r^2 \int_0^1 ((g\bar{R})^{(r)} - g\bar{R}^{(r)})^2(t) \frac{dt}{\bar{R}(t)} - k_r^2 \sum_{j=1}^{k} \kappa_j \int_0^1 ((g r_j)^{(r)} - g r_j^{(r)})^2(t) \frac{dt}{r_j(t)} \tag{2.4}$$

$$D_k = \Big[\int_{-1}^1 K^2(u)\, du - 2K(0) \Big] \sum_{j=1}^{k} \left(\int_0^1 \frac{\kappa_j \sigma_j^2(t) r_j(t)}{\bar{R}(t)}\, dt - \int_0^1 \sigma_j^2(t)\, dt \right), \tag{2.5}$$

the asymptotic variance is given by

$$\begin{aligned} \beta_k^2 = {} & 2 \int_{-1}^1 (2K - K * K)^2(u)\, du \Big\{ \sum_{j=1}^{k} \int_0^1 \sigma_j^4(t) \Big(\frac{\kappa_j r_j(t)}{\bar{R}(t)} - 1 \Big)^2 dt \\ & + \sum_{j=1}^{k} \sum_{\substack{l=1 \\ l \neq j}}^{k} \int_0^1 \sigma_j^2(t) \sigma_l^2(t) \frac{\kappa_j r_j(t) \kappa_l r_l(t)}{\bar{R}^2(t)}\, dt \Big\} \end{aligned}$$

and $\bar{R}(t) = \sum_{j=1}^{k} \kappa_j r_j(t)$ *denotes the convex combination of the underlying densities,.*
Under the alternative $g_i \neq g_j (\exists\ i, j \in \{1, \ldots, k\})$ *the statistic* T_N *satisfies*

$$\sqrt{N}(T_N - M_k^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \gamma_k^2)$$

where

$$M_k^2 = \sum_{j=1}^{k} \sum_{\substack{l=1 \\ l<j}}^{k} \int_0^1 (g_j - g_l)^2(t) \frac{\kappa_j r_j(t) \kappa_l r_l(t)}{\bar{R}(t)}\, dt. \tag{2.6}$$

and the asymptotic variance is given by

$$\gamma_k^2 = 4 \sum_{j=1}^{k} \int_0^1 \left(\sum_{\substack{l=1 \\ l \neq j}}^{k} (g_j(t) - g_l(t)) \frac{\kappa_l r_l(t)}{\bar{R}(t)} \right)^2 \sigma_j^2(t) \kappa_j r_j(t)\, dt. \tag{2.7}$$

3. General remarks and a small simulation study

Remark 1. As a consequence of the above theorem we obtain a consistent, asymptotic level α test by rejecting the hypothesis of equal regression curves whenever

$$N\sqrt{h}(T_N - B_k h^{2r} - \frac{D_k}{Nh}) > \frac{\Phi(1-\alpha)}{\beta_k} \tag{3.1}$$

where B_k, D_k and β have to be replaced by consistent estimators. We will illustrate the performance of a wild bootstrap version of this test at the end of this section.

Remark 2. It should be pointed out that the result in Section 2 does not depend on the special structures of the smoothing procedures used in the construction of the variance estimators. For example, a local polynomial estimator of odd order produces a different bias of the test statistic and the kernel in (**??**) has to replaced by the equivalent higher order kernel corresponding to the local polynomial estimator [see Wand and Jones (1995)]. Although local polynomial estimators have various advantages for the estimation of the regression function, our simulation results showed that this superiority is not reflected in the problem of testing the equality of regression functions using the statistic T_N. A heuristical explanation of this observation is that our approach avoids the direct estimation of the regression function and only uses estimates for quantities smoothed by a linear integral operator.

Example. We investigate the approximation of the level by a bootstrap version of the test (**??**) with uniform designs, quadratic regression functions, standard normal distributed errors and and two samples with size $n_1, n_2 = 10, 20, 30, 50$. The results are summarized in Table 1 which shows the simulated rejection probabilities of a wild bootstrap test with level 10% , 5% and 2.5%. For further simulation results we refer to Dette and Neumeyer (1999).

(n_1, n_2)	(10, 10)	(10,20)	(10,30)	(10, 50)	(20,20)
$\alpha = 10\%$	0.098	0.114	0.107	0.092	0.106
$\alpha = 5\%$	0.054	0.056	0.055	0.052	0.048
$\alpha = 2.5\%$	0.032	0.030	0.028	0.028	0.026
(n_1, n_2)	(20,30)	(20,50)	(30,30)	(30,50)	(50,50)
$\alpha = 10\%$	0.106	0.096	0.102	0.096	0.096
$\alpha = 5\%$	0.048	0.048	0.052	0.055	0.047
$\alpha = 2.5\%$	0.026	0.026	0.031	0.023	0.030

Table 1. *Simulated level of the test (??) for various sample sizes and standard normal errors.*

REFERENCES

H. Dette, N. Neumeyer (1999). Nonparametric analysis of covariance. Technical Report 262, Department of Mathematics, Ruhr-Universität Bochum. http://www.ruhr-uni-bochum.de/mathematik3/ preprint.htm).

P. Hall, J.S. Marron (1990). On variance estimation in nonparametric regression. Biometrika 77, 415-419.

P. Hall, J.D. Hart (1990). Bootstrap test for difference between means in nonparametric regression. J. Amer. Statist. Assoc. 85, 1039-1049.

E.C. King, J.D. Hart, T.E. Wherly (1991). Testing the equality of regression curves using linear smoothers. Statist. Probab. Lett. 12, 239-247.

K.B. Kulasekera (1995). Comparison of regression curves using quasi residuals. J. Amer. Stat. Assoc. 90, 1085-1093.

M.P. Wand, M.C. Jones (1995). Kernel Smoothing. Chapman and Hall, London.

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

RESUME

We consider the problem of testing the equality of k regression curves using independent samples. A new test is proposed, which is based on a linear combination of estimators for the integrated variance function in the individual samples and in the combined sample. Asymptotic normality of the introduced statistic is proved under the hypothesis of equality and under fixed alternatives. In contrast to most of the procedures proposed in the literature the method introduced in this paper is also applicable in the case of different design points in each sample and heteroscedastic errors.