Tests for the equality of conditional variance functions in nonparametric regression

Juan Carlos Pardo-Fernández*
Universidade de Vigo, Vigo, Spain - juancp@uvigo.es

María Dolores Jiménez-Gamero
Universidad de Sevilla, Sevilla, Spain - dolores@us.es

Anouar El Ghouch
Université catholique de Louvain, Louvain-la-Neuve, Belgium - anouar.elghouch@uclouvain.be

Abstract

In this piece of research we are interested in checking whether the conditional variances are equal in two or more location-scale regression models. Our procedure is fully nonparametric and is based on the comparison of the error distributions under the null hypothesis of equality of variances and without making use of this null hypothesis. We propose four test statistic based on empirical distribution functions (Kolmogorov-Smirnov and Cramér-von Mises type test statistics) and two test statistics based on empirical characteristic functions. The limiting distributions of these six test statistics are established under the null hypothesis and under local alternatives. We show how to approximate the critical values using either an estimated version of the asymptotic null distribution or a bootstrap procedure. Simulation studies are conducted to assess the finite sample performance of the proposed tests. We also apply our tests to data on household expenditures.

Keywords: empirical characteristic function; empirical distribution function; kernel smoothing; regression residuals.

1. Introduction

When comparing $k \geq 2$ populations it is interesting not only comparing the means, but also other characteristics like the variances. For example, in quality control, it is important to check the uniformity and the stability of the production process under different experimental and practical conditions. In biomedical research, detecting variation in gene expression levels is important for many reasons, for example, to identify experimental and environmental factors that affect a biological process. Equality of variances, when satisfied, can also be used to develop more powerful and simple ANOVA-type test statistics. Without controlling for the effect of covariates, there are a substantial number of tests available in the literature for the equality of (unconditional) variances from two or more populations. The standard procedures include the classical F-test and Levene's test (Levene, 1960) which is known to be more robust to the violation of normality; see Gastwirth *et al.* (2009) for a recent review and some interesting examples and applications. In this paper, we are interested in the comparison of conditional variances.
We assume that in each population, along with the variable of interest or response variable, Y, it is also observed another variable, X, the covariate, so that the mean and the variance of the response variable depend on the values of X. More specifically, let (X_j, Y_j), $1 \leq j \leq k$, be k independent random vectors satisfying general nonparametric regression models

$$Y_j = m_j(X_j) + \sigma_j(X_j)\varepsilon_j, \tag{1}$$

where $m_j(x) = E(Y_j \mid X_j = x)$ is the regression function, $\sigma_j^2(x) = Var(Y_j \mid X_j = x)$ is the conditional variance function and ε_j is the regression error, which is assumed to be independent of X_j. Note that, by construction, $E(\varepsilon_j) = 0$ and $Var(\varepsilon_j)$=1. The covariate X_j is continuous with density function f_j. Since the objective is to compare the variance functions, it is reasonable to assume that the covariates have common

support, say R. The regression functions, the variance functions, the distribution of the errors and the distribution of the covariates are completely unknown and no parametric models are assumed for them. Thus, our approach is completely nonparametric. In this conditional setting, the hypothesis of equality of variances is stated in terms of the conditional variance functions: $H_0 : \sigma_1^2(x) = \sigma_2^2(x) = \cdots = \sigma_k^2(x)$, for all $x \in R$. Or equivalently,

$$H_0 : \sigma_j(x)/\sigma_0(x) = 1, \quad \text{for } 1 \leq j \leq k,$$

where $\sigma_0^2(x)$ is the common variance that can be expressed as $\sigma_0^2(x) = \sum_{j=1}^k \pi_j(x)\sigma_j^2(x)$, for some positive functions $\pi_1, \ldots, \pi_k$ satisfying $\sum_{j=1}^k \pi_j(x) = 1$. The alternative hypothesis is

$$H_1 : \sigma_j(x)/\sigma_0(x) \neq 1, \quad \text{for some } j \in \{1 \ldots, k\}.$$

We will develop several test statistics and study their distribution under H_0 and under local alternatives converging to the null hypothesis at the rate $n^{-1/2}$, n being the sample size. Specifically, we consider the following local alternative hypothesis

$$H_{1,n} : \sigma_j(x)/\sigma_0(x) = 1 + n^{-1/2}\delta_j(x), \quad \text{for } 1 \leq j \leq k,$$

for some functions δ_j. To be more precise, in the previous expression we should have written $\sigma_{n,j}(x)/\sigma_{n,0}(x)$ instead of $\sigma_j(x)/\sigma_0(x)$, as this function depends on n. However, to short the notation we suppress this explicit dependence on n. Observe that as n increases $H_{1,n}$ becomes closer and closer to H_0. Also, when $\delta_j(x) = 0$, for $1 \leq j \leq k$, $H_{1,n}$ reduces to H_0.

Statistical literature concerning the problem of testing for common features in several regression models has mainly focused on testing for common regression curves or testing for common error distributions. The problem of testing for the equality of regression curves in nonparametric settings has been extensively treated; see for example Delgado (1993), Kulasekera (1995), Neumeyer and Dette (2003), Pardo-Fernández *et al.* (2007, 2014a), Srihera and Stute (2010) and González-Manteiga and Crujeiras (2013) for a recent review. On the other hand, testing for the equality of error distributions has been addressed in Pardo-Fernández (2007). To the best of our knowledge, the comparison of conditional variance functions has not been studied before. Most papers dealing with testing on the conditional variance function focus on studying the hypothesis of homoscedasticity (see for example Liero, 2003, or Dette and Marchlewski, 2010, and the references therein), or more in general, if the conditional variance function follows some fixed parametric form (see for example Dette *et al.*, 2007, or Koul and Song, 2010, and the references therein).

In order to construct a test for testing H_0, several approaches are possible. Here we follow the ideas in Pardo-Fernández *et al.* (2007, 2014a) for testing the equality of the regression functions, which consist of comparing the distribution of the errors of the regression models. Specifically, let

$$\varepsilon_j = \frac{Y_j - m_j(X_j)}{\sigma_j(X_j)}, \tag{2}$$

be the regression error in population j, $1 \leq j \leq k$. Define

$$\varepsilon_{0j} = \frac{Y_j - m_j(X_j)}{\sigma_0(X_j)} = \varepsilon_j \frac{\sigma_j(X_j)}{\sigma_0(X_j)} \tag{3}$$

to be the error under the null hypothesis, $1 \leq j \leq k$. Let $F_{\varepsilon_j}(t) = P(\varepsilon_j \leq t)$ and $F_{\varepsilon_{0j}}(t) = P(\varepsilon_{0j} \leq t)$ be the cumulative distribution function (CDF) of ε_j and ε_{0j}, respectively. The following theorem shows that H_0 is true if and only if the distributions of ε_j and ε_{0j} coincide.

Theorem 1. *Assume that σ_j is a continuous function in R and $0 < E(\varepsilon_j^4) < \infty$, $1 \leq j \leq k$.*

(a) *H_0 is true if and only if the random variables ε_j and ε_{0j} have the same distribution for all $1 \leq j \leq k$.*

(b) *Let $p_1, \ldots, p_k$ be such that $p_j > 0$, $1 \leq j \leq k$, and $\sum_{j=1}^k p_k = 1$. Let $F_\varepsilon(t) = \sum_{j=1}^k p_j F_{\varepsilon_j}(t)$ and $F_{\varepsilon_0}(t) = \sum_{j=1}^k p_j F_{\varepsilon_{0j}}(t)$. Assume also that $E(\varepsilon_1^4) = \cdots = E(\varepsilon_k^4)$. Then H_0 is true if and only if $F_\varepsilon(t) = F_{\varepsilon_0}(t)$, for all t.*

The assertions in the previous result can be interpreted in terms of the CDF or in terms of any other function characterizing a probability law, such as the characteristic function (CF). In this paper we will consider both cases, that is, to test H_0 we will compare consistent estimators of the CDFs and CFs of the random variables ε_j and ε_{0j}, $1 \leq j \leq k$.

2. The test statistics

Let (X_j, Y_j), $1 \leq j \leq k$, be k independent random vectors satisfying general nonparametric regression models (1). For $1 \leq j \leq k$, let ε_j and ε_{0j} be as defined in (2) and (3), respectively. As justified in Theorem 1, to test for H_0 we will compare consistent estimators of the CDFs and CFs of the random variables ε_j and ε_{0j}, $1 \leq j \leq k$, and also consistent estimators of the CDFs F and F_0 and of their associated CFs. Since neither ε_j nor ε_{0j} are observable, the inference must be based on the estimated residuals. Next we construct them. Let (X_{jl}, Y_{jl}), $1 \leq l \leq n_j$, be independent and identically distributed (iid) observations from (X_j, Y_j), $1 \leq j \leq k$ and let $n = \sum_{j=1}^{k} n_j$. Along the paper it will be assumed that $n_j/n \to p_j > 0$, $1 \leq j \leq k$. In order to estimate the errors, we first need to estimate the regression functions, $m_j(x) = E(Y_j|X_j = x)$, the variance functions, $\sigma_j^2(x) = E[\{Y_j - m_j(x)\}^2|X_j = x]$, and the common variance function under H_0, $\sigma_0^2(x)$. With this aim we use nonparametric estimators based on kernel smoothing techniques. We use the following estimators for the functions m_j, σ_j^2 and σ_0^2:

$$\hat{m}_j(x) = \sum_{l=1}^{n_j} w_{jl}(x) Y_{jl}, \quad \hat{\sigma}_j^2(x) = \sum_{l=1}^{n_j} w_{jl}(x) Y_{jl}^2 - \hat{m}_j^2(x), \quad \hat{\sigma}_0^2(x) = \sum_{j=1}^{k} \pi_j(x) \hat{\sigma}_j^2(x),$$

where the quantities w_{jl} are, either the local-linear weights or the Nadaraya-Watson weights (for details, see for example Fan and Gijbels, 1996). Based on these estimators, for each population j, $1 \leq j \leq k$, we construct two samples of residuals,

$$\hat{\varepsilon}_{jl} = \frac{Y_{jl} - \hat{m}_j(X_{jl})}{\hat{\sigma}_j(X_{jl})} \qquad \text{and} \qquad \hat{\varepsilon}_{0jl} = \frac{Y_{jl} - \hat{m}_j(X_{jl})}{\hat{\sigma}_0(X_{jl})},$$

$1 \leq l \leq n_j$. Then we can construct the corresponding empirical CDFs (ECDFs),

$$\hat{F}_{\varepsilon_j}(t) = \frac{1}{n_j} \sum_{l=1}^{n_j} I(\hat{\varepsilon}_{jl} \leq t) \qquad \text{and} \qquad \hat{F}_{\varepsilon_{0j}}(t) = \frac{1}{n_j} \sum_{l=1}^{n_j} I(\hat{\varepsilon}_{0jl} \leq t),$$

and empirical CFs (ECFs),

$$\hat{\varphi}_{\varepsilon_j}(t) = \frac{1}{n_j} \sum_{l=1}^{n_j} \exp(\mathrm{i}t\hat{\varepsilon}_{jl}) \qquad \text{and} \qquad \hat{\varphi}_{\varepsilon_{0j}}(t) = \frac{1}{n_j} \sum_{l=1}^{n_j} \exp(\mathrm{i}t\hat{\varepsilon}_{0jl}),$$

respectively. These ECDFs are consistent kernel based nonparametric estimators of the population CDFs $F_{\varepsilon_j}(t)$ and $F_{\varepsilon_{0j}}(t)$, respectively. Analogously, the above ECFs are consistent kernel based nonparametric estimators of the population CFs $\varphi_{\varepsilon_j}(t) = E\{\exp(\mathrm{i}t\varepsilon_j)\}$ and $\varphi_{\varepsilon_{0j}}(t) = E\{\exp(\mathrm{i}t\varepsilon_{0j})\}$, respectively. We can also consider the following ECDFs

$$\hat{F}_{\varepsilon}(t) = \frac{1}{n} \sum_{j=1}^{k} \sum_{l=1}^{n_j} I(\hat{\varepsilon}_{jl} \leq t) \qquad \text{and} \qquad \hat{F}_{\varepsilon_0}(t) = \frac{1}{n} \sum_{j=1}^{k} \sum_{l=1}^{n_j} I(\hat{\varepsilon}_{0jl} \leq t),$$

and ECFs,

$$\hat{\varphi}_{\varepsilon}(t) = \frac{1}{n} \sum_{j=1}^{k} \sum_{l=1}^{n_j} \exp(\mathrm{i}t\hat{\varepsilon}_{jl}) \qquad \text{and} \qquad \hat{\varphi}_{\varepsilon_0}(t) = \frac{1}{n} \sum_{j=1}^{k} \sum_{l=1}^{n_j} \exp(\mathrm{i}t\hat{\varepsilon}_{0jl}),$$

which estimate $F_{\varepsilon}(t) = \sum_{j=1}^{k} p_j F_{\varepsilon_j}(t)$, $F_{\varepsilon_0}(t) = \sum_{j=1}^{k} p_j F_{\varepsilon_{0j}}(t)$, $\varphi_{\varepsilon}(t) = \sum_{j=1}^{k} p_j \varphi_{\varepsilon_j}(t)$ and $\varphi_{\varepsilon_0}(t) = \sum_{j=1}^{k} p_j \varphi_{\varepsilon_{0j}}(t)$, respectively.

To test for H_0, we will construct Kolmogorov-Smirnov type statistics and Cramér-von Mises type statistics to compare the ECDFs, and weighted L_2-distances to compare the ECFs. More precisely, the considered statistics are

$$T_{KS}^1 = \sum_{j=1}^{k} \sqrt{n_j} \sup_t |\hat{F}_{\varepsilon_j}(t) - \hat{F}_{\varepsilon_{0j}}(t)|, \quad T_{CM}^1 = \sum_{j=1}^{k} n_j \int \{\hat{F}_{\varepsilon_j}(t) - \hat{F}_{\varepsilon_{0j}}(t)\}^2 d\hat{F}_{\varepsilon_{0j}}(t)$$

$$T_{KS}^2 = \sqrt{n} \sup_t |\hat{F}_{\varepsilon}(t) - \hat{F}_{\varepsilon_0}(t)|, \quad T_{CM}^2 = n \int \{\hat{F}_{\varepsilon}(t) - \hat{F}_{\varepsilon_0}(t)\}^2 d\hat{F}_0(t),$$

$$T_1 = \sum_{j=1}^{k} n_j \int \left|\hat{\varphi}_{\varepsilon_j}(t) - \hat{\varphi}_{\varepsilon_{0j}}(t)\right|^2 w(t)dt, \quad T_2 = n \int |\hat{\varphi}_{\varepsilon}(t) - \hat{\varphi}_{\varepsilon_0}(t)|^2 w(t)dt,$$

where w is a positive weight function that is needed to guarantee consistency. Note that in the case of T_1 and T_2, $|\cdot|$ represents the modulus of a complex number.

3. Summary of asymptotic results and conclusions

Due to the lack of space, we will not explain here in detail all the results obtained in this piece of research. The complete work can be found in Pardo-Fernández *et al.* (2014b). Below is a brief description of the results that will be explained in the presentation:

- Under certain regularity conditions, we have done a detailed study of the asymptotic distribution of the test statistics under the null hypothesis H_0 and under the local alternative $H_{1,n}$. The tests are consistent against any fixed alternative and are able to detect contiguous alternatives converging to the null at a rate $n^{-1/2}$. The assumptions needed to derive these properties are weaker for the ECF-based tests (specifically, no requirement is imposed on the distributions of the errors).
- We have proposed and studied approximations of the asymptotic null distribution the test statistics based on the ECF in order to obtain critical values for the test.
- We have studied the practical behaviour of the proposed test statistics by means of simulations. The simulation study includes the analysis of the tests based on critical values obtained from the asymptotic null distribution of the test statistics and from a bootstrap approximation.
- We have applied the proposed tests to a set of econometric data concerning households expenditures.

References

Delgado, M.A. (1993). Testing the equality of nonparametric regression curves. *Statistics and Probability Letters*, 17, 199–204.

Dette, H., & Marchlewski, M. (2010). A robust test for homoscedasticity in nonparametric regression. *Journal of Nonparametric Statistics*, 22, 723–736.

Dette, H., Neumeyer, N., & Van Keilegom, I. (2007). A new test for the parametric form of the variance function in non-parametric regression. *Journal of the Royal Statistical Society, Series B*, 69, 903–917.

Fan, J., & Gijbels. I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman & Hall.

Gastwirth, J.L., Gel, Y., & Miao, W. (2009). The impact of Levene's test of equality of variances on statistical theory and practice. *Statistical Science*, 24, 343–360.

González-Manteiga, W., & Crujeiras, R.M. (2013). An updated review of Goodness-of-Fit tests for regression models. *Test*, 22, 361–411.

Koul, H.L., & Song, W. (2010) Conditional variance model checking. *Journal of Statistical Planning and Inference*, 140, 1056–1072.

Kulasekera, K.B. (1995). Comparison of regression curves using quasi-residuals. *Journal of the American Statistical Association*, 90, 1085–1093.

Levene, H. (1960). Robust tests for equality of variances. In *Contributions to Probability and Statistics* (I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow and H.B. Mann, Eds.), 278–292. Stanford University Press, Stanford.

Liero, H. (2003). Testing homoscedasticity in nonparametric regression. *Journal of Nonparametric Statistics*, 15, 31–51.

Neumeyer, N., & Dette, H. (2003). Nonparametric comparison of regression curves: an empirical process approach. *The Annals of Statistics*, 31, 880–920.

Pardo-Fernández, J.C. (2007). Comparison of error distributions in nonparametric regression. *Statistics and Probability Letters*, 77, 350–356.

Pardo-Fernández, J.C., Jiménez-Gamero, M.D., & El Ghouch, A. (2014a). A nonparametric ANOVA-type test for regression curves based of characteristic functions. *Scandinavian Journal of Statistics*, in press (doi: 10.1111/sjos.12102).

Pardo-Fernández, J.C., Jiménez-Gamero, M.D., & El Ghouch, A. (2014b). Tests for the equality of conditional variance functions in nonparametric regression. Discussion paper 2014/42. Publications of the Institut de Statistique, biostatistique et sciences actuarielles, Université catholique de Louvain (available at: http://www.uclouvain.be/cps/ucl/doc/stat/documents/DP2013_50.pdf).

Pardo-Fernández, J.C., Van Keilegom, I., & González-Manteiga, W. (2007). Testing for the equality of k regression curves. *Statistica Sinica*, 17, 1115-1137.

Srihera, R., & Stute, W. (2010). Nonparametric comparison of regression functions. *Journal of Multivariate Analysis*, 101, 2039–2059.