Change-point Problems
IMS Lecture Notes – Monograph Series (Volume 23, 1994)

A NONPARAMETRIC TEST FOR HOMOGENEITY: APPLICATIONS TO PARAMETER ESTIMATION

BY K. GHOUDI AND D. MCDONALD
Université Laval and University of Ottawa

Testing for homogeneity has many applications in statistical analysis. For example, regression analysis may be viewed as determining the set of parameters that makes the residuals homogeneous. Assume for each $i = 1, \cdots, q$ a small sample $n(i)$ observations is collected. Let $N = \sum_{i=1}^{q} n(i)$, let F^i be the empirical distribution function of the ith sample and let the empirical distribution function of all the samples taken together be $\bar{F}$. The problem is to test if these samples are homogeneous. Lehmann (1951) considered the problem of testing the equality of the distributions of q samples. He proposed the statistic

$$\sum_{i=1}^{q} \int (F^i(s) - \bar{F}(s))^2 d\bar{F}(s).$$

The asymptotic properties of Crámer-Von Mises statistics like the above were studied by Kiefer (1959). He considered the case where $n(i) \to \infty$ while q stayed fixed. McDonald (1991) considered the situation where $q \to \infty$ and $n(i)$ stays fixed for univariate observations for a more general family of statistics called *randomness statistics*. In case of multivariate observations similar asymptotics are discussed in Ghoudi (1992).

Here we present an application of the above statistics to the estimation of the parameters of regression models with independent additive errors. The main novelty of our approach is the use of blocking to contrast the empirical distribution of the residuals of observations whose independent variables are in the same block with the empirical distribution of all the residuals taken together. Our estimated regression surface is the one whose residuals minimize a *randomness statistic* like Lehmann's. Confidence intervals and a test for the model follow without assumptions on the error distribution.

Introduction. Consider a multivariate linear regression model. Observations of a dependent vector $\mathbf{y} = (y_1, y_2, \cdots, y_d)$ and independent variables $\mathbf{x} = (x_1, x_2, \cdots, x_p)$ are indexed by $t = 1, \cdots, N$ and are governed by the model

$$\mathbf{y}_t = (\mathbf{x}_t)'(\beta) + \epsilon_t$$

AMS 1991 Subject Classification: Primary 62G10; Secondary 62J02.

Key words and phrases: Homogeneity tests, nonparametric regression.

with p unknown parameters $(\beta) = (\beta_1, \beta_2, \cdots, \beta_p)'$. We assume throughout that the errors ϵ_t are independent identically distributed d dimensional vectors with unknown distribution F.

In many experimental designs there are replicate experiments for a fixed value of $\mathbf{x}_t$. In this case, group into blocks $i = 1, \cdots, q$ all the observations having the same independent variables $(x_i^1, x_i^2, \cdots, x_i^p)'$. Index the observations in block i by $j = 1, \cdots, n(i)$. The $n(i)$ observations in block i satisfy

$$\mathbf{y}_{ij} = (\mathbf{x}_{ij})'(\beta) + \epsilon_{ij}$$

where $(\mathbf{x}_{ij})' = (x_i^1, x_i^2, \cdots, x_i^p)'$. Even if there are no replicates we may simply pave the parameter space of the dependent variables with contiguous blocks indexed by $i = 1, \cdots, q$. For any choice of β we may calculate the residuals of the dependent vector. Denote the j^{th} residual in the ith block by U_{ij}. The true (unknown) value for β, say β_0, makes the residuals i.i.d. hence the best fit or best choice of β is the one that makes these residuals homogeneous.

Throughout we consider Lehmann's statistic (but any *randomness statistic* would do):

$$\begin{aligned}\Re(\beta) &= \sum_{i=1}^{q} \int \cdots \int \left(F^i(s_1, \cdots, s_d) - \overline{F}(s_1, \cdots, s_d)\right)^2 n(i) dF^i(s_1, \cdots, s_d) \\ &= \sum_{i=1}^{q} \sum_{j=1}^{n(i)} \left(F^i(U_{ij}) - \overline{F}(U_{ij})\right)^2.\end{aligned}$$

In principle the above statistic should be small when the residuals are homogeneous since the empirical distribution of the residuals of each group would then have the same empirical distribution as the entire sample. The best estimate for β should then be the value that minimizes the above *randomness* statistic of the residuals.

In Section 2 we exploit an efficient algorithm for finding the value $\hat{\beta}$ minimizing the statistic $\Re(\beta)$ as a function of β which compares well with the standard least squares estimator. Since the statistic $\Re(\beta)$ is based on empirical distributions it works well when the errors are not normal and even if outliers are present. The estimate $\hat{\beta}$ can be shown to be a consistent estimator of β.

Having found $\hat{\beta}$ we calculate the N residuals. If $\hat{\beta}$ is close to β_0 these residuals should be homogeneous. More precisely, since $\hat{\beta}$ is a consistent estimator of β it can be shown the empirical distribution of the marginals, denoted by $F_{\hat{\beta}}$, converges to the true sampling distribution F. We now consider all $N!$ permutations of the N residuals and for each permutation we reconstruct the blocks $\{n(i) : i = 1, 2, \cdots, q\}$ and we call this a permutation sample. We now

recalculate $\Re$ for each permutation sample. The distribution of these $\Re$ values is called the permutation distribution $\mathcal{P}(\Re, F_{\hat{\beta}})$ (In practice we sample at random from the set of permutation samples to approximate $\mathcal{P}(\Re, F_{\hat{\beta}})$). If the regression model is valid, it follows from the asymptotic normality of *randomness statistics* and the consistency of $\hat{\beta}$ that the quantiles of the permutation distribution $\mathcal{P}(\Re, F_{\hat{\beta}})$ converge to those of the distribution of $\Re(\beta_0)$.

If, in fact, homogeneity is violated one would expect $\Re(\hat{\beta})$ to lie in the upper percentiles of the permutation distribution $\mathcal{P}(\Re, F_{\hat{\beta}})$. If this proves to be the case we reject homogeneity and hence the regression model. Numerical studies in Section 2 show this procedure has good power for detecting deviations from the regression model. Finally if the linear model is accepted, the 95% confidence region for β_0 is simply the set of β such that $\Re(\beta)$ is less than the 95% percentile of the permutation distribution $\mathcal{P}(\Re, F_{\hat{\beta}})$. Note that in the univariate case the quantiles of the distribution $\Re(\beta_0)$ may be calculated exactly since they are distribution free. In the multivariate case the permutation procedure is the only one possible. In fact it is always advisable to use the quantiles of $\mathcal{P}(\Re, F_{\hat{\beta}})$ since the model may have some undetected nonlinear term so using the theoretical quantiles based on the distribution of $\Re(\beta_0)$ might lead to an unjustifiably tight confidence interval.

Nonparametric regression has been treated by Adichie (1967), Jurečkovà (1971), Jaeckel (1972), Hettmansperger and McKean (1977), Koenker and Bassett (1978, 1982) and Gutenbrunner and Jurečkovà (1992).

2. Numerical Results. In this section we present an algorithm for computing the estimate $\hat{\beta}$ described above. Consider all possible subsamples of p different points and index them by J; $J = 1, \ldots, \binom{n}{p}$. Let β_J be the vector of coefficients of the regression surface passing through the p points of subsample J. The computation of such a β_J amounts to the solution of a linear system of p equations with p unknowns. The search procedure goes as follows: for each β_J we compute the corresponding residuals and then the corresponding statistics $\Re(\beta_J)$ which allows us to determine the argument $\hat{\beta}$ giving the minimum of $\Re(\beta)$ and it provides us also with the curve of $\Re(\beta)$ as a function of β.

Note that the number of β's that should be examined is $\binom{n}{p}$. Rousseeuw and LeRoy (1984) reduce computation by randomly selecting m subsamples where m is chosen in a way to insure a high probability of selecting a "good subsample". We can do the same.

We first consider the case $d = 1$ of univariate dependent variables. First we consider the model $y = 4x + \epsilon$ where ϵ is normal with mean 0 and variance 1. We make 5 observations $\{y_{ij} : j = 1, 2, \cdots, 5\}$ at $x_i = i; i = 1, 2, \cdots, 10$.

The graph of $\Re(\beta)$ is given in Figure 1 and the minimum $\Re(\hat{\beta}) = 1.12111$ is obtained at $\hat{\beta} = 3.9033$.

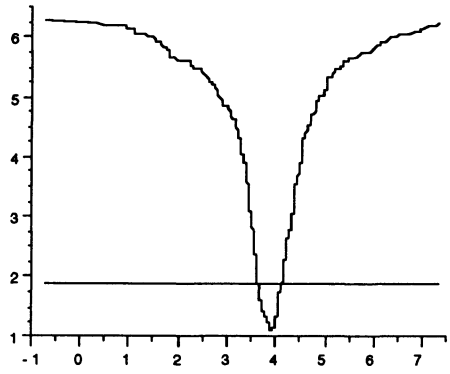

Figure 1: $\Re$ as a function of β

Next we calculate the residuals $y_{ij} - \hat{\beta}x_i$. We then sample at random from the set of permutations of these residuals and then arrange these values in 10 blocks of 5 and recalculate $\Re$. The histogram of these resampled values of $\Re$ is given in Figure 2. The distribution obtained in this way is approximately the same as the permutation distribution $\mathcal{P}(\Re, F_{\hat{\beta}})$.

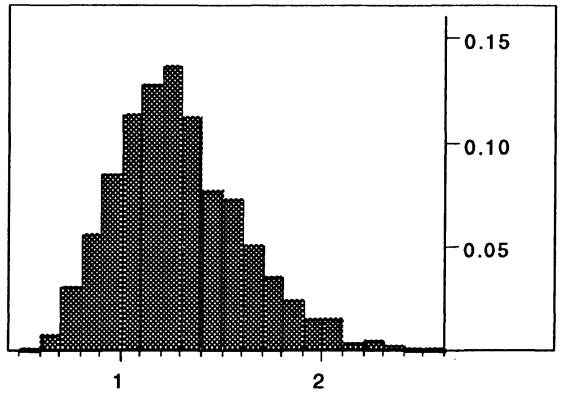

Figure 2: Histogram of the resampled values of $\Re$

We notice that the value $\Re(\hat{\beta}) = 1.12111$ obtained above is in the center of this histogram (at 33.6 percentile) so we conclude the linear model is compatible with our results. Finally the 95th percentile of the estimated permutation distribution, $L_{0.05} = 1.88111$, is plotted on Figure 1 and the associated confidence interval is [3.647,4.096]. The corresponding least squares estimate and confidence interval are 3.892 and [3.782,4.001]. We repeated this experiment 50 times and the histogram of the values of $\hat{\beta}$ is given in Figure 3.

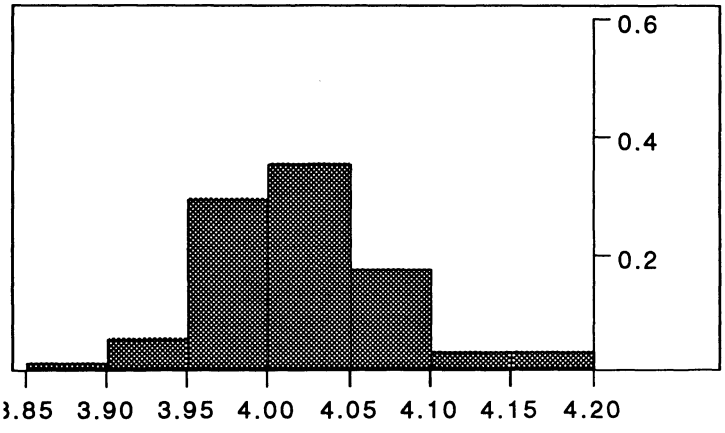

Figure 3: Histogram of the values of $\hat{\beta}$

Suppose the above model is modified to $y = 4x+0.1x^2+\epsilon$ but we continue to fit the model: $y = \beta x + \epsilon$. We repeat the above experiment 50 times and each time we calculate the percentile of $\Re(\hat{\beta})$ in the estimated permutation distribution. The result is summarized in Figure 4 which shows the linear model is (correctly) rejected in each experiment at least at the 80% level.

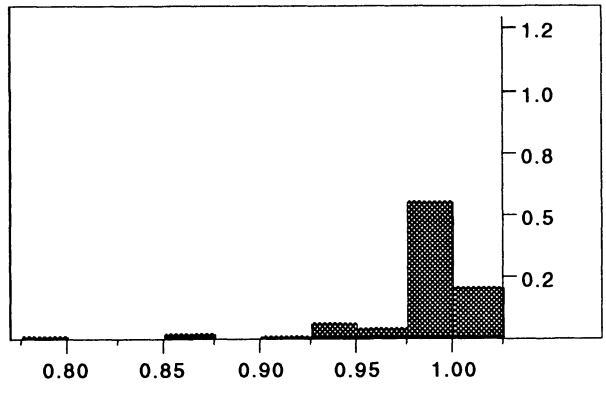

Figure 4: Histogram of the percentile of the observed value of $\Re(\hat{\beta})$

Now modify the first model by changing the distribution of the error ϵ from a standard normal to a standard Cauchy distribution. Figure 3 gives the histogram of the observed values of $\hat{\beta}$ for 50 replications of the experiment. It is clear that our nonparametric procedure is much more successful than the traditional least squares procedure which in some of the above experiments gave estimates like -30.

Consider the model $(y(1), y(2)) = (4, 1)x + (\epsilon(1), \epsilon(2))$ where the components of ϵ are normal with mean 0 and variance 1 and the correlation is 0.5. We make 5 observations $\{(y_{ij}(1), y_{ij}(2)) : j = 1, 2, \cdots, 5\}$ at $x_i = i; i = 1, 2, \cdots, 10$.

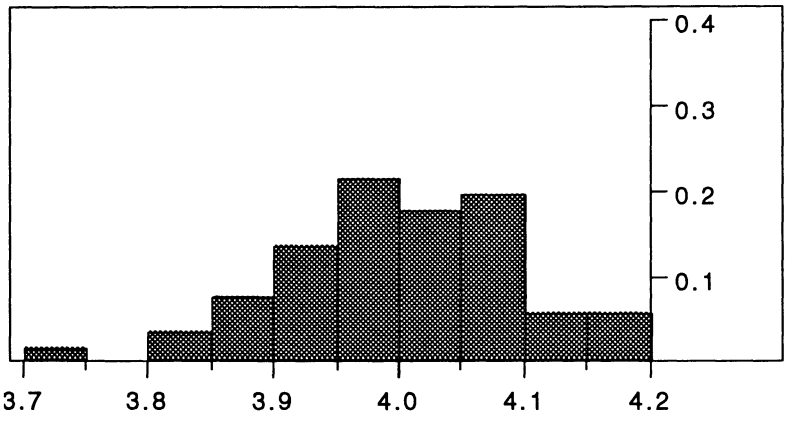

Figure 5: Histogram of the values of $\hat{\beta}$

The graph of $\Re(\beta)$ is given in Figure 6 and the minimum $\Re(\hat{\beta}) = 4.78$ is obtained at $\hat{\beta} = (4.07, 1.08)$.

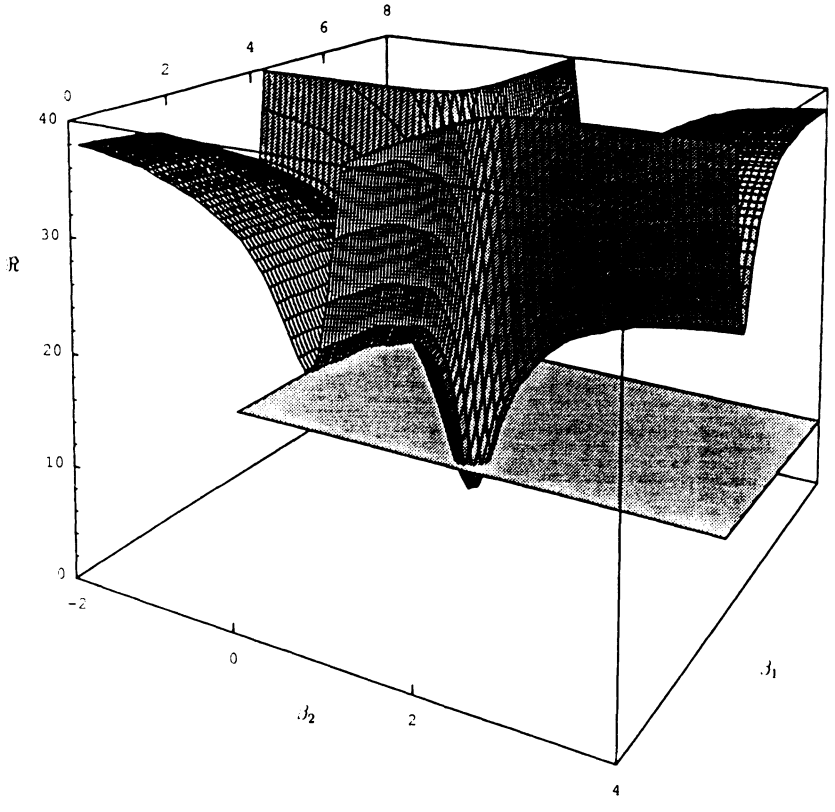

Figure 6: $\Re$ as a function of β

Next we calculate the residuals $(y_{ij}(1), y_{ij}(2)) - \hat{\beta}x_i$. We then sample at random from the set of permutations of these residuals and then arrange these values in 10 blocks of 5 and recalculate $\Re$. The histogram of these resampled values of $\Re$ is given in Figure 7. We notice that the value $\Re(\hat{\beta}) = 4.78$ obtained

above is in the center of this histogram so we conclude the linear model is compatible with our results. Finally the value $L_{0.05} = 6.552$, the 95^{th} percentile of the histogram in Figure 7, is plotted on Figure 6 to give a confidence region for (β_1, β_2). The comparable least squares estimate for the slope is (4.08,1.07).

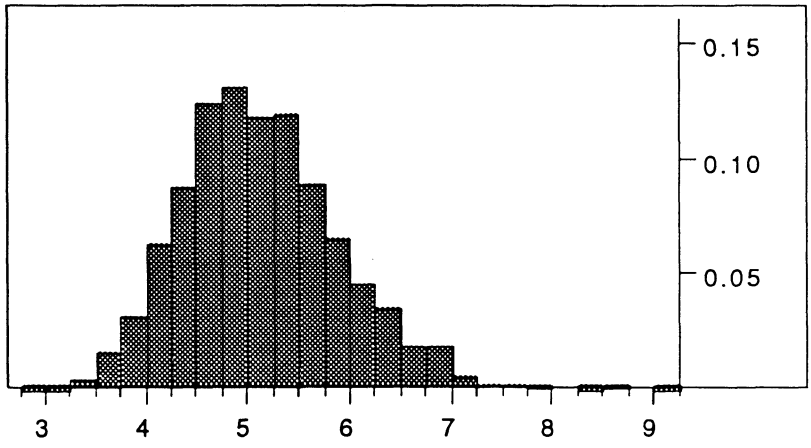

Figure 7: Histogram of the resampled values of $\Re$

REFERENCES

ADICHIE, J. N. (1967). Estimation of regression coefficients based on rank tests. *Ann. Math. Statist.* **38**, 894–904.

BASSETT, G. and KOENKER, R. (1982). An empirical quantile function for linear models with iid errors. *J. Amer. Statist. Assn.* **77**, 407–415.

GHOUDI, K. (1990). *Multivariate Nonparametric Quality Control Statistics.* Master's thesis, University of Ottawa, Ottawa, Ontario Canada.

GHOUDI, K. (1992). *Multivariate Randomness Statistics.* PhD thesis, University of Ottawa, Ottawa, Ontario Canada.

GHOUDI, K. (1992) Asymptotics of multivariate randomness statistics. Preprint.

GUTENBRUNNER, C. and JUREČKOVÁ, J. (1992). Regression rank scores and regression quantiles. *Ann. Statist.* **20**, 305–330.

HETTMANSPERGER, T. P. and MCKEAN, J. W. (1977). A robust alternative based on ranks to least squares in analyzing linear models. *Technometrics* **19**, 275–284.

JAECKEL, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of residuals. *Ann. Math. Statist.* **43**, 1449–1458.

JUREČKOVÁ, J. (1971). Nonparametric estimate of regression coefficients. *Ann. Math. Statist.* **42**, 1328–1338.

KIEFER, J. (1959). K-sample analogues of the Kolmogorov-Smirnov and Cramer-Von Mises tests. *Ann. Math. Statist.* **30**, 420–447.

KOENKER, R. and BASSETT, G. (1978). Regression quantiles. *Econometrica* **46**, 33–50.

KOENKER, R. and BASSETT, G. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica* **50**, 43–61.

LEHMANN, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests. *Ann. Math. Statist.* **22**, 165–179.

MCDONALD, D. (1991). On the asymptotics of randomness statistics. *Can. J. Statist.* **19**, 209–217.

ROUSSEEUW, P. J. and LEROY, A. M. (1987). *Robust Regression and Outlier Detection.* John Wiley, New York.

SIEGEL, A. F. (1982). Robust regression using repeated medians. *Biometrika* **69**, 242–244.

DÉPARTMENT DE MATHÉMATIQUES
ET DE STATISTIQUE
UNIVERSITÉ LAVAL
QUÉBEC, QUÉBEC G1K 7P4, CANADA

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF OTTAWA
OTTAWA, ONTARIO K1N 6N5, CANADA