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

TESTS FOR A CHANGE-POINT IN LINEAR REGRESSION

BY HYUNE-JU KIM

Syracuse University

This paper considers a problem of detecting a change-point in a linear model. We discuss analytic properties of the likelihood ratio statistic and study its asymptotic behavior. An approximation for the significance level of the test is provided assuming values of the independent variables are effectively random. We also discuss the power and the robustness of the likelihood ratio test.

1. Introduction. The problem of detecting a change-point in a linear regression model has been addressed by many authors. While likelihood ratio statistics to test for parameter changes in a broader sense, have been derived (Quandt (1960), Worsley (1983)) for simple and multiple regression models, mathematical difficulties associated with the sampling distribution of the likelihood ratio statistic have hindered its application in the past. As discussed in Feder (1975a,b) and Quandt (1958, 1960), a proposed chi-squared approximation for the null distribution of the likelihood ratio statistic is very poor and maximum likelihood estimators are not asymptotically normal.

Brown, Durbin, and Evans (1975) introduced recursive residuals to test changes in multiple regression models. Although the sampling distribution of the cumulative sum and cumulative sum squares of the recursive residuals are relatively simple under the null hypothesis of no change, Brown et al. left the alternative hypothesis unspecified and its power to detect specific changes has remained in question. Bayes type tests, first introduced by Chernoff and Zacks (1964), are extended in the regression model by Jandhyala and MacNeill (1991). They discussed asymptotic distribution theory of the Bayes type statistics and suggested a numerical method to compute its critical values. The test of Jandhyala and MacNeill, however, considers a change with the same size in every component of the regression coefficient vector and the computation of the critical values is not simple. Hawkins (1989) considered a union and and intersection approach to detect parameter shifts in a linear regression

AMS 1991 Subject Classification: Primary 62F05; Secondary 62J05

Key words and phrases: Two-phase regression, likelihood ratio test, boundary crossing probability .

model and showed that his statistic converges to a functional of the standard Brownian bridge process. There are Bayesian approaches discussed in Ferreira (1975), Broemeling and Moen (1984), Holbert (1982) D. Kim (1991), and others.

Kim and Siegmund (1989) considered the likelihood ratio test for a change-point in simple linear regression and derived analytic approximations for the significance level of the test. Loader (1992) also used likelihood ratio approaches for modeling nonhomogeneous Poisson processes. Using large deviation methods, Loader obtained approximations for the significance level, power and the confidence level.

This paper considers a problem of detecting a change-point in multiple linear regression. Our aim in this paper is to study analytic properties of the likelihood ratio statistics and to indicate its applicability in practice. In Section 2, we will define a generalized likelihood ratio test and study the asymptotic behavior of the test statistic. An approximation for the significance level of the test is also provided. Section 3 includes further discussion on the power and the robustness of the likelihood ratio test (LRT).

2. Likelihood Ratio Statistics and Asymptotic Properties. Suppose that we have a sequence of observations $(\mathbf{x_j}, y_j)$ $j = 1, \ldots, n$, where, given $\mathbf{x_j'} = (1, x_{j,1}, \ldots, x_{j,p-1})$, the y_j's are normally distributed with mean $\mu(\mathbf{x_j})$ and variance σ^2. The null hypothesis of interest is that these observations satisfy a linear regression model: $H_0 : \mu(\mathbf{x_j}) = \mathbf{x_j'}\boldsymbol{\beta}$ for $j = 1, \ldots, n$. Under the alternative H_1, there is a change-point ρ such that $\mu(\mathbf{x_j}) = \mathbf{x_j'}\boldsymbol{\beta}$ if $j \leq \rho$ and $\mu(\mathbf{x_j}) = \mathbf{x_j'}\boldsymbol{\beta}^*$ if $j > \rho$.

Worsley (1983) studied the LRT for H_0 against H_1, and provided an upper bound for the significance level of the test. Numerical examples indicate that his approximate upper bounds are reasonably accurate for small samples and for moderate sizes of the significance level. However, the approximation loses its accuracy for large samples and its accuracy in multiple regression has not been supported in the paper.

For the formulation of the likelihood ratio statistics under a variety of assumptions, we refer to Worsley (1983). Now we introduce a useful alternative expression for the likelihood ratio statistic. For $\kappa = 1, \ldots, p$, let $\mathbf{a}_\kappa'$ be a $(p+k) \times 1$ vector with 1 and -1 at the $(2\kappa - 1)$st and the 2κth component, respectively, and 0 in all other components, and let $\mathbf{X}_{\kappa,j}$ be the $n \times (p + \kappa)$ matrix such that

$$\mathbf{X}_{\kappa,j}\mathbf{e}_{\kappa,i} = (\mathbf{1_j'}, \mathbf{0_{n-j}'}) \quad \text{for} \quad i = 1,$$
$$\mathbf{X}_{\kappa,j}\mathbf{e}_{\kappa,i} = (\mathbf{0_j'}', \mathbf{1_{n-j}'}) \quad \text{for} \quad i = 2,$$

$$\mathbf{X}_{\kappa,j}\mathbf{e}_{\kappa,i} = (x_{(i-1)/2,1}, \ldots, x_{(i-1)/2,j}, \mathbf{0}'_{\mathbf{n-j}}) \quad \text{for} \quad i = 3, 5, \ldots, 2\kappa - 1,$$
$$\mathbf{X}_{\kappa,j}\mathbf{e}_{\kappa,i} = (\mathbf{0}'_{\mathbf{j}}, x_{(i/2-1),j+1}, \ldots, x_{(i/2-1),n}) \quad \text{for} \quad i = 4, 6, \ldots, 2\kappa,$$
$$\mathbf{X}_{\kappa,j}\mathbf{e}_{\kappa,i} = (x_{(i-1)/2,1}, \ldots, x_{(i-1)/2,n}) \quad \text{for} \quad i = 2\kappa + 1, 2\kappa + 3 \ldots,$$

and

$$\mathbf{X}_{\kappa,j}\mathbf{e}_{\kappa,i} = (x_{i/2,1}, \ldots, x_{i/2,n}) \quad \text{for} \quad i = 2\kappa + 2, 2\kappa + 4 \ldots.$$

where $\mathbf{1_j}$ is the $j \times 1$ vector of 1's, $\mathbf{0_j}$ is the $j \times 1$ vector of 0's, and $\mathbf{e}_{\kappa,i}$ is the $(p + \kappa) \times 1$ vector with 1 in the ith component and with 0 in all other components. Define

$$U_\kappa(j) = [\mathbf{a}'_{\kappa,j}(\mathbf{X}'_{\kappa,j}\mathbf{X}_{\kappa,j})^{-1}\mathbf{X}'_{\kappa,j}\mathbf{Y}]/\{\mathbf{a}'_{\kappa,j}(\mathbf{X}'_{\kappa,j}\mathbf{X}_{\kappa,j})^{-1}\mathbf{a}_{\kappa,j}\}^{1/2}$$

for $\kappa = 1, \ldots, p$, where $\mathbf{Y}' = (y_1, \ldots, y_n)$.

Then the LRT to test H_0 against H_1 can be based on

$$\hat{\sigma}^{-2} \max_{p \le j \le n-p} \|\mathbf{U}(j)\|^2 = \hat{\sigma}^{-2} \max_{p \le j \le n-p} \{U_1^2(j) + \cdots + U_p^2(j)\},$$

where $\hat{\sigma}^{-2}$ is the maximum likelihood estimate of σ^2 under H_0. Note that U_κ is the test statistic to test for a change also in β_κ assuming that there is a change in $(\beta_1, \ldots, \beta_{\kappa-1})$. It is easy to see that under the null hypothesis of no change, $U_\kappa(j)$ is a standard normal random variable for each κ and j, and the covariance between $\mathbf{U}(i)$ and $\mathbf{U}(j)$ can be found by using a straightforward computation.

As discussed in Maronna and Yohai (1978), the likelihood ratio statistics converge to infinity in distribution as $n \to \infty$. To achieve a valid limiting distribution, Kim and Siegmund (1989) suggested a generalized LRT, which rejects H_0 for a large value of $\hat{\sigma}^{-2} \max_{n_0 \le j \le n_1} \|\mathbf{U}(j)\|^2$, where $1 < n_0 < n_1 < n$. The introduction of n_0 and n_1 also improves the power of the LRT provided the change-point is not near 1 and n. For more discussion on the choice of n_0 and n_1, see Kim and Cai (1993a).

Let $\mathbf{X}_j$ be the first j rows of the design matrix $\mathbf{X}$, and define $Q_t = \lim_{j/n \to t} j^{-1}(\mathbf{X}'_j\mathbf{X}_j)$. If there exist a positive definite matrix Q such that $Q_t \equiv Q$ for all t, then straightforward convergence argument shows that as n, n_0, and $n_1 \to \infty$, in such a way that $n_i/n \to t_i$ $(i = 0, 1)$,

$$\hat{\sigma}^{-2} \max_{n_0 \le j \le n_1} \| \mathbf{U}(j) \|^2 \to \max_{t_0 \le t \le t_1} \| \mathbf{W}(t) \|^2/\{t(1-t)\}^{1/2} \quad \text{in distribution,}$$

where $\mathbf{W}$ is a p-dimensional Brownian bridge process. Then a simpler large sample approximation for the significance level of the test can be obtained by considering the maximum of the limiting process. However, this approximation usually overestimates the true probability about forty to one hundred percent,

and thus we provide the following discrete approximation by extending the arguments in Siegmund (1985, Chap. 12): as $b \to \infty$ in such a way that $b^2/n \to c^2$,

$$P\{\max_{n_0 \le j \le n_1} \| \mathbf{S}_j - j\mathbf{S}_n/n \| / \{j(1-j/n)\}^{1/2} > b\} \sim \frac{b^p \exp(-b^2/2)}{2^{1-p/2}\Gamma(p/2)} \int_{c\sqrt{t_1^{-1}-1}}^{c\sqrt{t_0^{-1}-1}} r^{-1} \cdot \nu(r + c^2/r)\, dr, \tag{2.1}$$

where $\mathbf{S}_j$ is the partial sum of j independent multivariate normal random vectors with zero mean vector and identity covariance matrix, and

$$\nu(x) = 2x^{-2} \exp[-2 \sum_{1}^{\infty} n^{-1} \Phi(-x\sqrt{n}/2)] \qquad (x > 0).$$

A particularly simple case where $Q_t \equiv Q$ for all t occurs if $\mathbf{x_1}, \ldots, \mathbf{x_{p-1}}$ are observed values of random vectors. Table 1 shows the accuracy of the approximation, (2.1), for $p = 3$. $\mathbf{x}_1$ is taken to be a standard normal random variable, and $\mathbf{x}_2$ is the normal random variable with mean 5 and variance 4. The critical values are estimated by a 10,000 repetition Monte Carlo experiment and (2.1) is evaluated at the estimated percentiles. As Table 1 shows, the approximations are quite accurate in small samples but tend to be less accurate for larger sample sizes and for larger values of p. An *ad hoc* modification suggested in James, James, Siegmund (1992) might be desirable to improve its accuracy.

Table 1. Accuracy of Approximation (2.1)
$n_0 = .1 \times n, \qquad n_1 = .9 \times n$

n	Significance level	Estimated critical values (b)	Approximation (2.1)
20	0.10	3.2164	0.1033
	0.05	3.4596	0.0534
	0.01	3.9778	0.0102
40	0.10	3.3299	0.1220
	0.05	3.5761	0.0613
	0.01	4.0645	0.0130

When all of $\mathbf{x}_\kappa$, $\kappa = 1, \ldots, p-1$, are fixed, approximate significance levels can be obtained by extending the argument in Kim and Siegmund (1989).

However, it involves p-dimensional integration, which is not computationally efficient. Furthermore, the computation of the integrand of the multiple integration is not easy, and thus we do not report the result for the case of 'fixed' x's in this paper. If one of the regressor variables is fixed while all others are random, the situation seems a lot simpler than the case of 'fixed' x's and will be considered in a future paper. In the case of simple linear regression with 'fixed' x's, the significance levels can be approximated by using the results of Kim and Siegmund (1989).

REMARK. In the case that the observations are normally distributed, it is not necessary to introduce n_0 and n_1. The significance level of the original likelihood ratio test can still be considered as a boundary crossing probability by a discrete Gaussian process. Then a large deviation method can be used to obtain a similar asymptotic expression for the significance level of the test.

3. Discussion. Our concern in this section is the power and the robustness of the LRT. James, James, and Siegmund (1987) compare the powers of the likelihood ratio, the test based on the forward cumulative sum (CUSUM) of the recursive residuals, and backward CUSUM test to detect a mean change in a sequence of the independent random variables. They showed that the forward CUSUM test, which is the one suggested by Brown et al. is much less powerful than the likelihood ratio test and proposed the backward CUSUM test to improve the power of the CUSUM test. Kim (1992) considered the problem of detecting a change in the intercept term of simple linear regression and derived an approximation for the power of the LRT. In her study, Kim showed that, regardless of the inclusion of the covariate, the LRT outperforms the forward CUSUM test and achieves almost the same power as the backward CUSUM test. Kim and Cai (1993a) discussed the distributional robustness of the LRT and observed that the LRT achieves almost the same level and power regardless of the underlying distribution. Two-phase regression with nonhomogeneous errors has been studied in Kim (1993b), who concluded that the effects of the nonhomogeneity of variance may not be disastrous to the use of the LRT, unless the disproportionality between the variance is particularly severe.

Our aim in this paper has been to indicate the applicability of the LRT whose use has presented some intractable analytic difficulties in the past. There are still many open problems left such as a full derivation of the asymptotic significance level, asymptotic power, confidence regions, and so on, and will be discussed in a future paper.

REFERENCES

BROEMELING, L. D. and MOEN, D. H. (1984). Testing for a change in the regression matrix of a multivariate linear model. *Comm. Statist.* **A 13**, 1521–1531.

BROWN, R. L., DURBIN, J. and EVANS, J. M. (1975). Techniques for testing the constancy of regression relationships over time. *J. R. Statist. Soc.* **B 37**, 149–192.

CHERNOFF, H. and ZACKS, S. (1964). Estimating the current mean of a normal distribution which is subject to changes in time. *Ann. Math. Statist.* **35**, 999–1018.

FEDER, P. L. (1975a). On asymptotic distribution theory in segmented regression problems-identified case. *Ann. Statist.* **3**, 49–83.

FEDER, P. L. (1975b). The log likelihood ratio in segmented regression. *Ann. Statist.* **3**, 84–97.

FERREIRA, P. E. (1975). A Bayesian analysis of a switching regression model: Known number of regimes. *J. Amer. Statist. Assn.* **70**, 370–374.

HAWKINS, D. L. (1989). A U-I approach to retrospective testing for shifting parameters in a linear model. *Comm. Statist.* **18**, 3117–3134.

HOLBERT, D. (1982). A Bayesian analysis of a switching linear model. *J. Econometrics* **19**, 77–87.

JANDHYALA, V. K. and I. B. MACNEILL (1991). Tests for parameter changes at unknown times in linear regression models. *J. Statist. Plan. Inf.* **27**, 291–316.

JAMES, B., JAMES, K. L. and SIEGMUND, D. (1992). Asymptotic approximation for likelihoos ratio tests and confidence regions for a change-point in mean of a multivariate normal distribution. *Statistica Sinica* **2**, 69–90.

KIM, D. (1991). A Bayesian significance test of the stationarity of regression parameters. *Biometrika* **78** , 667–675.

KIM, H. J. (1992). The power of the likelihood ratio and cusum tests for a level shift in simple linear regression. Preprint, Syracuse University.

KIM, H. J. and CAI, L. (1993a). Robustness of the likelihood ratio test for a change in simple linear regression. *J. Amer. Statist. Assn.* **88**, 864–871.

KIM, H. J. (1993b). Two-phase regression with nonhomogeneous errors. *Comm. Statist. – Theory and Methods* **22**, 647–658.

KIM, H. J. and SIEGMUND, D. (1989). The likelihood ratio test for a change-point in simple linear regression. *Biometrika* **76**, 409–423.

LOADER C. R. (1992). A log-linear model for a Poisson process change-oint. *Ann. Statist.* **20**, 1391–1411.

MARONNA, R. and YOHAI, V. J. (1973). A bivariate test for the detection of a systematic change in mean. *J. Amer. Statist. Assn.* **73**, 640–645.

QUANDT, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *J. Amer. Statist. Assn.* **53**, 873–880

QUANDT, R. E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. *J. Amer. Statist. Assn.* **55**, 324–30

SIEGMUND, D. (1985). *Sequential Analysis: Test and Confidence Intervals.* New York: Springer-Verlag.

WORSLEY, K. J. (1983). Testing for a two-phase multiple regression. *Technometrics* **25**, 35–42.

DEPARTMENT OF MATHEMATICS
SYRACUSE UNIVERSITY
SYRACUSE, NY 13244