On Designs for Recursive Least Squares Residuals to Detect Alternatives

Wolfgang Bischoff

Abstract Linear regression models are checked by a lack-of-fit (LOF) test to be sure that the model is at least approximatively true. In many practical cases data are sampled sequentially. Such a situation appears in industrial production when goods are produced one after the other. So it is of some interest to check the regression model sequentially. This can be done by recursive least squares residuals. A sequential LOF test can be based on the recursive residual partial sum process. In this paper we state the limit of the partial sum process of a triangular array of recursive residuals given a constant regression model when the number of observations goes to infinity. Furthermore, we state the corresponding limit process for local alternatives. For specific alternatives designs are determined dominating other designs in respect of power of the sequential LOF test described above. In this context a result is given in which e^{-1} plays a crucial role.

1 Introduction

In order to guarantee the quality of each delivery of contract goods companies take samples to decide whether the quality is or is not constant. If the goods are sequentially produced this problem can be modelled by the regression model

$$Y(t) = g(t) + \varepsilon(t), \quad t \in [0,1], \tag{1}$$

where g is the true, but unknown mean function of the quality, $\varepsilon(t)$ is a real random variable with expectation 0 and variance $\sigma^2 > 0$, and $[0,1]$ is the period of production. Since our results keep true when σ^2 is replaced by a consistent estimator for σ^2 we can put $\sigma^2 = 1$ without loss of generality.

Wolfgang Bischoff

Katholische Universitaet Eichstaett-Ingolstadt, Mathematisch-Geographische Fakultaet, D-85071 Eichstaett, Germany, e-mail: wolfgang.bischoff@ku.de

We consider the problem more generally. We like to test if the model (1) is a linear model with respect to known and linearly independent functions $f_1, \ldots, f_d$, i.e. if there exist suitable constants $\beta_1, \ldots, \beta_d \in \mathbf{R}$ such that $g(t) = \sum_{i=1}^{d} \beta_i f_i(t), t \in [0,1]$. Hence, we look for a test of the null hypothesis

$$H_0 : g = \sum_{i=1}^{d} \beta_i f_i = f^\top \beta \text{ for some } \beta = (\beta_1, \ldots, \beta_d)^\top \in \mathbf{R}^d, \qquad (2)$$

where $f^\top = (f_1, \ldots, f_d)$, against the alternative that model (1) is not a linear model with respect to $f_1, \ldots, f_d$, that is (2) is not fulfilled.

In quality control we are interested in testing

$$H_0 : g(t) = \beta = \mathbf{1}_{[0,1]}(t)\beta, \quad t \in [0,1], \ \beta \in \mathbf{R} \text{ unknown constant}, \qquad (3)$$

where $\mathbf{1}_{[0,1]}$ is the function identical 1 on $[0,1]$, against the alternative

$$K : g \neq constant. \qquad (4)$$

Such a function g under the alternative has typically the following form:

$$g(t) = \beta \text{ for } t \in [0, t_0], \ g \text{ increasing or decreasing for } t \in (t_0, 1].$$

This form of g means that the quality keeps constant up to a fixed, known or unknown change–point $t_0 \in (0,1)$, then the quality is getting worse or is getting better.

In the literature on 'detecting changes' in regression models, it is common to consider (recursive) residual partial sum processes or variants thereof; see for instance, Gardner [8], Brown, Durbin and Evans [5], Sen and Srivastava [18], MacNeill [13] and [14], Sen [17], Jandhyala and MacNeill [9], [10] and [11], Watson [19], Bischoff [1], Jandhyala, Zacks and El-Shaarawi [12], Bischoff and Miller [3], Xie and Mac-Neill [20], Bischoff and Somayasa [4], Bischoff and Gegg [2]. The asymptotics of the partial sum of recursive residuals is investigated by Sen [17] only. Sen [17], however, assumed a time series sampling for his asymptotic result. For our problem we need a triangular array approach.

In Section 2 we discuss some asymptotic results for the partial sum process of recursive least squares residuals. Assuming a constant regression model we state such a result for a triangular array of design points on one hand if the null hypothesis (3) is true and on the other hand under certain assumptions if a local alternative (4) is true. With this we are in the position to establish an asymptotic size α test to test the null hypothesis (3). Furthermore, in Section 3, we can discuss the power of this test for certain alternatives. There, we determine designs that have uniformly more power than other designs. For one of the results e^{-1} occurs as crucial constant.

2 Recursive Residuals

Recursive (least squares) residuals were described in Brown et al. [5], for some history see Farebrother [7]. Brown et al. considered recursive residuals for a linear regression model given a time series sampling. To this end let $t_1 < t_2 < \ldots$ be a sequence of (design) points (in time), let $\varepsilon_1, \varepsilon_2, \ldots$ be iid real random variables with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = 1$, let $n \in \mathbb{N}, n > d$, be the number of observations where d is the number of known regression functions. Moreover, we put

$$X_i = (f(t_1), \ldots, f(t_i))^\top, \, d \leq i \leq n,$$

where $f = (f_1, \ldots, f_d)^\top$. So we get $n - d + 1$ linear models, namely for the first i, $d \leq i \leq n$, observations each

$$\mathbf{Y}_i = (Y_1, \ldots, Y_i)^\top = X_i \beta + (\varepsilon_1, \ldots, \varepsilon_i)^\top, \, d \leq i \leq n.$$

Let $t_1, \ldots, t_d$ be chosen in such a way that $rank(X_d) = d$. Then, for each $i \geq d$ we estimate β by the least squares estimate $\hat{\beta}_i$ using the first i observations $\mathbf{Y}_i$. Now we can define $n - d$ recursive residuals

$$e_i = \frac{Y_i - f(t_i)^\top \hat{\beta}_{i-1}}{\left(1 + f(t_i)^\top (X_{i-1}^\top X_{i-1})^{-1} f(t_i)\right)^{1/2}}, \, i = d+1, \ldots, n.$$

To state Sen's and our result it is convenient to define the partial sum operator T_n : $\mathbb{R}^n \longrightarrow C[0,1], \mathbf{a} = (a_1, \ldots, a_n)^\top \mapsto T_n(\mathbf{a})(z), \, z \in [0,1]$, where

$$T_n(\mathbf{a})(z) = \sum_{i=1}^{[nz]} a_i + (nz - [nz]) a_{[nz]+1}, \, z \in [0,1].$$

Here we used $[s] = \max\{n \in \mathbb{N}_0 \mid n \leq s\}$ and $\sum_{i=1}^0 a_i = 0$. Let $\mathbf{a} = (a_1, \ldots, a_n)^\top \in \mathbb{R}^n, b_i = a_1 + \ldots + a_i, i = 1, \ldots, n$, then the function $T_n(\mathbf{a})(\cdot)$ is shown in Figure 2.
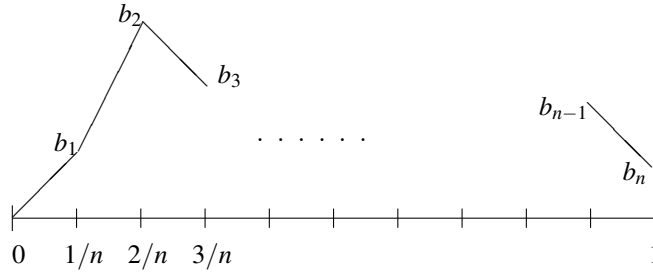

Fig 2. The function $T_n(\mathbf{a})(\cdot)$.

By Donsker's Theorem the stochastic process $\frac{1}{\sqrt{n}} T_n((\varepsilon_1, \ldots, \varepsilon_n)^\top)$ converges weakly to Brownian motion B for $n \to \infty$. For recursive residuals Sen [17] proved the following result.

Theorem 1 (Sen [17]). *For the regression model given in (2) let* $\mathbf{e}_n = (e_{d+1}, \ldots, e_n)^\top$. *If H_0 given in (2) is true, then under certain assumptions it holds true*

$$\frac{1}{\sqrt{n-d}} T_{n-d}(\mathbf{e}_n)(z) \text{ converges weakly to } B(z), \ z \in [0,1], \text{ for } n \to \infty.$$

Sen, however, could not determine the limit process for a local alternative.

The time series sampling approach described above cannot be applied to problems of experimental design. Instead we need the asymptotic result for a triangular array of design points under the null hypothesis and under local alternatives. To this end let $n_0 \in \mathbb{N}$, $n_0 > d$, be the number of observations. We assume that the data are taken at the design points $0 \leq t_{n_0 1} \leq t_{n_0 2} \leq \ldots \leq t_{n_0 n_0} \leq 1$. These design points can be embedded in a triangular array of design points: $0 \leq t_{n1} \leq t_{n2} \leq \ldots \leq t_{nn} \leq 1$, $n \in \mathbb{N}$. Furthermore, let $\varepsilon_{n1}, \ldots, \varepsilon_{nn}, n \in \mathbb{N}$, be a triangular array of real random variables where $\varepsilon_{n1}, \ldots, \varepsilon_{nn}$ are iid with $\mathbf{E}(\varepsilon_{ni}) = 0$ and $Var(\varepsilon_{ni}) = 1$ for each $n \in \mathbb{N}$. Accordingly, we get a corresponding triangular array of observations for model (1):

$$Y_{nj} = g(t_{nj}) + \varepsilon_{nj}, \ 1 \leq j \leq n, \ n \in \mathbb{N}.$$

We put

$$\varepsilon_i^n = (\varepsilon_{n1}, \ldots, \varepsilon_{ni})^\top, \ d \leq i \leq n,$$
$$X_i^n = (f(t_{n1}), \ldots, f(t_{ni}))^\top, \ d \leq i \leq n.$$

So we get $n - d + 1$ linear models under the null hypothesis H_0 given in (2), namely for the first i, $d \leq i \leq n$, observations each

$$\mathbf{Y}_i^n = (Y_{n1}, \ldots, Y_{ni})^\top = X_i^n \beta + \varepsilon_i^n.$$

Let $t_{n1}, \ldots, t_{nd}$ for all $n \geq d$ be chosen in such a way that $rank(X_d^n) = d$. Moreover, let $\hat{\beta}_i^n, d \leq i$, be the least squares estimate for β using the first i observations $\mathbf{Y}_i^n$. Then the $n - d$ recursive least squares residuals for the triangular array are defined by

$$e_{ni} = \frac{Y_{ni} - f(t_{ni})^\top \hat{\beta}_{i-1}^n}{\left(1 + f(t_{ni})^\top (X_{ni-1}^\top X_{ni-1})^{-1} f(t_{ni})\right)^{1/2}}, \ i = d+1, \ldots, n.$$

Assuming the constant regression model the next result states the limit of the recursive residual partial sum process if H_0, see (3), is true and if a local alternative is true. In case (3) is true, the location of the design points has no influence. So the result is true for any triangular array of design points. If a local alternative is true, we give the result for a uniform array of design points only to avoid further technical notation. This result will be sufficient for our purposes below.

Theorem 2 (Master Thesis Rabovski [15] under the supervision of the author and Frank Miller). *For the constant regression model let* $\mathbf{e}^n = (e_{nd+1}, \ldots, e_{nn})^\top$ *be the vector of the $n - d$ recursive residuals of a triangular array of design points.*

a) *If H_0 given in (3) is true, then for any triangular array of design points*

$$\frac{1}{\sqrt{n-d}}T_{n-d}(\mathbf{e}^n)(z) \text{ converges weakly to } B(z), \ z \in [0,1], \text{ for } n \to \infty.$$

b) *Let $g : [0,1] \to \mathbb{R}, g \neq \text{constant, have bounded variation and let the triangular array of design points be given by a uniform design}$*

$$t_{n1} = 0, t_{n2} = \frac{1}{n-1}, t_{n3} = \frac{2}{n-1}, \ldots, t_{nn} = 1, n \in \mathbb{N}.$$

Then, if the local alternative $\frac{1}{\sqrt{n-d}}g$ is true,

$$\frac{1}{\sqrt{n-d}}T_{n-d}(\mathbf{e}^n)(z) \text{ converges weakly to } h(z) + B(z), \ z \in [0,1], \text{ for } n \to \infty,$$

where
$$h(z) = \int_0^z g(t)dt - \int_0^z \frac{1}{s} \int_0^s g(t)dt\,ds, \ z \in [0,1].$$

Theorem 2 part a) can be used to establish an asymptotic size α test of Kolmogorov(-Smirnov) or Cramér-von Mises type. As an example we state a one-sided test of Kolmogorov type, to detect a negative deviation h from Brownian motion.

Theorem 3. *For the constant regression model let $\mathbf{e}^n = (e_{nd+1}, \ldots, e_{nn})^\top$ be the vector of the $n - d$ recursive residuals of an arbitrary triangular array of design points. Then an asymptotic size α test is given by*

$$\text{reject } H_0 \text{ given in (3)} \iff \exists t \in [0,1] : \frac{1}{\sqrt{n-d}}T_{n-d}(\mathbf{e}^n)(t) < \Phi^{-1}(\frac{\alpha}{2}),$$

where $\Phi^{-1}(\frac{\alpha}{2})$ is the $\alpha/2$ quantile of the standard normal distribution.

Proof. Since $P(\exists t \in [0,1] : B(t) < \Phi^{-1}(\frac{\alpha}{2})) = \alpha$, see, for instance, Shorack [16] p.314, the theorem is proved.

The above test is not constructed sequentially. The test statistic $\frac{1}{\sqrt{n-d}}T_{n-d}(\mathbf{e}^n)(t)$ can be calculated sequentially for each new recursive residual e_{ni} and so the null hypothesis can be rejected as soon as the test statistic is less than $\Phi^{-1}(\frac{\alpha}{2})$.

3 Designs for Detecting Alternatives

We look for designs being useful for the quality problem discussed in the introduction. Therefore we consider the constant regression model. Usually in the context of quality control certain properties of the alternative are often known.

We begin with the alternative

$$g_{t_0}(t) = g_{t_0;c_0,c_1}(t) = c_0 \mathbf{1}_{[0,t_0]}(t) + c_1 \mathbf{1}_{(t_0,1]}(t), \ t \in [0,1], \tag{5}$$

where $c_0, c_1 \in \mathbb{R}$ are unknown constants and the change-point $t_0 \in (0,1)$ is a known or unknown constant. We assume $c_0 > c_1$ to get a negative trend h, see Theorem 4. (Note that Theorem 3 states a test for detecting negative trends h.) Let a triangular array of design points be given with

$$q := \lim_{n \to \infty} \frac{\text{number of } \{t_{ni}|t_{ni} \le t_0, 1 \le i \le n\}}{n} \in (0,1).$$

We call such a triangular array of design points an asymptotic q-design. The proof of the following result is given in the next section.

Theorem 4. *For a constant regression model let $\mathbf{e}^n = (e_{nd+1}, \dots, e_{nn})^\top$ be the vector of the $n - d$ recursive residuals of a triangular array of design points being an asymptotic q-design. Let the alternative g_{t_0} given in (5) be true. Then, we have for the local alternative $\frac{1}{\sqrt{n-d}} g_{t_0}$:*

$$\frac{1}{\sqrt{n-d}} T_{n-d}(\mathbf{e}^n)(z) \text{ converges weakly to } h(z) + B(z), \ z \in [0,1], \text{ for } n \to \infty,$$

where
$$h(z) = q(c_1 - c_0)(\ln(z) - \ln(q))\mathbf{1}_{(q,1]}(z), \ z \in [0,1].$$

For an asymptotic q-design the power of the test given in Theorem 3 with respect to the alternative (5) is given by

$$P(\exists z \in [0,1] : B(z) - q(c_0 - c_1)(\ln(z) - \ln(q))\mathbf{1}_{(q,1]}(z) \le \Phi^{-1}(\frac{\alpha}{2})).$$

Therefore, we call an asymptotic q^*-design uniformly better than an asymptotic q-design if for all $z \in (0,1]$

$$-q^*(\ln(z) - \ln(q^*))\mathbf{1}_{(q^*,1]}(z) \le -q(\ln(z) - \ln(q))\mathbf{1}_{(q,1]}(z) \tag{6}$$
$$\text{with } ' <' \text{ at least for one } z \in [0,1].$$

The proof of the following result is given in the next section.

Theorem 5. *Let the situation considered in Theorem 4 be given and let $q_1, q_2 \in [e^{-1}, 1)$. Then an asymptotic q_1-design is uniformly better than an asymptotic q_2-design, if $q_1 < q_2$.*

The author does not know whether the result stated above has some relation to the famous e^{-1}-law for the best choice problem, see, for instance, Bruss [6]. For n_0 design points we consider the design d^* with the fractional part of about e^{-1} design points as near as possible at 0 (let t_1^* be the largest of these design points) and the fractional part of about $1 - e^{-1}$ design points as near as possible at 1 (let t_2^* be

the smallest of these design points). Then, by Theorem 5, d^* is asymptotically the uniformly best applicable design, if $t_1^* < t_2^*$ and $t_1^* \leq t_0$.

Finally, we consider the alternative

$$g(t) = c_0 \mathbf{1}_{[0,t_0]}(t) + (c_0 + c_1 t_0 - c_1 t) \mathbf{1}_{(t_0,1]}(t), \ t \in [0,1], \tag{7}$$

where $t_0 \in [0,1)$ is a known or unknown constant and $c_0 \in \mathbb{R}, c_1 \in (0,\infty)$ are unknown constants. The last result follows in an analogous way as above.

Theorem 6. *The asymptotic 0-design which is uniformly distributed on $[t_0,1]$ is uniformly better with respect to the alternative (7) than an asymptotic q-design, $q \in (0,1)$, whose fractional part of design points on $[t_0,1]$ is uniformly distributed.*

For an unknown change-point t_0 the above result is of theoretical interest only.

4 Some Proofs

The following relation between an arbitrary design and a uniform design is crucial for the next proof. To this end let the alternative (5) and an arbitrary triangular array of design points $t_{n1},\ldots,t_{nn}$ with $0 \leq t_{n1} \leq \ldots \leq t_{ns} \leq t_0 < t_{ns+1} \leq \ldots \leq t_{nn} \leq 1$ be given. Moreover, let $q := s/n$. Then we have

$$g_{t_0}(t_{ni}) = g_q\left(\frac{i-1}{n-1}\right), \ i = 1,\ldots,n.$$

Thus instead to analyze the alternative g_{t_0} and an arbitrary design with s design points equal to or less than t_0 we can analyze the alternative g_q with the change-point $q = s/n$ and a uniform design.

Proof (of Theorem 4). By the above considerations the limit process of the recursive residual partial sum process with respect to the local alternative $\frac{1}{\sqrt{n-d}} g_q$ and a uniform design coincides with the limit process with respect to the local alternative $\frac{1}{\sqrt{n-d}} g_{t_0}$ and an asymptotic q-design. The trend h given in Theorem 2 part b) can be obtained for the local alternative $\frac{1}{\sqrt{n-d}} g_q$ and a uniform design after some calculations:

$$h(z) = 0, \qquad\qquad\qquad z \in [0,q],$$
$$h(z) = \int_q^z c_1 - \frac{1}{s}(qc_0 + (s-q)c_1)ds = q(c_1 - c_0)(\ln(z) - \ln(q)), \ z \in (q,1].$$

Proof (of Theorem 5). For $z = 1$ the expression $-q(\ln(z) - \ln(q))\mathbf{1}_{(q,1]}(z)$ considered in (6) takes on its minimum for $q = e^{-1}$ and, furthermore, it is strictly increasing on $[e^{-1},1)$.

Let $e^{-1} \leq q_1 < q_2 < 1$. Then we have for all $z \in (q_2,1]$

$$\frac{d}{dz}(-q_1(\ln(z) - \ln(q_1))) = -\frac{q_1}{z} > -\frac{q_2}{z} = \frac{d}{dz}(-q(\ln(z) - \ln(q))).$$

This together with the first result of the proof provides the statement of Theorem 5.

References

1. Bischoff, W.: A functional central limit theorem for regression models. Ann. Statist. **26**, 1398–1410 (1998)
2. Bischoff, W., Gegg, A.: Partial sum process to check regression models with multiple correlated response: with an application for testing a change point in profile data. J. Multivariate Anal. **102**, 281–291 (2011)
3. Bischoff, W., Miller, F.: Asymptotically optimal tests and optimal designs for testing the mean in regression models with applications to change-point problems. Ann. Inst. Statist. Math. **52**, 658–679 (2000)
4. Bischoff, W., Somayasa, W.: The limit of the partial sums process of spatial least squares residuals. J. Multivariate Anal. **100**, 2167–2177 (2009)
5. Brown, R.L., Durbin, J., Evans, J.M.: Techniques for testing the constancy of regression relationships over time. J. Roy. Statist. Soc. Ser. B **37**, 149–192 (1975)
6. Bruss, F.T.: A unified approach to a class of best choice problems with an unknown number of options. Ann. Probab. **12**, 882–889 (1984)
7. Farebrother, R.W.: An historical note on recursive residuals. J. Roy. Statist. Soc. Ser. B **40**, 373–375 (1978)
8. Gardner, L.A.: On detecting changes in the mean of normal variates. Ann. Math. Statist. **40**, 116–126 (1969)
9. Jandhyala, V.K., MacNeill, I.B.: Residual partial sum limit process for regression models with applications to detecting parameter changes at unknown times. Stochastic Process. Appl. **33**, 309–323 (1989)
10. Jandhyala, V.K., MacNeill, I.B.: Tests for parameter changes at unknown times in linear regression models. J. Statist. Plann. Inference **27**, 291–316 (1991)
11. Jandhyala, V.K., MacNeill, I.B.: Iterated partial sum sequences of regression residuals and tests for changepoints with continuity constraints. J. Roy. Statist. Soc. Ser. B **59**, 147-156 (1997)
12. Jandhyala, V.K., Zacks, S., El-Shaarawi, A.H.: Change–point methods and their applications: Contributions of Ian MacNeill. Environmetrics **10**, 657–676 (1999)
13. MacNeill, I.B.: Properties of sequences of partial sums of polynomial regression residuals with applications to tests for change of regression at unknown times. Ann. Statist. **6**, 422–433 (1978)
14. MacNeill, I.B.: Limit processes for sequences of partial sums of regression residuals. Ann. Probab. **6**, 695–698 (1978)
15. Rabovski, O.: Asymptotische Tests basierend auf rekursiven Residuen von Regressionsmodellen, Diplomarbeit (2003)
16. Shorack, G.R.: Probability for Statisticans. Springer, New York [u.a.] (2000)
17. Sen, P.K.: Invariance principles for recursive residuals. Ann. Statist. **10**, 307–312 (1982)
18. Sen, A., Srivastava, M.S.: On tests for detecting change in mean when variance is unknown. Ann. Inst. Statist. Math. **27**, 479–486 (1975)
19. Watson, G.S.: Detecting a change in the itercept in multiple regression. Statist. Probab. Lett. **23**, 69–72 (1995)
20. Xie, L., MacNeill, I.B.: Spatial residual processes and boundary detection. South African Statist. J. **40**, 33–53 (2006)