# Discussion: Clinical measurements of periodontitis

**Edith C. Morrison and Charles J. Kowalski**

National Institute of Dental Research, National Institutes of Health, Bethesda, Maryland USA and Dental Research Institute, School of Dentistry, University of Michigan, Ann Arbor, Michigan, USA

My comments will be directed at the use of sequential linear regression analysis by Goodson and colleagues to monitor patients with periodontal disease. When using *any* statistical procedure, it is important to be reasonably confident that the underlying model is at least relatively realistic and that its attendant assumptions are satisfied. When using SLR on time series data, the model is

$$Y_t = \alpha + \beta t + \varepsilon_t$$

and the assumptions are that the $\varepsilon_t$ are *independently* normally distributed with mean zero and (common) variance $\sigma^2$, $\varepsilon_t \sim \text{NID}(0,\sigma^2)$.

Focusing first on the structure of the model itself (Fig. 1), to test $H:\beta=0$, is to test no change versus linear change and thus, according to the model, to reject H is to conclude that linear change has occurred, i.e., progressive change at a constant rate. *To the extent that the data fit a line, the data are giving testimony favoring the slow, continuous disease hypothesis.*

To test the "random burst" hypothesis, we would like a procedure that would test no change versus change in the level of a stationary time series (Fig. 2), but this is *not* consistent with the choice of the SLR model. If one believes in the random burst hypothesis, choosing the SLR model is choosing a model *believed to be inappropriate*. Indeed, the SLR model would be the model of choice only if one believed that periodontal disease was a continuous, progressive process. Thus without questioning any of the assumptions of SLR, without considering the sequential application of SLR, without considering the choice of sites as experimental units, there is a problem: the SLR model is logically inconsistent with the random burst hypothesis.

Consider next the assumptions underlying the appropriate use of SLR. My remarks are confined to *independence* of the $\varepsilon_t$. If this assumption is not satisfied, if the $\varepsilon_t$ are autocorrelated, a dramatic increase in the type-I error will result. The use of polynomial regression on longitudinal observations is considered by (Hoel 1964). The SLR and TSR models receive careful attention in (Ostrom 1978).

We considered the behavior of SLR tests of $H:\beta=0$ in models when a known correlation structure is imposed on the error terms. In particular, we generated normally distributed random variables with constant variances (so as to satisfy the other requirements of SLR) and with the same means (so that $H:\beta=0$ is true) and so that the correlation between observations separated by one time point is $\rho$, that between observations two time units apart is $\rho^2$, etc. We counted the number of times H is rejected for various combinations of $\rho$ and $T$. In each case, we generated 1000 sets of observations and performed the tests at the 1% level of significance (the level used by Goodson et al. (1982)) and at the 5% level of significance (the more traditional level).

Specifically, for each value of $\rho$ considered, 1000 sets of observations

$$Y_1, Y_2, \ldots, Y_T$$

were generated where each $y_t$ has the same mean ($\mu=5$), the same variance ($\sigma^2=\sigma=1$) and a normal distribution. The results for $T=10$ and $\rho=0.1,0.2,\ldots,0.9$ are shown in Table 1. Also shown are the expected numbers of times that H would be rejected when $\rho=0$. When the 1% level of significance is used, we expect 10 type-I errors and when the 5% level of significance is used, we expect 50 type-I errors.

It is seen for $T=10$, even moderate values of $\rho$ cause pronounced increases in the type-I error rates. If $\rho=0.5$ and $T=10$, we see that 64 type-I errors are committed at the 1% level of significance, over 6 times the expected number.

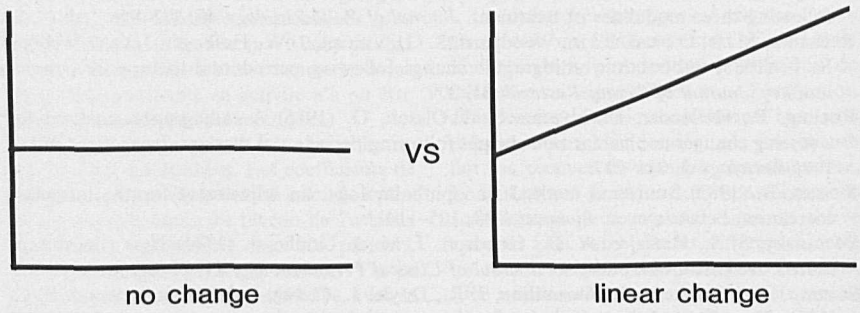


*Fig. 1.* Simple linear regression model.
*Das Modell einfacher linearer Regression.*
*Modèle de régression linéaire simple.*

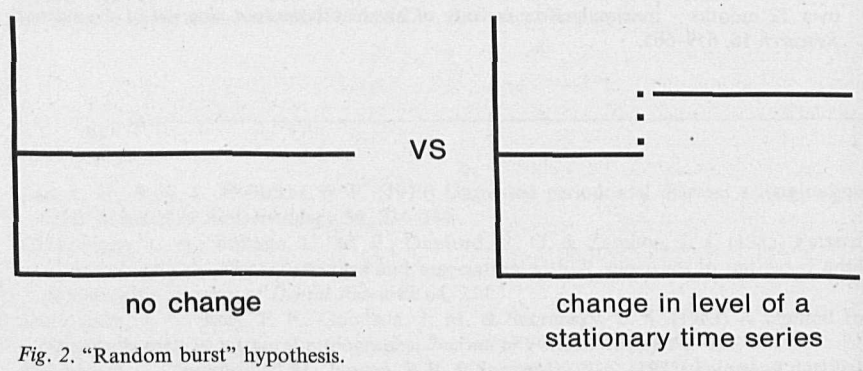


*Fig. 2.* "Random burst" hypothesis.
*Die Hypothese des "zufälligen Ausbruches".*
*Hypothèse du "flamboiement randomisé".*

*Table 1.* Numbers of type-I errors in 1000 trials committed for various values of $\rho$ when $T=10$

| $\rho$ | 0.05 | 0.01 |
|---|---|---|
| 0 | 50 | 10 |
| 0.1 | 58 | 13 |
| 0.2 | 89 | 27 |
| 0.3 | 130 | 35 |
| 0.4 | 152 | 48 |
| 0.5 | 193 | 64 |
| 0.6 | 257 | 126 |
| 0.7 | 326 | 167 |
| 0.8 | 403 | 231 |
| 0.9 | 486 | 334 |

Goodson (1982) is apparently aware of this problem but counters by saying that he checked on this assumption and found no autocorrelation in his measurements. While we were unable to check this on a comparable data set of our own, consider the following data (Figs. 3a, b) epitomizing the sort of data we might well expect under the random burst hypothesis.

This shows 2 things: first, that data consistent with the random burst hypothesis can show substantial autocorrelation and, second, that $r_1(y_t) \neq r_1(\hat{\varepsilon}_t)$.

The regression analyses carried out by Goodson et al. (1982) were performed on an iterative or sequential basis by starting with the first 3 measurements at each site and "increasing the number of points fitted until data from all appointments at a particular periodontal site were included." Using regression analysis in this manner – even if all the other assumptions concerning SLR are satisfied – again may dramatically increase the type-I error rate. The problem here is that repeated testing of the hypothesis, one time point being included at each stage, increases the chance that the hypothesis will be rejected. Simply stated, multiple testing provides more opportunity for error so that, again, more type-I errors are to be expected and the risk of erroneously declaring a site to be "active" is increased.

We also simulated the behavior of sequentially testing $H:\beta=0$ over a number of sites. Specifically, we generated 100 sets of observations and counted the number of times $H:\beta=0$ was rejected at at least one point in time at the 1% and/or 5% level of significance. We used only 100 sets of measurements since, for each set, the regression analysis was performed at each of the time points $T=3,4,...,10$ for a total of 800 regression analyses. Given 100 sets of observations, we would expect 5 type-I errors at 5% and 1 type-I error at 1%. The results are summarized in Table 2. "*" indicates significance at the 5% level and "**" at the 1% level. 21 of the 100 sites showed significance at the 5% level and 5 of these were also significant at the 1% level. These 21 sites are shown in Table 2 along with the results of the SLR test at each of the time points $T=3,4,...,10$. For example, no. 10 was significant at $T=4$ with $P<0.01$ and again at $T=6$ with $0.01<P<0.05$. Table 2 shows that we got 21 and 5, respectively, i.e., roughly 4–5 times the expected number.

In order to assess the combined effects of sequential SLR's and correlated errors, we simulated the situation in which 1000 sites were monitored over $T=10$ time points, the correlation between adjacent time points being $\rho=0.5$. That is, each of the 1000 sets of observations

$$Y_1, Y_2, ..., Y_{10}$$

were generated to be normally distributed, each having mean value 5, standard deviation 1 and such that the correlation between adjacent $y$'s was 0.5, those two time units apart having correlation $(0.5)^2=0.25$, etc. The tests were performed sequentially at $T=3,4,...,10$ and we found that 203 of the 1000 sites showed "change" at the 1% level of significance at at least one of these time

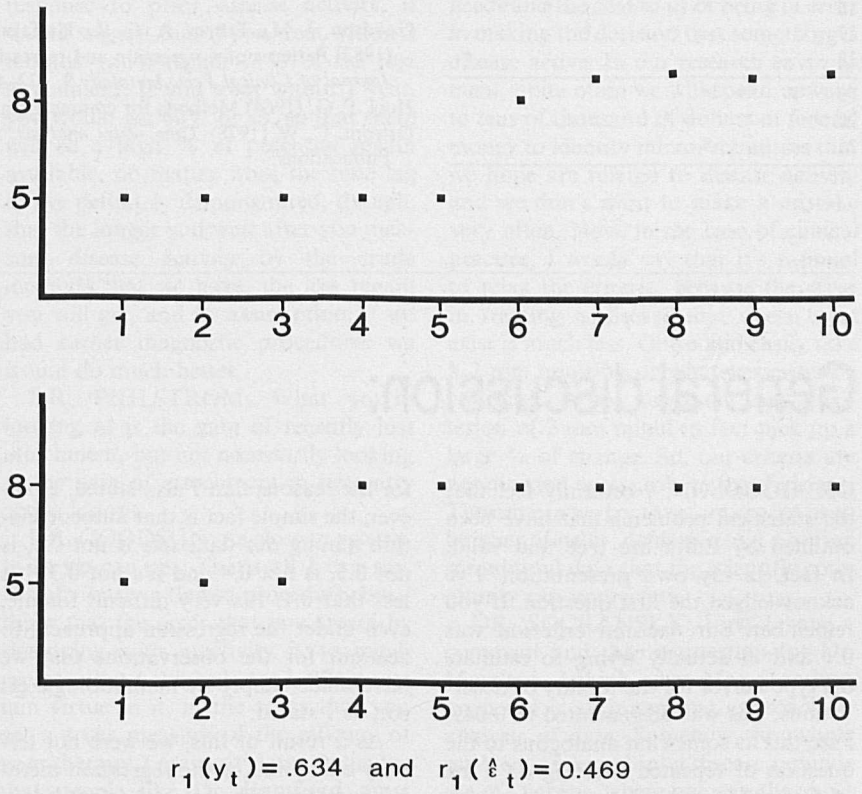


*Fig. 3.* Autocorrelation "random burst" hypothesis. The $y$-axis represents attachment level (mm). The $x$-axis is time. (a) The observations themselves have first-order autocorrelations $r_1=0.70$ while the residuals $(\hat{\varepsilon}_t)$ have $r_1=0.20$. (b) Note also that if the burst had occurred earlier (later), $r_1(\hat{\varepsilon}_t)$ would be higher $r_1=0.469$.

*Die Eigenkorrelation (autocorrelation) bei der "Hypothese des zufälligen Ausbruches". Die Ordinate bezeichnet das Attachmentniveau (mm). Auf der Abzisse ist die Zeit abgetragen. (a) Bei den Beobachtungen selbst ist eine Eigenkorrelation ersten Ranges von $r_1=0.70$ vorhanden, während bei den Resten (den Residualen $\hat{\varepsilon}_t$) eine $r_1=0.20$ vorliegt. (b) Beachten Sie ebenfalls, dass bei einem früheren (späteren) Ausbruch ein höherer Korrelationskoeffizient $r_1=0.469$ vorhanden gewesen wäre.*

*Autocorrélation de l'hypothèse du "flamboiement randomisé". L' ordonnée représente le niveau d'attache (mm), l'abscisse le temps. (a) Les observations elles-mêmes ont une autocorrélation de premier ordre $r_1=0.70$ tandis que les résidus $(\hat{\varepsilon}_t)$ ont $r_1=0.20$. (b) Noter également que si le flamboiement avait eu lieu plus tôt (plus tard) $r_1(\hat{\varepsilon}_t)$ serait plus élevé $r_1=0.469$.*

*Table 2.* Results of SLR tests at $T=3,4,...,10$ for the 21 of 100 sites which were significant at some point in time ($P<0.05$)

| Site no. | Time 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| 1 | * | | | | * | * | | |
| 2 | | | | * | | | | |
| 3 | | | | | * | * | | |
| 4 | | | | | | | * | * |
| 5 | ** | | | | | | | |
| 6 | | ** | | | | | | |
| 7 | | | | | | | | * |
| 8 | | | * | | | | * | |
| 9 | | | * | | | | | |
| 10 | | ** | | * | | | | |
| 11 | | | * | | | | | |
| 12 | * | | | | | | | |
| 13 | * | | | | | | | |
| 14 | | | * | | | | | |
| 15 | | | | * | | * | * | ** |
| 16 | | * | | | | * | * | |
| 17 | | | | | * | | | |
| 18 | | | | | | | | * |
| 19 | ** | * | | | | | | |
| 20 | | | * | | | | | |
| 21 | | * | | | | | | |

Table 3. Results of sequential SLR's and correlated errors in 1000 sites

| Time | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| $\alpha$ 1% | 15 | 16 | 27 | 43 | 65 | 63 | 75 | 78 |
| $\alpha$ 5% | 59 | 92 | 127 | 164 | 172 | 179 | 206 | 196 |

points. This 20.3% change is slightly greater than that reported by Goodson et al. (1982) (17.2%) and so the changes detected by them can be identified as type-I errors in the context of the simulation model employed above. Indeed, while we view $\rho=0.5$ to reflect but moderate correlation among the $\varepsilon_t$, we are even able to account for all the "changes" with the even smaller values of $\rho$ of 0.4 or 0.3. A summary of the test history of the 203 sites found to be significant at the 1% level is shown in Table 3 using the same format as employed in Table 2.

In summary, then, all of the changes detected by Goodson et al. (1982) can be accounted for in the context of a *no change* model in which the errors are moderately correlated ($\rho=0.4$ or 0.5) when sequential SLR's are used to monitor the sites over time.

## References

Goodson, J. M., Tanner, A. C. R., Haffajee, A. D., Sornberger, G. C. & Socransky, S. S. (1982) Patterns of progression and regression of advanced destructive periodontal disease. *Journal of Clinical Periodontology* **9**, 472–481.

Hoel, P. G. (1964) Methods for comparing growth type curves. *Biometrics* **20**, 859–872.

Ostrom, C. W. (1978) *Time series analysis: regression techniques.* Beverly Hills, USA: Sage Publications.

# General discussion:

DR. GOODSON: I certainly feel that the statistical problems that have been outlined by Edith are true and valid. In fact, in my own presentation, I've acknowledged the first question. If you remember, our decision criterion was 0.1 and in actually trying to estimate the type I error for the totality of observations, this was decremented to 0.049. I see this as somewhat analogous to the question of repeated *t*-testing on a set of data. This was not surprising to me. Actually, at the time we wrote the paper, we made no statement of what the overall sensitivity or type-I error rate might have been. All that we did was to list what our decision criteria were at each point. The second issue, however, the well-known autocorrelation, which I think certainly is a valid concern, something that we should be aware of, – I think Edith has very elegantly demonstrated that the effect of autocorrelation, particularly in her regression model, can have rather dramatic effects for the reasons that I also stated. However, the simple fact is that autocorrelation among our data sets is not 0.3, is not 0.5, is not 0.4 and it's not 0.3. It's less that 0.1. It's very difficult for me, even under the regression approach, to account for the observations that we have made simply by methodologic error, as I stated.

As a result of this, we were not terribly enamored of the regression method ourselves. That was the first paper we published on disease activity and, as I have acknowledged, there are several very serious shortcomings. From the clinical standpoint, one of the biggest shortcomings is the fact that you have to wait for at least three visits before you can even begin to detect a change, and the type II error seems to be much higher than what we would like it to be. Shortly after publication of that paper, we suggested two other methods which we do feel to be superior, and strangely, disease activity didn't disappear because of the fact that we started other methods. I think that, though I'm not enamored of the regression method myself, the results that we report with our regression method cannot be accounted for by the indicated two sources of error, because simply stated, the true biology of the system does not reflect that degree of autocorrelation. If in simulating a data set, a statistician were to assume that our standard deviation was, instead of 0.8 mm, something like 5.0 mm, we would never be able to make a measurement at all. I feel that the estimation of the parameters as it actually occurs in biological systems is relevant. However, I feel that the statements that Edith Morrison has made are completely valid and in making the statements, she has certainly encouraged our group to further investigate our methods to try to document the degree of error that we think exists.

DR. FLEISS: I think that both Dr. Goodson and his colleagues and Dr.