STATISTICS IN MEDICINE, VOL. 16, 373–384 (1997)

ASSESSMENT OF LACK OF FIT IN SIMPLE LINEAR REGRESSION: AN APPLICATION TO SEROLOGIC RESPONSE TO TREATMENT FOR SYPHILIS

M. SARAH ROSE AND GORDON H. FICK

Department of Community Health Sciences, Faculty of Medicine, University of Calgary, 3330 Hospital Drive N.W., Calgary, Alberta, T2N 4N1, Alberta, Canada

SUMMARY

A patient treated for infectious syphilis is cured when serologic tests become non-reactive, which may take years to achieve. Our objective is to develop a method to determine, within months, whether the patient has responded adequately to treatment. Previous research and our exploratory graphical analysis suggested that treatment response is linear when we applied logarithmic transformations of the axes. If the response to treatment is linear, titres recorded within the first few months of treatment will determine the slope of the line and one can develop an action line in future research. We used a non-parametric method to assess whether the logarithmic transformation improved the linearity and then we applied three different methods of testing lack of fit in linear regression. Based upon a sample size that reflects a clinically reasonable number of data points, the results of these tests provided no evidence against linearity.

1. INTRODUCTION

Patients treated for infectious syphilis have periodic follow-up by their physicians, often for more than two years, to assess whether they have adequately responded to treatment. There are two general types of serologic tests for syphilis: treponemal and non-treponemal. Assessment of serologic response to treatment entails non-treponemal tests. Non-treponemal tests provide a quantitative measure of antibody response to a substance, reagin, formed in the sera of patients with syphilis. The non-treponemal test is reported as non-reactive, reactive or weakly reactive. If the result is reactive, serial dilutions are carried out to quantitate the result. The result of the test, the titre, is reported as $1{:}n$ where n is the highest dilution at which the test remains reactive. A return to the non-reactive state is referred to as seroreversion.

A patient is considered cured when he/she reaches the non-reactive state, but this may take years to achieve, depending upon the stage of the disease and the magnitude of the initial titre.[1–7] Our overall objective was to develop a method to determine within a few months after treatment whether the patient had responded adequately to treatment or whether he/she needs to be retreated. Previous research has suggested that serologic response to treatment for syphilis, for patients classified as clinically cured, follows an exponential path.[8] Our initial exploratory graphical analysis suggested that with logarithmic transformations of the axes we could describe this response with a straight line. Assuming that the response to treatment is linear, titres recorded within the first few months of treatment determine the slope of the treatment response

CCC 0277–6715/97/040373–12

Received September 1995
Revised February 1996

line. This suggests the development of an action line which could be used to determine, with high probability, the need for retreatment based on the early response. Since the development of such an action line depends upon the assumption that the decline of titre is linear on a log-time scale, the first stage in the analysis is to assess this assumption. Thus, the objective of the analysis presented here is to provide a rigorous test of this assumption.

2. DATA

We used a subset of data from a study described by Romanowski *et al.*[1] In Alberta, during the years 1983–1985, there was a sudden sustained epidemic of infectious syphilis. Data were collected on all cases of syphilis and their contacts who were treated in Alberta during the years 1981–1987 inclusive. Each patient with infectious syphilis was treated and advised to return for follow-up serologic testing at varying intervals, usually at 3, 6 and 12 months, until either achievement of seroreversion or the physician's satisfaction with the patient's response. At each visit, the patient had one non-treponemal test, the rapid plasma reagin (RPR) test, and two treponemal tests. The study sample included patients at each stage of infectious syphilis (primary, secondary or early latent syphilis) either with a first or a repeat episode. Further details of the study sample and a descriptive analysis appear in Romanowski *et al.*[1] This prior analysis used life table methods to provide cumulative seroreversion rates by stage of disease, initial RPR titre and disease episode.

3. METHODS

For each individual with N pairs of observations, (x_i, y_i), $i = 1, \ldots, N$, we fitted the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (i = 1, \ldots, N) \tag{1}$$

where y_i is the response variable measured on the ith occasion as the logarithm to base 2 of the RPR titre, x_i is the logarithm (to base 10) of time since treatment on the ith occasion, and ε_i is the model error. We assume that the ε_i are random with mean zero and constant variance σ^2. We wish to assess whether this model provides an appropriate description of the response to treatment.

We generated two exploratory plots for each individual, the first plotted the logarithm (to base 2) of the titre against time, the second, log titre against the logarithm (to base 10) of time. Visual inspection of these plots suggested that there was indeed evidence of non-linearity for the plot of titre against time but that the logarithmic transformation of the time variable rectified this. To assess whether the logarithmic transform did in fact improve the linearity of the fit, we used an approximate F-test of linearity described by Hastie and Tibshirani and which involves the use of scatter plot smoothers.[9] As these authors pointed out, the exact distributional results are as yet unavailable for these tests, but they claim that simulation studies have suggested that these approximations are useful 'at least as rough guides' (p. 65). Because of these approximations and because the type of scatter plot smoother also affects the F-statistics, we chose to use this method to assess whether the linearity of the plot improved with the use of the logarithmic transformation (by comparing F-statistics) rather than to use it as an objective statistical test of non-linearity. We describe this test in Section 3.1.

In addition to assessment of this potential improvement we also wished to test in an objective way whether the linear model was an adequate representation of the response to treatment with use of the logarithmic transformation. Determination of the adequacy of a fitted regression model

is called a test for lack of fit (LOF).[10,11] Joglekar *et al.*[12] provide a review of recent developments in this field. To test whether the apparent linearity of the response stands up to objective statistical testing, we chose three different methods to compare for LOF in the hypothesized model: the near-neighbour approach of Neill and Johnson;[13,14] the groupwise regression approach of Breiman and Meisal;[15] and the test of Utts.[16] We describe these methods in Sections 3.2 to 3.5. The advantage of the tests we describe here is that each employs standard statistical software packages.

3.1. Assessment of the improvement in linearity using the logarithmic transformation of time

A scatter plot smoother is a tool to summarize the trend of a response measurement Y as a function of an explanatory variable X. Unlike simple linear regression, the scatter plot smoother does not assume a rigid form for the dependence of Y on X, thus it is often referred to as a non-parametric regression.[9] Using the same notation as for the linear regression in (1) we describe the model using the scatter plot smoother as

$$y_i = f(x_i) + \varepsilon_i \qquad (i = 1, \ldots, N) \tag{2}$$

where f is some unknown and arbitrary function of the predictor variable X and the ε_i represent zero mean, independently distributed errors. A detailed account of scatter plot smoothers appears in Hastie and Tibshirani.[9] We used a cubic smoothing spline to test whether the linearity of the plot improved with the use of the logarithmic transformation. Generally speaking, the cubic spline is a composite function that is a piecewise cubic polynomial defined on several regions joined by a sequence of knots (or breakpoints). These cubic polynomials are constrained to be continuous and to join smoothly at the knots. A natural cubic spline is a cubic spline with the additional constraint that the function is linear beyond the boundary points.

Specifically, the cubic smoothing spline fit to our data is that function $f(x)$, with continuous first and integrable second derivatives, that minimizes the penalized residual sum of squares

$$\sum_{i=1}^{N} (y_i - f(x_i)) + \lambda \int_a^b (f''(t))^2 \, \mathrm{d}t$$

where λ is a fixed constant (the smoothing parameter) and $a \leqslant x_1 \leqslant \ldots \leqslant x_n \leqslant b$. The parameter λ controls the amount of smoothing which is applied to the fit; the smaller the value of λ, the more wiggly the resulting function is. Conversely, larger values of λ result in smoother functions and as $\lambda \to \infty$ the solution is the least-squares line. There exists an explicit unique solution to this minimization, which is a natural spline with interior and boundary knots at the values of x_i; $i = 1, \ldots, N$. The equivalent degrees of freedom for a smooth fit are inversely related to the smoothing parameter, and should always be greater than 1 with 1 implying a linear fit. The smoothers (dashed line) applied to the data in Figures 1 and 2 are examples of the cubic smoothing spline.

To test the hypothesis that the regression is linear, we can use an approximate F-test that compares two smoothers by thinking of the least square regression line as an 'infinitely smooth' function where $f(x_i) = \beta_0 + \beta_1 x_i$ in (2).

Suppose we wish to compare two smooths $\hat{f}_1(x)$ and $\hat{f}_2(x)$ where $\hat{f}_1(x)$ is the linear regression function and $\hat{f}_2(x)$ is a non-parametric smooth. Let RSS_1 and RSS_2 be the residual sum-of-squares for the two models, and γ_1 and γ_2 the degrees of freedom in the linear and smooth fits,

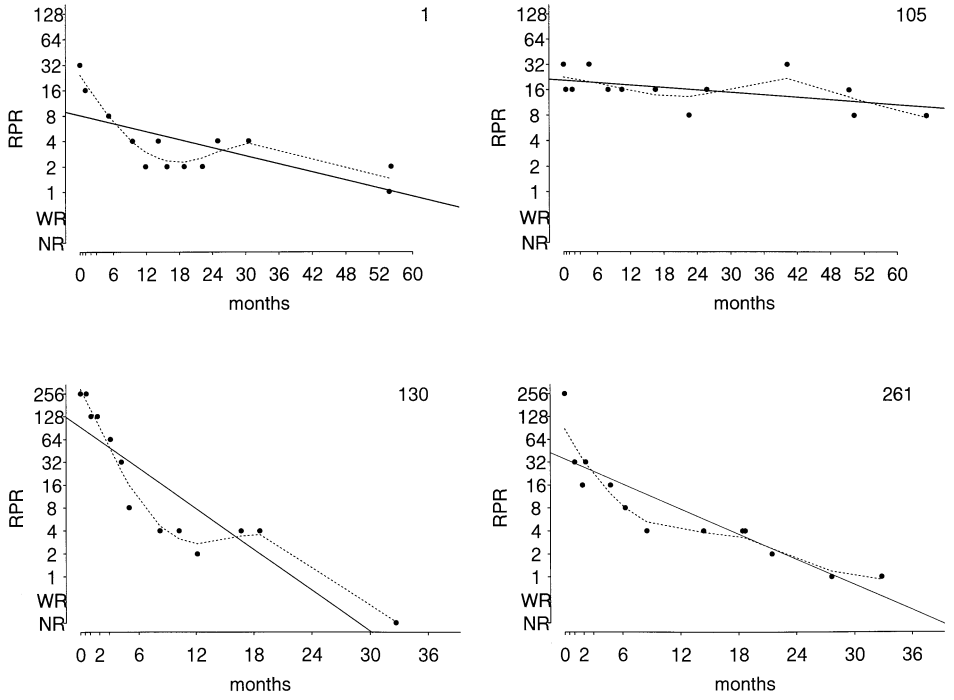

Figure 1. Serologic response to treatment as a function of time (linear) for four patients treated for infectious syphilis. The linear regression line is indicated by the solid line and the fitted scatter plot smoother (cubic spline) is indicated by the dashed line. The number in the upper right hand corner refers to the study number for comparison with Figure 2 and Tables I and II

respectively. We can test the hypothesis that the regression is linear by comparing the residual sum-of-squares for the two models. The statistic for this approximate F-test[9] is

$$F = \frac{(\mathrm{RSS}_1 - \mathrm{RSS}_2)/(\gamma_2 - \gamma_1)}{\mathrm{RSS}_2/(N - \gamma_2)}$$

with $\gamma_2 - \gamma_1$ and $N - \gamma_2$ degrees of freedom.

3.2. Assessment of the lack of fit of the linear regression model

The classical LOF procedure[10,11] described below is well known in situations where samples have replicate measurements in the space of predictor variables. If the regression model is incorrect, the residual mean square will tend to be inflated and will not provide a satisfactory measure of the random variation in the observations. If there is a prior estimate of σ^2 available and the residual mean square is significantly greater than this prior estimate, then this provides evidence of a lack of fit of the model to the data. Alternatively, when there are replicate measurements of Y available, one can use these to obtain an estimate of σ^2.[10]

Joglekar *et al.*[12] and Neill and Johnson[13] provided detailed reviews of methods developed in recent years to test for LOF when neither of these two options is available. Joglekar *et al.* classified the different approaches to testing for LOF as follows: the near-neighbour approach; the groupwise-regression approach; the checkpoint method; and 'others'.[12] As explained by

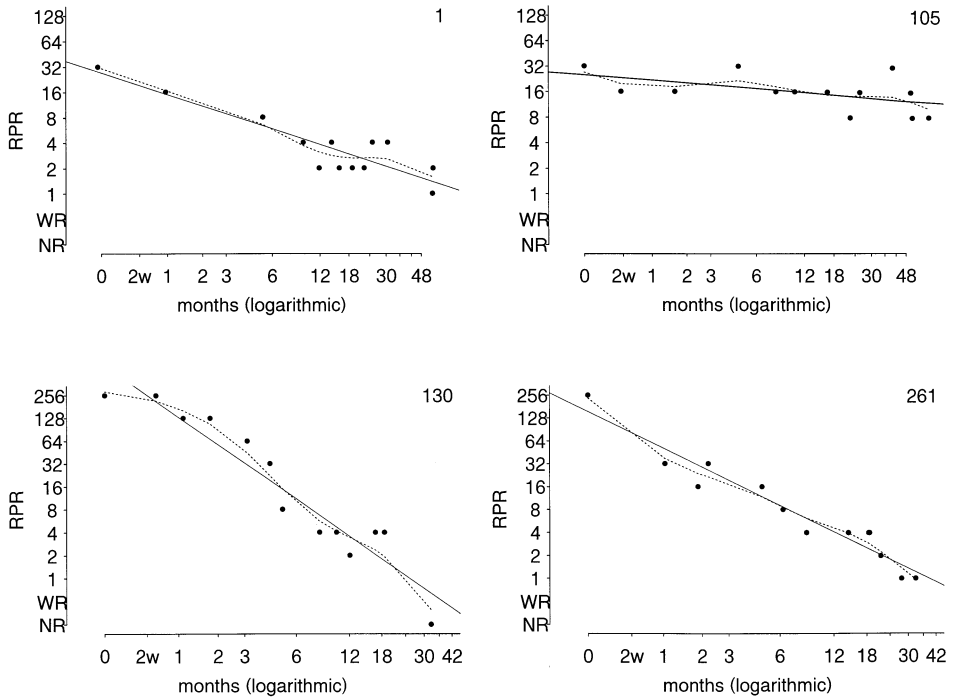

Figure 2. Serologic response to treatment as a function of time (logarithmic) for four patients treated for infectious syphilis. The linear regression line is indicated by the solid line and the fitted scatter plot smoother (cubic spline) is indicated by the dashed line. The number in the upper right hand corner refers to the study number for comparison with Figure 1 and Tables I and II

Joglekar *et al.*, in the near-neighbour approach one obtains an estimate of σ^2 by grouping the 'near' observations in the space of predictor variables (pseudoreplicates or near neighbours). The rationale of the groupwise regression approach is one can well approximate the true relationship between Y and the predictor variables by piecewise polynomial approximation and thus use piecewise or groupwise regression to obtain an estimate of σ^2.

By the nature of our data, replication was not possible and our sample sizes were very small ($N \leqslant 13$ for any one individual) so that we were limited in our choice of methods. We chose from the near-neighbour category a test developed by Neill and Johnson,[14] from the groupwise-regression approach a method developed by Breiman and Meisal,[15] and from the other category a test suggested by Utts[16] that does not require replicates or near-replicates. For a full review of tests for LOF, we refer the reader to Neill and Johnson and Joglekar *et al.*[12,13]

The checkpoint approach concerns a testing of the predictive ability of the regression model. We did not use the checkpoint method to assess the fit of our model, since we took care not to extend the estimated regression coefficients beyond the range of the observed data. In this article we assess the LOF of the straight line over the time interval for data collection. In a subsequent analysis[17] we assess the usefulness of the 'first year slope' (that is, the slope of the serologic response for the first year post-treatment) by including this as a predictor variable in a proportional hazards model whose dependent variable was time to seroreversion. For patients who seroreverted, we recorded the time to seroreversion as the midpoint of the interval between the last reactive and the first non-reactive test results.

3.3. The method of Neill and Johnson

Neill and Johnson provide a review of procedures to assess the adequacy of a proposed regression model in the case of non-replication.[13] They indicate that the power of tests based on a pseudo pure error estimator of the error variance (including Utts[16]) may be adversely affected since the proposed estimators are biased under the hypothesis and/or the alternatives.

Neill and Johnson generalized the pure error lack of fit test to accommodate the case of non-replication.[14] Using a pseudo pure error estimator, which is consistent whether or not the specified model is correct, they developed a test statistic for model adequacy. They showed that this test statistic is asymptotically equivalent to the pure error LOF test with replication. The test is general in the sense that one need not specify the alternative except for power calculations. In this paper we describe this test as presented by Neill and Johnson, but in terms of the simple linear model. It is helpful to first recall the classical LOF test. A representation of simple linear regression with replication is

$$y_{ik} = \beta_0 + \beta_1 x_i + \varepsilon_{ik} \tag{3}$$

where $i = 1, 2, \ldots, M$, $k = 1, 2, \ldots, n_i$ and $n_i > 1$ for at least one i. For simplicity we limit our discussion to the case when $n_i = n$ for all $i = 1, \ldots, M$. Thus the total number of observations in the sample is $N = Mn$. We assume that the random errors ε_{ik} are independent and identically distributed with $E(\varepsilon_{ik}) = 0$ and $E(\varepsilon_{ik}^2) = \sigma^2$. The ratio

$$F = \frac{\mathrm{MS_{LOF}}}{\mathrm{MS_{PE}}} = \frac{\mathrm{SS_{LOF}}/M - 2}{\mathrm{SS_{PE}}/N - M} \tag{4}$$

where $\mathrm{MS_{LOF}}$ is the lack of fit mean square and $\mathrm{MS_{PE}}$ is the pure error mean square, follows an F-distribution with $M - 2$ and $N - M$ degrees of freedom.

A representation of simple linear regression without replication is

$$y_{ik} = \beta_0 + \beta_1 x_{ik} + \varepsilon_{ik} \tag{5}$$

where $i = 1, 2, \ldots, M$, $k = 1, 2, \ldots, n_i$. As above, we consider only the case where $n_i = n$ for $i = 1, \ldots, M$. Note that the only difference between the two models (3) and (5) is the additional subscript k in model (5) that allows for non-replication, that is, we group the data into M mutually exclusive (near-neighbour) groups.

Let x_{ik} denote the regressor variable for the kth observation in the ith group. We consider that x_{ik} is of the form $x_i + \delta_{ij}$ where x_i and δ_{ik} are fixed observable real numbers. We can think of δ_{ik} as the perturbation of the regressor variable for the kth observation in the ith group.

Let $\bar{X} = (x_1, \ldots, x_M)$ and $\Delta = (\delta_{ik})$. We naturally chose x_i as the mean of the ith group and hence $\Delta = (x_{ik} - \bar{x})$ is the vector of deviations of the kth observation in the ith group from the mean of the group. Suppose model (5) is correct and let $Y^* = Y - \Delta\beta$. If Y^* were observable, it would conform to the model with replication as given in (3). Then we could assess model adequacy of the usual LOF test with Y replaced by Y^*. Since Y^* is not observable, Neill and Johnson suggest its replacement with an observable vector

$$\hat{Y}^* = Y - \Delta\hat{\beta}$$

where $\hat{\beta}$ is the least squares estimate of ß under model (5). Neill and Johnson show that under certain conditions $\hat{Y}^*$ is asymptotically equivalent to Y. We replace Y in (5) by $\hat{Y}^*$, compute a lack of fit F-statistic as in (4), and denote this $\hat{F}^*$. We claim evidence against model adequacy if the observed value of $\hat{F}^*$ exceeds $F_{1-\alpha, N-p, N-M}$.

Neill and Johnson show that $\hat{F}^*$ is asymptotically equivalent, under general alternatives, to the test statistic obtained when replication actually exists.[14] Neill and Johnson claim that a simulation study suggests that this test is useful for small sample sizes. The simulation study included extreme cases in which the points were not near replicates but were uniformly separated. We implemented this test by grouping the regressor variable into adjacent pairs and by letting x_i be the mean of the ith group.

3.4. The method of Breiman and Meisal

Breiman and Meisal proposed a groupwise regression approach in which they developed a data-splitting algorithm to form the groups.[15] We adapted this method to deal with our small sample sizes and we review it here, using the terminology of Joglekar *et al.*,[12] in terms of simple linear regression.

Given the data set R that consists of N points (x_i, y_i), $i = 1, \ldots, N$, we fit the simple linear regression model for which we wish to test lack of fit. Let SSE denote the residual sum of squares for this regression model. We then subdivide R into two subregions R_1 and R_2. In each of these two subregions we fit a separate linear least squares regression and let SSE_j denote the residual sum of squares for subregion j; $j = 1, 2$. The idea behind this test is that if the true relationship is strongly non-linear in R, then we obtain a much better fit by conducting a separate linear regression model in each of the two subregions. Breiman and Meisal suggest using the test statistic

$$F = \frac{(\mathrm{SSE} - \mathrm{SSE}_1 - \mathrm{SSE}_2)/3}{(\mathrm{SSE}_1 + \mathrm{SSE}_2)/N - 6}$$

to test the significance of the reduction in the residual sum of squares due to the splitting, and which follows an F-distribution with $(3, N - 6)$ d.f. when the true relationship is linear in R. They point out, however, that even if the true relationship is strongly non-linear over R, the random subdivision of R into R_1 and R_2 does not necessarily produce a significantly better fit. Hence, they recommend repetition of this process, at most K times, each time randomly choosing a new vector to split R. In their simulation work they used a significance level of 0·01 and K equal to 5.

Breiman and Meisal recommend no further splitting of a subregion if it contains 6 or fewer points for simple linear regression. They also recommend division in half for points in any one region, or as close as possible to one half. We decided to investigate lack of fit in the linear model only for those cases where there were 8 or more data points and when our splitting of these subregions would yield subregions of 4 or more data points. Although we should have conducted the subdivision by choosing a random vector, we had severe limitations of our sample sizes. The maximum number of points for any one individual was 13. Because of the above restrictions this meant that the maximum number of possible vectors, m, was no more than 5. For each individual we therefore used all of the m possible vectors. If we find any one of the tests significant, then we will consider this evidence of lack of fit of the line for this individual.

Although our analysis stopped at this point, the Breiman and Meisal procedure extends the above procedure by splitting the above region into the two subregions if any one of the tests is significant. One then repeats the procedure within the subregions until one produces k 'terminal' subregions $R_1, \ldots, R_k$. A region becomes terminal if it either has (i) six or fewer points or (ii) more than six points but none of the attempted splittings produces a significant reduction in the residual sum of squares. If SSE_j is the residual sum of squares for subregion j, Breiman and Meisal suggest use of

$$\mathrm{BM_w} = (1/k) \sum_{j=1}^{k} \mathrm{SSE}_j$$

as an estimate of σ^2, which we can then compare to the estimate of σ^2 obtained from the fitted model. As pointed out by Breiman and Meisal[15] and by Joglekar *et al.*,[12] however, 'the goodness of fit of the fitted model will have to be decided by a subjective decision on the part of the experimenter' since the distributional properties of the test statistic are hard to obtain.

3.5. The Rainbow test for lack of fit in regression

Utts[16] suggests an LOF test, that does not require replicates or a prior estimate of error. This 'rainbow' test is based on comparison of a fit over low leverage points with the fit over the entire data set. As before, we describe the test with the author's notation but in terms of simple linear regression.

Suppose that we fit the model as in (1) but that the correct model includes additional terms that can be expressed as

$$y_i = \beta_0 + \beta_1 x_i + \sum_{h=1}^{q} \theta_h w_{hi} + \varepsilon_i$$

where the $\boldsymbol{\theta} = (\theta_h)$ is a $q \times 1$ vector of unknown parameters and the w_{hi} are fixed observable real numbers. We can write the rainbow test for lack of fit as the test of $H_0: \boldsymbol{\theta} = \mathbf{0}$ versus $H_1: \boldsymbol{\theta} \neq \mathbf{0}$. The test is general since one need not specify either $\mathbf{w} = (w_{hj})$ or $\boldsymbol{\theta}$ to carry out the test.

Let $\mathrm{SSE}_{\mathrm{FULL}}$ be the error sum of squares from fitting model (1) to the entire data set. We then form a subset of the data that consists only of the data points that have low leverage. Let m be the number of points in the central region. Let $\mathrm{SSE}_{\mathrm{CENTRAL}}$ be the residual sum of squares obtained from fitting model (1) to these m observations. The numerator for the test is based on the difference between these, that is, $\mathrm{SSE}_{\mathrm{LOF}} = \mathrm{SSE}_{\mathrm{FULL}} - \mathrm{SSE}_{\mathrm{CENTRAL}}$. Utts defined the test statistic as

$$F = \frac{\mathrm{SSE}_{\mathrm{LOF}}/N - m}{\mathrm{SSE}_{\mathrm{CENTRAL}}/m - 2}.$$

Utts[16] showed that this test statistic has a double non-central F-distribution under H_1 and a central F-distribution under H_0.

Utts recommends using about half the data points in the central region, since this will approximately minimize the variance of F under H_0 and will provide some robustness if a few outliers are present in either region. To implement this test, we took points with leverage less than or equal to the median of the diagonal elements of the hat matrix to form the central region. Utts recommends that for larger data sets, points with leverage less than $2/n$ could be used. She also recommends that evidence of non-linearity should be followed by a more detailed investigation of the data, since rejection of the null hypothesis could also be caused by incorrect assumptions (such as normally distributed errors and homoscedasticity) rather than lack of fit. In particular she presents the results of a simulation study which show that the rainbow test has an inflated α level when the errors are $\mathrm{N}(0, x^2)$ (extreme heteroscedasticity) or when they are log-normally distributed. The test, however, behaves well when the errors are normally distributed (N(0, 1)) or mildly contaminated (N(0, 1) with probability 0·9 and N(0, 9) with probability 0·1.)

4. RESULTS AND DISCUSSION

We have provided in Figure 1 an illustration of the serologic response to treatment for syphilis for four patients, as a function of time, and in Figure 2 as a function of the logarithm of time. We chose these four patients simply because they had the maximum number (13) of data points. The

Table I. Approximate F-statistics to compare linearity before and after transformation

Study number	Number of points	Untransformed data F	Untransformed data p	Transformed data F	Transformed data p
1	13	13·81	0·002	1·37	0·320
105	13	2·00	0·193	0·99	0·445
130	13	28·55	$< 0·001$	3·56	0·067
261	13	4·86	0·033	2·66	0·119
11	10	4.52	0·046	5·20	0·034
231	10	9·97	0·01	11·76	0·006
27	10	18·39	0·004	1·10	0·430
34	10	13·02	0·008	3·63	0·100
70	10	8·04	0·023	2·81	0·148
264	10	23·76	0·002	15·39	0·006
364	10	2·79	0·149	297·79	$< 0·001$
17	9	7·63	0·039	1·19	0·419
104	9	21·77	0·006	3·61	0·124
133	9	11·90	0·018	8·52	0·033
151	9	4·38	0·094	2·02	0·254
193	9	4·89	0·080	1·52	0·339
228	9	12·74	0·016	1·20	0·416
253	9	1·92	0·268	3·12	0·150
255	9	1·08	0·452	1·23	0·408
96	8	3·84	0·149	4·20	0·135
2	8	119·37	0·001	31·75	0·009
9	8	20·43	0·017	2·79	0·211
12	8	148·63	$< 0·001$	290·21	$< 0·001$
14	8	8·93	0·051	1·04	0·488
24	8	1·69	0·338	1·28	0·422
30	8	1·86	0·311	0·52	0·696
75	8	7·60	0·065	2·28	0·258
92	8	7·33	0·068	2·43	0·242
100	8	3·13	0·187	1·84	0·314
134	8	12·95	0·032	2·72	0·216
225	8	3·57	0·162	5·07	0·108
251	8	6·43	0·080	2·98	0·197
259	8	4·27	0·132	0·30	0·824
283	8	9·22	0·050	3·31	0·176
311	8	4·69	0·118	5·75	0·092
318	8	13·82	0·029	0·66	0·627
330	8	4·74	0·117	0·55	0·682

linear regression line is the solid line and the fitted scatter plot smoother (cubic spline) the dashed line.

As noted above, we felt it necessary that an individual have at least $N = 8$ data points before we could realistically assess whether or not the regression line was linear. There were 37 individuals who met this criterion.

Table I shows comparisons of the approximate F-ratios for the linearity of the regression lines before and after the log transformation. The F-statistic has 3 and $N-5$ degrees of freedom, where N is the number of points for that individual. We have presented the p-values associated with these approximate tests but we emphasize that our use of these tests is to compare the evidence of non-linearity before and after application of the log transformation rather than to

assess significance *per se*. Our reasons for this are three-fold: (i) the tests are approximate and as yet the distributional results of the tests are unavailable; (ii) the degree of significance depends upon the type of smoother chosen, and (iii) the significance also depends upon the degree of smoothing used. By keeping the type of smoother used and the amount of smoothing the same, we can compare the tests for linearity before and after the logarithmic transformation of time within individuals. In 28 out of the 37 cases, the size of the F-ratio was smaller with the logarithmic transform of time. This indicates that the difference between the linear regression line and the non-linear spline is smaller (that is, the linear regression line provides a closer fit to the data) with the use of the logarithmic transformation of time. Using the 0·05 level of significance as a baseline for comparisons, rather than an absolute level to declare significant evidence against the hypothesis of linearity, we can see that in six cases there was evidence of non-linearity both before and after the transformation. In 17 cases there was no evidence of non-linearity either before or after. In one case, there was evidence of non-linearity after but not before and in 14 cases there was evidence of non-linearity before but not after. On the basis of these comparisons it appears that the logarithmic transformation does improve the linearity of the fit.

Having decided that the logarithmic transformation does not provide a better fit, in terms of linearity, we wished to assess whether this fit was adequate. Table II contains the results of the three tests described above to assess lack of fit after application of the logarithmic transformation. For the Breiman and Meisal test, subjects who had eight points had only one test performed in which the data were split into two groups of four points each. In this case the F-statistic and p-value for these tests are given. For subjects with more than eight points, however, we performed more than one test using a different split of the data each time. For example, if an individual had 13 points, we performed four different tests with the following four splits (5 + 8), (6 + 7) (7 + 6) (8 + 5). In this case, to be conservative we present the maximum F-statistic obtained. Breiman and Meisal recommend use of an adjusted significance level of α/m in the case of m different splits. In this case, however, we caution against such formal adjustment procedures, especially since these tests are not independent.

Perusal of Table II indicates that we have very little evidence of lack of fit of the straight line response. None of the cases provided consistently small p-values (that is $< 0{\cdot}05$) across all three tests and only two cases (264 and 12) provided small p-values for two out of the three tests. For both of these cases, visual inspection alone of the plots casts aspersions on our hypothesis. It is important, however, to view these two cases in the context of our multiple testing of the hypothesis either with our use of multiple comparison procedures or recognition of our expected error rate. We can view our data here as an opportunity to examine the null hypothesis that a straight line adequately describes the serologic response to treatment for syphilis in 37 independent sets of data. Thus, for each of the tests we have performed we can either impose a p-value of $0{\cdot}05/37 = 0{\cdot}00135$ to claim evidence of lack of fit for any one individual or recognize the expected error rate. That is, if the relationship is truly linear we can expect to find two cases that produce 'significant' evidence against the null hypothesis at the 5 per cent level. This is exactly our observation for both the Utts and Breiman and Meisal tests. For the Neill and Johnson test, we observed four cases with a p-value less than 0·05. None of these p-values, however, was below the 0·00135 Bonferroni-adjusted level.

Conversely, in the light of the small sample sizes typical with data of this type, we would find it helpful to consider whether we had the power to detect evidence against the null hypothesis. Unfortunately, this is not possible since, for each of the three tests used to assess lack of fit, we need to specify the alternative hypothesis to assess the power of the test and we lack an alternative hypothesis any more specific than the vague 'non-linear'. With our data, it is not feasible to

Table II. F-statistics and p-values for each of the three lack of fit tests (transformed data)

Study number	Number of points	Neill and Johnson		Breiman and Meisal		Utts	
		F	p	F	p	F	p
1	13	2·46	0·152	0·86	0·506	1·13	0·456
105	13	2·73	1·127	0·37	0·780	0·54	0·766
130	13	0·79	0·593	1·24	0·366	4·33	0·065
261	13	6·75	0·019	0·89	0·502	1·94	0·242
11	10	1·42	0·334	3·06	0·113	1·98	0·265
231	10	13·00	0·008	3·08	0·129	3·66	0·116
27	10	1·07	0·441	0·34	0·800	1·72	0·347
34	10	1·70	0·282	4·80	0·082	1·85	0·325
70	10	1·58	0·306	1·09	0·451	7·81	0·061
264	10	3·93	0·087	10·57	0·023	19·04	0·018
364	10	0·32	0·812	1·26	0·401	–*	–*
17	9	0·54	0·682	2·20	0·267	1·04	0·508
104	9	7·89	0·037	2·23	0·263	0·26	0·887
133	9	5·86	0·060	5·05	0·108	1·93	0·309
151	9	1·86	0·278	0·30	0·823	1·57	0·371
193	9	3·04	0·156	2·45	0·240	1·31	0·428
228	9	1·78	0·291	0·53	0·690	9·39	0·048
253	9	2·03	0·253	2·33	0·253	3·52	0·164
255	9	3·04	0·158	2·01	0·290	1·28	0·438
96	8	0·92	0·470	1·34	0·454	1·67	0·407
2	8	4·60	0·092	29·70	0·033	1·21	0·500
9	8	2·05	0·244	2·03	0·347	2·42	0·313
12	8	36·74	0·003	8·21	0·111	20·61	0·047
14	8	0·63	0·579	0·13	0·935	3·61	0·229
24	8	0·27	0·773	0·26	0·854	0·88	0·595
30	8	0·50	0·639	1·57	0·413	0·07	0·987
75	8	0·70	0·549	0·52	0·713	3·14	0·256
92	8	0·20	0·198	0·30	0·829	3·94	0·213
100	8	1·77	0·282	1·06	0·518	0·41	0·798
134	8	0·35	0·724	1·28	0·469	2·75	0·284
225	8	2·64	0·186	2·76	0·277	2·72	0·287
251	8	1·15	0·404	0·37	0·789	2·43	0·312
259	8	0·24	0·800	0·36	0·792	0·95	0·570
283	8	0·64	0·576	0·19	0·894	2·32	0·323
311	8	1·09	0·420	4·24	0·197	2·85	0·276
318	8	1·97	0·254	0·39	0·774	1·38	0·462
330	8	0·90	0·476	0·20	0·889	0·91	0·585

* Nine points had the same y-values, which made the calculation of the Utts statistic impossible

increase the number of data points for any one individual since it is unreasonable to request more frequent follow-up of patients for serology.

We acknowledge that our data are serial observations on each patient and therefore do not satisfy the independence assumption. The tests that we employed, however, are all approximations to varying (unknown) degrees, especially in the light of our small sample sizes. In particular, we used the Hastie and Tibshirani test primarily to compare the lack of fit before and after the log transformation. Although we cannot view the results of this test as 'accurate' in terms of the p-values, they did provide evidence of a change in the lack of fit of the straight line model with application of the log transformation.

In summary, based upon a sample size that reflects a clinically feasible number of data points, the results of these lack of fit tests provide no evidence to contradict our assumption that the decline of RPR titre in patients treated for syphilis is linear on a log-time scale.

ACKNOWLEDGEMENTS

This research was supported in part by the National Health and Welfare Research and Development program through a Ph.D. Training Fellowship and the University of Calgary through a William Davies Medical Sciences Scholarship.

REFERENCES

1. Romanowski, B., Sutherland, R., Fick, G. H., Mooney, D. and Love, E. 'Serological response to treatment of infectious syphilis', *Annals of Internal Medicine*, **114**, 1005–1009 (1991).
2. Fiumara, N. J. 'The treatment of seropositive primary syphilis: An evaluation of 196 patients', *Sexually Transmitted Diseases*, **4**, 92–95 (1977).
3. Fiumara, N. J. 'The treatment of secondary syphilis: An evaluation of 204 patients', *Sexually Transmitted Diseases*, **4**, 96–99 (1977).
4. Fiumara, N. J. 'Reinfection primary and secondary syphilis', *Sexually Transmitted Diseases*, **5**, 85–88 (1977).
5. Fiumara, N. J. 'Treatment of primary and secondary syphilis: Serologic response', *Journal of the American Medical Association*, **243**, 2500–2502 (1980).
6. Capinski, T. Z., Lebioda, J., Koalasa, B. and Budzanouska, E. 'Antibiotics in the treatment of early syphilis', *in* Luger, A. (ed), *Current Problems in Dermatology*: *vol II*. *Antibiotic Treatment of Venereal Diseases*, Karger, Basel, 1968.
7. Schroeter, A. L., Lucas, J. B., Price, E. V. and Falcone, V. H. 'Treatment for early syphilis and reactivity of serologic tests', *Journal of the American Medical Association*, **221**, 471–476 (1972).
8. Brown, S. T., Zaidi, A., Larsen, S. A. and Reynolds, G. H. 'Serological response to syphilis treatment: a new analysis of old data', *Journal of the American Medical Association*, **253**, 1296–1299 (1985).
9. Hastie, T. J. and Tibshirani, R. J. *Generalized Additive Models*, Chapman and Hall, London, 1990.
10. Draper, N. and Smith, H. *Applied Regression Analysis*, 2nd edn, Wiley, New York, 1981.
11. Fisher, R. A. 'The goodness of fit of regression formulae and the distribution of regression coefficients', *Journal of the Royal Statistical Society*, **85**, 597–612 (1922).
12. Joglekar, G., Schuenmeyer, J. H. and LaRiccia, V. 'Lack-of-fit testing when replicates are not available', *American Statistician*, **43**, 135–143 (1989).
13. Neill, J. W. and Johnson, D. E. 'Testing for lack-of-fit in regression: a review', *Communications in Statistics – Theory and Methods*, **13**, 485–511 (1984).
14. Neill, J. W. and Johnson, D. E. 'Testing linear regression function adequacy without replication', *Annals of Statistics*, **13**, 1482–1489 (1985).
15. Breiman, L. and Meisal, W. S. 'General estimates of the intrinsic variability of data in nonlinear regression models', *Journal of the American Statistical Association*, **7**, 301–307 (1976).
16. Utts, J. M. 'The rainbow test for lack of fit in regression', *Communications in Statistics – Theory and Methods*, **A11**, 2801–2815 (1982).
17. Rose, M. S., Fick, G. H., Romanowski, B. and Love, E. J. 'Serologic response to treatment for syphilis', Presented at the 9th International meeting of the International Society for Sexually Transmitted Disease Research, Banff, Canada, October 1991.

.