On the relation between initial value and slope

K. BYTH*

NHMRC Clinical Trials Centre, University of Sydney, Locked Bag 77, Camperdown, NSW 1450, Australia and Westmead Millennium Institute, Westmead Hospital, Westmead, NSW 2145, Australia
kbyth@ctc.usyd.edu.au

D. R. COX

Nuffield College, Oxford, OX1 1NF, UK

SUMMARY

Suppose measurements of a particular feature are collected at baseline and at a number of subsequent time points and that for each individual there is a roughly linear trend in time. This paper takes three approaches to testing whether there is a relation between the initial value and the slope. It also considers whether the initial value for an individual is a useful predictor of the slope for that individual. The problems are formulated in terms of regression models with random coefficients. The solutions are illustrated using data from an observational study of clinical correlates of disability and progression in Huntington’s disease.

Keywords: Huntington’s disease; Initial value; Longitudinal data; Regression model with random coefficients; Regression to the mean; Slope.

1. INTRODUCTION

Suppose that patients are measured for a particular feature at baseline ($t = 0$) and at a number of subsequent time points and that for each patient there is a roughly linear trend in time. Is there a relation between the initial value and the slope, and is the initial value for an individual a useful predictor of the slope for that individual? Any analysis must take account of the phenomenon of regression to the mean first noted by Galton (1869). For a recent account from a number of perspectives, see the series of papers edited by Senn (1997) in the issue of *Statistical Methods in Medical Research* devoted to this topic.

The present paper was motivated by a specific application to be described in Section 2. Three versions of the problem are given and formulated in terms of a regression model with random coefficients; see, for example, Laird and Ware (1982). A theoretical discussion more from first principles is outlined in Section 3 and the conclusions from the specific study are sketched in Section 4. Section 5 discusses some further developments.

The multilevel model approach to longitudinal data analysis is well known and our suggested solution implicit in its formulation; see, for example, Burton *et al.* (1998), Omar *et al.* (1999), Verbeke and Molenberghs (2000), or Leyland and Goldstein (2001). Some care, however, is needed in applying the approach. For example, we are unaware of papers in the medical statistical literature where such a model

*To whom correspondence should be addressed.
has been correctly used to address the questions posed above. Even when a multilevel model is properly fitted, it is easy to overlook the effect of regression to the mean in interpretation.

In this paper we identify three different formulations of the issue and indicate solutions which avoid the regression to the mean effect.

2. A STUDY OF HUNTINGTON’S DISEASE

The data used to illustrate the above ideas were collected as part of an observational study of clinical correlates of disability and progression in Huntington’s disease (Mahant et al., 2003). Patients with a definite diagnosis of Huntington’s disease were assessed at their initial and subsequent routine clinic visits for motor, cognitive, and functional impairment using the Unified Huntington’s Disease Rating Scale (UHDRS) (Huntington Study Group, 1996). We concentrate here on the total motor score (TMS) and the total functional capacity (TFC) of the UHDRS. The TMS can take integer values from 0 (normal) to 124 (maximally abnormal) while the TFC takes integer values between 0 (maximal disability) and 13 (normal).

Our study population consists of 83 patients who had more than one clinic visit and were followed up for at least 1 year at Westmead Hospital, a major Sydney teaching hospital and tertiary referral center. The median duration of follow-up was 5.2 years (interquartile range, 3.1–6.8 years). The median number of visits per patient was 9 (interquartile range, 5–12). The median baseline TMS and TFC scores were

Fig. 1. Profile plots of the TMS and TFC scores over time for 10 randomly selected subjects.
38 (interquartile range, 26–55) and 8 (interquartile range, 5–10), respectively. Figure 1 illustrates typical TMS and TFC profiles for a random selection of 10 patients.

3. THEORETICAL DISCUSSION

3.1 Formulation

For the $i$th patient ($i = 1, \ldots, k$), suppose observations $Y_{ij}$ ($j = 0, \ldots, r_i$) are available at times $t_{ij} \geq 0$, where $t_{i0} = 0$ defines the time origin for individual $i$ and is taken at the first observation. We take as an initial model

$$Y_{ij} = \mu_i + \gamma_i t_{ij} + \varepsilon_{ij}, \quad (3.1)$$

where the $\varepsilon_{ij}$ are uncorrelated random terms of zero mean and variance $\sigma^2_\varepsilon$, and $(\mu_i, \gamma_i)$ are, respectively, the expected value at $t_{ij} = 0$ and the slope for the $i$th individual. We write

$$\mu_i = \mu + \zeta_i, \quad \gamma_i = \gamma + \eta_i, \quad (3.2)$$

where $(\zeta_i, \eta_i)$ are random terms of zero mean and variances $\sigma^2_{\zeta}$ and $\sigma^2_{\eta}$, and $\text{cov}(\zeta_i, \eta_i) = \rho_{\zeta\eta}\sigma_{\zeta}\sigma_{\eta}$.

We now distinguish different versions of the problem sketched in Section 1. In the first, we are interested in the relation between the expected value $\mu_i$ at time zero and the slope $\gamma_i$. This relation is summarized in $\rho_{\zeta\eta}$ or by the regression coefficient $\beta_{\gamma\mu} = \rho_{\zeta\eta}\sigma_{\zeta}/\sigma_{\eta}$. (3.3)

Especially if the $\varepsilon_{ij}$ represent primarily measurement error or uninteresting ‘noise,’ the dependence of interest may be best encapsulated in $\beta_{\gamma\mu}$. Note that the conditional variance of $\eta_i$ given $\zeta_i$ is $\sigma^2_{\eta}(1 - \rho^2_{\zeta\eta})$.

A second possibility, directly relevant for empirical prediction, is to relate $\gamma_i$ to $Y_{i0}$. Under assumptions (3.1) and (3.2), the regression coefficient of $\gamma_i$ on $Y_{i0}$ is

$$\beta_{\gamma Y_0} = \beta_{\gamma\mu}(1 + \sigma^2_\varepsilon/\sigma^2_{\zeta})^{-1}. \quad (3.4)$$

A different approach to assessing dependence on the baseline value is to consider $Y_0$ purely as an explanatory variable.

3.2 Statistical analysis

The model (3.1)–(3.2) can be fitted by PROC MIXED in SAS or lme in R or S-PLUS, the defining parameters being estimated preferably by restricted maximum likelihood (REML). In particular, estimates of $\beta_{\gamma\mu}$ and $\beta_{\gamma Y_0}$ can be found and confidence intervals (CIs) obtained via the asymptotic standard errors. We now give a discussion from first principles which has the advantage of showing explicitly the relation with regression to the mean and with the contributions of individual patients.

For the analysis to study $\beta_{\gamma\mu}$, we start by fitting linear least square regressions to the data from each patient. For the $i$th patient this gives $\hat{\gamma}_i$ and $\hat{\mu}_i = \bar{Y}_i - \hat{\gamma}_i\bar{t}_i$. Direct evaluation under the model (3.1)–(3.2) gives the covariance matrix of $(\hat{\mu}_i, \hat{\gamma}_i)$ as

$$\begin{bmatrix}
\sigma^2_\varepsilon + \sigma^2_{\zeta}(1/r_i + \bar{t}_i^2/S_i) & \sigma_{\gamma\eta} - \sigma^2_{\zeta}\bar{t}_i/S_i \\
\sigma_{\gamma\eta} - \sigma^2_{\zeta}\bar{t}_i/S_i & \sigma^2_\eta + \sigma^2_{\zeta}/S_i
\end{bmatrix},$$

where $S_i = \sum_j (t_{ij} - \bar{t}_i)^2$ and $(\bar{Y}_i, \bar{t}_i)$ are averages for the $i$th individual. Note that if $\sigma_{\gamma\eta} = 0$, that is, if baseline mean and slope are uncorrelated, the estimated covariance is negative, a manifestation of regression to the mean.
A pooled estimate of the parameter $\sigma^2_e$ is obtained from the residual mean square after fitting separate regression lines for each patient. Various simple, if somewhat inefficient, estimates of the remaining parameters are now available. In particular, we may calculate the sums of squares and products about the mean of the individual ($\hat{\mu}_i$, $\hat{\gamma}_i$) and equate these to their expectations. This gives for the expectations of, respectively, the sum of squares of the $\hat{\mu}_i$, the $\hat{\gamma}_i$ and the sum of products of $\hat{\mu}_i$, $\hat{\gamma}_i$:

\[
(k - 1)\sigma^2_\epsilon + (k - 1)\sigma^2_e \sum_i (1/r_i + \bar{t}_i^2/S_i)/k, \tag{3.5}
\]

\[
(k - 1)\sigma^2_\epsilon + (k - 1)\sigma^2_e \sum_i 1/(kS_i), \tag{3.6}
\]

\[
(k - 1)\sigma^2_\eta - (k - 1)\sigma^2_e \sum_i \bar{t}_i/(kS_i). \tag{3.7}
\]

As noted above, the version of the problem in which the regression of slope on the observed baseline value is of interest can be approached via the above analysis, taking $\beta_\gamma \gamma_0$ in (3.4) as the primary parameter of interest. An alternative approach, which assumes less, is as follows. We use the baseline value as an explanatory variable, and hence to be treated conditionally when modeling the possibility that the baseline value, while a predictor of slope, is not derived from the system (3.1)–(3.2).

For example, when the baseline value is measured at diagnosis it may be subject to a bias that does not, however, affect its usefulness as a predictor of slope. Thus, we fit the following linear regression models to TMS and TFC. Because TFC takes only integer values over the narrow range $[0, 13]$, a suitably transformed version, namely

\[
\text{TFC}_{\text{trans}} = \log \frac{(0.5 + \text{TFC})}{(13.5 - \text{TFC})},
\]

The random variable $\delta_i$ in the model $\gamma_i = \gamma + \beta(y_i0 - \bar{y}_0) + \delta_i$ represents the variation in slope ($t > 0$) not accounted for by linear regression on $y_i0$. The sum of squares for estimating $\sigma^2_\delta = \text{var}(\delta_i)$ is the difference between the sums of squares for fitting constants in models (3.8) and (3.9). The expected value of this difference under the model

\[
y_{ij} = \mu_i + \gamma(y_i0 - \bar{y}_0) + \delta_i + \epsilon_i \ (t > 0), \tag{3.11}
\]

is of the form $A\sigma^2_\delta + (n_2 - 2)\sigma^2_e$, where for $S_i = \sum_j (t_{ij} - \bar{t}_i)^2$, ($t > 0$),

\[
A = \sum_i S_i - \frac{\sum (t_{ij} - \bar{t}_i)^4}{\sum_i S_i} - \frac{\sum (y_{i0} - \bar{y}_0)^2(t_{ij} - \bar{t}_i)^4}{\sum_i (y_{i0} - \bar{y}_0)^2S_i}. \tag{3.12}
\]

It is therefore possible to estimate $\sigma^2_\delta$ and $\sigma^2_e$ and to test the significance of the $\sigma^2_\delta$ term. Additional explanatory variables could be included in the regression term. One way to set out the calculations is in the form of an analysis of variance.

4. RESULTS FOR HUNTINGTON’S DISEASE STUDY

If normal error structures are assumed in (3.1)–(3.2), R or S-PLUS can be used to fit linear mixed effects models to TMS and TFC. Because TFC takes only integer values over the narrow range $[0, 13]$, a suitably transformed version, namely
Table 1. *Estimates of parameters in (3.1)–(3.4) obtained using REML for TMS, TFC, and TFC transformed = log \(\frac{(0.5 + \text{TFC})}{(13.5 - \text{TFC})}\)

| Parameter | TMS Estimate | 95% CI | TFC Estimate | 95% CI | TFC transformed Estimate | 95% CI |
|-----------|--------------|--------|--------------|--------|---------------------------|--------|
| \(\mu\)   | 41.71        | 37.18, 46.24 | 7.26         | 6.39, 8.13 | 0.33          | -0.01, 0.68 |
| \(\gamma\) | 3.66         | 3.16, 4.15  | -0.6         | -0.69, -0.50 | -0.25        | -0.28, -0.21 |
| \(\sigma_{\zeta}\) | 20.59 | 17.55, 24.17 | 3.66         | 3.08, 4.37  | 1.44         | 1.21, 1.71  |
| \(\sigma_\eta\) | 1.68         | 1.26, 2.25  | 0.29         | 0.22, 0.39  | 0.1          | 0.07, 0.14  |
| \(\sigma_\epsilon\) | 6.66         | 6.28, 7.07  | 1.47         | 1.39, 1.56  | 0.57         | 0.54, 0.61  |
| \(\rho_{\zeta \eta}\) | 0.06         | -0.22, 0.34 | -0.37        | -0.62, -0.04 | -0.22        | -0.52, 0.14 |
| \(\beta_{\gamma \mu}\) | 0.011        | -0.022, 0.043 | -0.049  | -0.087, -0.012 | -0.026    | -0.060, 0.008 |
| \(\beta_0\) | 0.009        | -0.019, 0.037 | -0.042    | -0.074, -0.011 | -0.022    | -0.051, 0.006 |

*All available patients used in each tabulated analysis.

Fig. 2. (a) Residuals in slope \(\eta_i\) versus residuals in intercept \(\zeta_i\) for the fitted TMS model (3.1)–(3.2). (b) Individual regression coefficients for TMS by time \((t > 0)\) versus baseline \((t = 0)\) value.

was also considered. The resulting estimates of parameters in (3.1)–(3.4) and their associated 95% CIs obtained using REML are shown in Table 1. The maximum likelihood estimates are virtually the same.

First consider the results for TMS. Since \(\hat{\rho}_{\zeta \eta}\) is here 0.06 with a 95% CI \((-0.23, 0.34)\), there is no evidence of association between the individual expected value \(\mu_i\) at time zero and the slope \(\gamma_i\). A plot of the residuals \(\eta_i\) versus \(\zeta_i\) for the fitted model (3.1)–(3.2) is shown in Figure 2(a). This plot suggests no
obvious departures from the underlying assumptions and appears consistent with the independence of $\eta_i$ and $\zeta_i$ and hence of the intercept $\mu_i$ and slope $\gamma_i$.

Similarly, there is no evidence of association between the baseline value of TMS and its rate of change over subsequent visits since $\hat{\beta}\gamma_0$ is 0.009, 95% CI $(-0.019, 0.037)$. The analysis of variance associated with (3.8)–(3.10) together with (3.12) yields estimates of 6.48 and 1.81 for $\sigma_\epsilon$ and $\sigma_\delta$, respectively. These are consistent with the earlier REML estimates for the system (3.1)–(3.2) ($t \geq 0$) given in Table 1. There the residual error estimate was 6.66, 95% CI (6.28, 7.07), and the estimated error for individual slopes was 1.68, 95% CI (1.26, 2.25). Figure 2(b) illustrates the individual TMS regression coefficients (for $t > 0$) plotted against the baseline value. This plot is consistent with the absence of appreciable association between these variables.

Now consider the TFC scores. Here $\hat{\rho}_{\zeta \eta}$ is $-0.37$, 95% CI $(-0.62, -0.04)$, and $\hat{\beta}\gamma_0$ is $-0.042$, 95% CI $(-0.074, -0.011)$. At first glance there seems to be reasonable evidence of a negative association between the initial value and the slope, that is, more rapid deterioration of TFC among those with better initial scores. On the transformed scale, however, $\hat{\rho}_{\zeta \eta}$ is $-0.22$, 95% CI $(-0.52, 0.14)$, and $\hat{\beta}\gamma_0$ is $-0.022$, 95% CI $(-0.051, 0.006)$, both not quite significant at the 5% level. The normal probability Q–Q plots of the residuals $\epsilon_{ij}$, $\zeta_i$, and $\eta_i$ obtained fitting model (3.1)–(3.2) to the raw and to the transformed TFC data are shown in Figure 3. There is some evidence of departure from normality which is not entirely removed by the transformation. This departure may be a result of ‘floor and ceiling effects’ due to the limited possible range of observations.

Figure 4(a) is a plot of the residuals $\eta_i$ versus $\zeta_i$ for the model (3.1)–(3.2) fitted to the raw TFC data. Note that patients with $\zeta_i < -3$ all have $\eta_i > 0$ while those with $\zeta_i > -3$ have, on average, negative $\eta_i$. 

![Fig. 3. Normal probability Q–Q plots of the residuals $\epsilon_{ij}$, $\zeta_i$, and $\eta_i$ obtained fitting model (3.1)–(3.2) to the raw and the transformed TFC data (all patients).](https://academic.oup.com/biostatistics/article-abstract/6/3/395/241336)
On the relation between initial value and slope

Fig. 4. Residuals in slope $\eta_i$ versus residuals in intercept $\zeta_i$ for model (3.1)–(3.2) fitted to (a) TFC (all patients) and (b) TFC transformed (omitting two profoundly ill patients with TFC $\equiv 0$).

Closer examination revealed that patients corresponding to points in the top left of Figure 4(a) had TFC scores $\leq 3$ at baseline and could deteriorate over time only to a score of zero, the lowest possible for TFC. In particular, there were two profoundly affected subjects who entered the study with zero TFC scores and remained at this level throughout (a total of 3.0 and 7.8 years). Such behavior would be expected clinically since Huntington’s disease is a chronic progressive illness. These patients correspond to the larger points on the far left of Figure 4(a).

Omission of these two patients from the analysis effectively normalized the residuals from the model of the transformed TFC scores, producing the scatterplot of $\eta_i$ versus $\zeta_i$ shown in Figure 4(b). On the transformed scale, the reduced data set yielded estimates for $\hat{\rho}_{\zeta \eta}$ and $\hat{\beta}_{\eta \gamma_0}$ of $-0.30$, 95% CI $(-0.60, 0.08)$, and $0.003$, 95% CI $(-0.017, 0.024)$, respectively. The estimated correlation between the individual rate of change of transformed TFC values and the expected transformed TFC at time zero is negative and not quite significant at the 5% level. There is no obvious association between this rate of change and the baseline transformed value.

Figure 5 shows the individual regression coefficients (for $t > 0$) plotted against the baseline value for (a) raw TFC scores and (b) transformed TFC. Note that the larger points at the bottom of the plots are associated with the three patients who each had only two postbaseline readings and therefore the least accurate slope estimates. There is no obvious relationship between the individual slopes for $t > 0$ and the baseline values for either the raw or transformed TFC scores. The estimates of $\sigma_{\epsilon}$ and $\sigma_{\delta}$ obtained using (3.12) and the related analysis of variance are 1.430 and 0.298 for the raw TFC and 0.550 and 0.121 for the transformed TFC, respectively. These values are virtually equivalent to the earlier REML estimates for the system (3.1)–(3.2) ($t \geq 0$) given in Table 1.
5. DISCUSSION

The substantive conclusion is that for TMS there is no firm evidence of a relation between initial value and slope. For TFC, a scale with a very limited range, it is desirable both to transform the response scale to reduce the effects of end-point constraint, and also to exclude the two patients whose initial (and all subsequent) scores were zero. For the remaining patients, although there is a suggestion of a negative relationship between the estimated individual rate of change of transformed TFC values and the expected transformed TFC at time zero, the evidence for this conclusion is not decisive. Analysis of a larger set of data is desirable. It seems unlikely that, even if confirmed, the relation is strong enough for predictive purposes unless the slope for an individual patient could be shown to depend appreciably on an explanatory variable, thereby eliminating a substantial source of random variation. A final point is that in any protocol for a new study in which TFC is an important outcome variable it might be advisable either to exclude patients with a very low initial score or at least to give provision for them to be considered separately.

The analysis illustrates a number of methodological issues. Quite apart from the important need to escape entrapment by regression to the mean, there are two broad distinctions of formulation. One is that between regression on the initial underlying population mean versus regression on the initial observed value. The other is the distinction between treating the initial value as part of the underlying random system and treating the initial value as distinct, for example, related to the reasons why a patient enters the study. For these data any association with observed value at entry is so weak as to be useless for predicting slope. There is, however, some evidence of a modest relationship between the underlying initial mean and slope for TFC. In future studies, it may be worthwhile trying to measure the initial value more precisely.

The various estimates reported in Table 1 are readily available by fitting a carefully formulated multi-level model using likelihood-based methods. This approach on its own is likely to be rather dangerous, however, and may overlook important aspects of the data. Our more elementary analyses are based on fitting separate regressions to each patient, examining the results graphically and constructing an analysis of variance table. This table is closely connected to the analysis of covariance table common in the older literature. The estimate of the component of variance between regression coefficients obtained via the analysis of variance table equating mean squares to their expectation is very similar to that obtained by the slightly more efficient REML procedure.
The details of the statistical analysis illustrate a number of general points. First is the need to avoid totally a regression to the mean effect. Then there is the more subtle aspect of distinguishing between regression of slope on the observed value of the initial response and regression on some underlying notional true value. The former permits the possibility that the initial observation is predictive of slope even though it is not part of the main stochastic system. The suggested approach to analysis also highlights other potential problems such as apparently anomalous individuals. These latter issues are likely to arise quite often when multilevel random variability is present.

ACKNOWLEDGMENTS

We would like to thank Elizabeth McCusker, Neil Mahant, and Shanti Graham of the Department of Neurology, Westmead Hospital, for bringing this problem to our attention and for providing the data used in the examples.

REFERENCES

BURTON, P., GURRIN, L. AND SLY, P. (1998). Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modeling. Statistics in Medicine 17, 1261–1291.

GALTON, F. (1869). Hereditary Genius. London: Macmillan.

HUNTINGTON STUDY GROUP (1996). Unified Huntington’s Disease Rating Scale: reliability and consistency. Movement Disorders 11, 136–142.

LAIRD, N. M. AND WARE, J. H. (1982). Random-effects models for longitudinal data. Biometrics 38, 963–974.

LEYLAND, A. H. AND GOLDSTEIN, H. (2001). Multilevel Modelling of Health Statistics. Chichester: John Wiley and Sons.

MAHANT, N., MCCUSKER, E. A., BYTH, K. AND GRAHAM, S. (2003). Huntington’s disease: clinical correlates of disability and progression. Neurology 61, 1085–1092.

OMAR, R. Z., WRIGHT, E. M., TURNER, R. M. AND THOMPSON, S. G. (1999). Analysing repeated measurements data: a practical comparison of methods. Statistics in Medicine 18, 1587–1603.

SENN, S. (1997). Editorial. Statistical Methods in Medical Research 6, 99–102.

VERBEKE, G. AND MOLENBERGHS, G. (2000). Linear Mixed Models for Longitudinal Data. New York: Springer-Verlag.

[Received May 28, 2004; first revision October 21, 2004; second revision December 7, 2004; accepted for publication January 6, 2005]