Quantifying Regression to the Mean Using Proc Mixed

Thaddeus J. Haight, School of Public Health-University of California, Berkeley, CA
Long Ngo, School of Public Health-University of California, Berkeley, CA
Ira Tager, School of Public Health-University of California, Berkeley CA

Abstract-In an analysis of change using longitudinal data, adjusting for initial value can lead to biased results if regression to the mean (RTM) distorts the relationship of the change and initial value. The effects of RTM can be visually evaluated by graphically comparing the change in a variable across time to its initial value. Our aim is to assess RTM quantitatively by measuring the coefficient of reliability. This represents the amount of between person variation over the total variation (between person + within person variation) in longitudinal data. We determine this quantity using PROC MIXED, and show its connection to the results obtained by graphing change against the initial value where RTM occurs. Using our results, we apply a method to study change adjusted for initial value which corrects for RTM (Vollmer, 1986).

INTRODUCTION

The relationship of the change in an outcome measure and the initial value of that measure is of particular interest to investigators in areas such as clinical trials and epidemiology. In clinical trials, for example, an investigator may want to study the effectiveness of a treatment on some outcome, but to do so requires knowing what patients responses were preceeding the treatment. The investigator has to "adjust" for the pretreatment status of the patients to accurately assess the efficacy of the treatment. In an epidemiological study, an investigator may want to study change as a function of a set of covariates, including the initial value, because the initial value might play an important role in the change of the outcome over time. The investigator may be interested in seeing the adjusted effects of other covariables on change after adjusting for the effect of the initial value.

One of the potential pitfalls when analyzing change adjusted for initial value is regression to the mean (RTM) (Vollmer). RTM is a phenomenon that obscures the true underlying relationship between two measuements, because of the effects of measurement error on those measurements (Brand). It occurs in the case when one of the measures is compared to another, for instance when a comparison is made of a follow-up value to the initial value. We are specifically interested in the change in an outcome measure and its relationship with the initial value which occurs as a consequence of adjusting the change for the initial value. This also poses a potential RTM problem because the bias from the comparison of a secondary measurement and initial measurement, implicit in studying change, would affect the relationship of the change and initial value (Brand). Therefore, in a model with change as the outcome, we can assume the expected change in a measure adjusted for the measure's initial value will be biased to the extent that RTM distorts the relationship of the change and initial value.

In the context of clinical trials, the results of a drug's efficacy could be distorted if the RTM isn't controlled. The reason is that in assessing the effectiveness of a drug, the change due to treatment has to account, or adjust, for pretreatment status. If RTM affects the relationship of the change in level due to treatment and the pretreatment level, then the true underlying effectiveness of the drug is questionable. RTM can also potentially lead to biased results in epidemiological studies of change involving longitudinal data. Here the problem would occur in the context of estimating expected change given a set of covariates that includes initial value, where RTM affects the true underlying relationship of the change and initial value. Then the estimated change may be larger or smaller than the true change. This latter problem is explored in the current paper using an example in which change in lung function is investigated as of function of a set of covariates that includes baseline (initial) lung function. It is suspected that baseline lung function plays an important role in the change in lung function over time. It is also suspected that RTM might possibly obscure the relationship of the change in lung function and baseline lung function.

One method of identifying potential effects due to RTM in such a case is to simply plot the observed change against the observed initial value. Change would be negatively correlated with the initial value. That is, as the initial value increases there would be a corresponding decrease in the change (Brand). Although this is an efficient and useful method for determining whether RTM may be present, it does not provide a complete diagnosis of the problem, nor does it provide a solution. There is no way to disentangle how much of the negative relationship is the result of RTM and how much of it is the true underlying relationship of the change and initial value. A quantitative measure of the effect that is directly the result of RTM would clarify the extent of the bias due to RTM in the relationship. It would also make different techniques available in addition to the ones typically used to control RTM.

We discuss a method for quantifying RTM using the information from the between- and within-subject

variability for a set of repeated measures data, and directly calculate the effect due to RTM. This is easily achieved through the PROC MIXED procedure of SAS, which can derive unbiased estimates of these variance components using compound-symmetry (CS) variance structure and restricted maximum likelihood (REML) estimation.

Once RTM is quantified, it can be corrected. Vollmer proposes a solution that evaluates change adjusted for initial value, while correcting for the potential effects of RTM. By applying this method, two useful results are obtained. One is that we can reassess the relationship of the change and initial value after RTM is corrected, to see if any relationship remains after the effect of RTM is removed. The other is that the observed change can be transformed using the information about the relationship of the change and initial value after RTM is corrected. These transformed, or adjusted, change values can then be modelled as a function of a set of covariates, including initial value, without the distorting effects of RTM.

METHODS AND RESULTS

Study Population

The Study of Physical Performance and Age-Related Changes in Sonomans (SPPARCS) is a community-based study of persons 55 and older in Sonoma, California to investigate changes in function, health, and activity in the elderly. Among 2092 participants in the study, 725 healthy individuals (419 females, 306 males) completed at least 8 minutes of exercise on a treadmill (Bruce protocol) on 2 occasions separated by an average of 2 years. During exercise, several pulmonary and cardiovascular readings were recorded at 15 second intervals for each individual.

$OUES_{ij}$ is defined as the least squares (LS) slope estimate of the oxygen consumption $VO2_{ij}$ (ml/min) readings regressed on the log ventilation VE_{ij} (L/min) readings for the *i*th person at the *j*th timepoint. $OUES_{ij}$ is utilized as a measure of each subject's lung function at a given time. In addition, other measurements (peak heart rate, exercise duration) and variables (age, smoking history) were collected for each individual.

Studying Change as a Function of Initial Value

All analyses were sex-specific. Plots comparing observed change in $OUES_i$ ($OUES_{i2}$ - $OUES_{i1}$) and observed initial value, $OUES_{i1}$, were generated for females and males (Figure1, Figure2). Given the linearity of the relationships,slope parameter LS estimates were obtained β_{raw} (Table 2). The plots and slope estimates for both sexes led us to suspect that RTM could potentially be affecting their respective relationships of the change and initial value. Brand and Vollmer outlined a series of steps that can be used in assessing the effects of RTM in a problem of change and initial value. These steps were applied to the current problem of OUES to determine the extent that RTM affected the change and initial value of the measure separately for each sex. For the*i*th person, let observed OUES at the 2 timepoints be represented as

$$\begin{aligned} Y_{i1} &= U_{i1} + e_{i1} \\ Y_{i2} &= U_{i2} + e_{i2} \end{aligned} \qquad (1)$$

where U_{i1} and U_{i2} represent the true underlying OUES values, and e_{i1}, e_{i2} represent the measurement error associated with the observed values.

$$\begin{aligned} U_{ij} &\sim N(\eta_j, \gamma^2_j) \\ e_{ij} &\sim N(0, \delta^2) \end{aligned}$$

If A_i is the average of U_{i1} and U_{i2} then (1) above can be rewritten as

$$\begin{aligned} Y_{i1} &= A_i + E_{i1} \\ Y_{i2} &= A_i + E_{i2} \end{aligned} \qquad (2)$$

where $A_i \sim N(\mu, \sigma^2)$, $E_{ij} \sim N(0, \delta^2)$, and A_i, E_{i1} and E_{i2} are statistically independent.

We assume the joint distribution of Y_1 and Y_2, across all subjects, to be multivariate normal:

$$\begin{array}{lll} Y_1 & \mu & \sigma^2+\delta^2 \quad \sigma^2 \\ Y_2 & \mu & \sigma^2 \quad \sigma^2+\delta^2 \end{array}$$

That is, if we select one individual from this distribution, their Y_{i1} and Y_{i2} would be distributed according to an overall mean OUES μ, a Var $(Y_{ij}) = \sigma^2 + \delta^2$, and Cov $(Y_{ij}, Y_{ij'}) = \sigma^2$.

Given the distribution of the random variables Y_1 and Y_2,

$$E[Y_2 \mid Y_1 = y_1] = \mu + G(Y_1 - \mu) \qquad (3)$$

where $G = (\sigma^2/\sigma^2 + \delta^2)$, $0 \le G \le 1$, represents the coefficient of reliabiltiy, or proportion of between-subject variability to the total variability, as developed by Brand. See appendix for development of (3). Note that if G=1, then $E[Y_2 \mid Y_1 = y_1] = Y_1$. That is putting aside any mean differences between Y_1 and Y_2 that occur as a function of some set of covariates X, if all the variability in Y_1 and Y_2 can be explained by just between-subject variability and not any within-subject variability, then we would expect $Y_1 = Y_2$. This is represented by a graph of Y_2 and Y_1 with a slope of 1. Brand proves that

$$\begin{aligned} E[Y_2 - Y_1 \mid Y_1 = y_1] &= \mu + G(Y_1 - \mu) - Y_1 \\ &= -(Y_1 - \mu) + G(Y_1 - \mu) \\ &= -(1-G)(Y_1 - \mu) \end{aligned} \qquad (4)$$

where $1 - G = (\delta^2/\sigma^2 + \delta^2)$ represents the coefficient of unreliabiltiy, or proportion of within-subject variability to the total variability. If G = 1, the expected change $Y_2 - Y_1$ given the initial value Y_1 is 0, which can be represented graphically by change on initial value with a zero slope.

Again, putting aside any mean changes that occur as a function of X, if only between-subject variability explained the variability in Y_1 and Y_2, we would expect to see zero change given Y_1. But if some of the total variability was explained by within-subject variability, we would expect this to be reflected in the coefficient 1 – G. This coefficient represents the bias affecting the change and initial value as a result of within-subject variability, or RTM, and determines the extent to which the slopes of change in OUES on initial OUES seen in Figures 1 and 2 are the result of RTM.

Quantifying RTM

1 – G, which we will rename λ, can be calculated directly from esitmates of σ^2 and δ^2 which can be derived using the PROC MIXED procedure of SAS.Littell et al. describe two types of models for estimating σ^2 and δ^2 from repeated measures data. One model utilizes a random effects term so that (2) can be rewritten as a single line

$$Y_{ij} = \mu + \alpha_i + E_{ij} \qquad (5)$$

where A_i is now partitioned into the fixed effect μ and α_i, the random effect due to subject *i*, where $\alpha_i \sim N(0, \sigma^2)$. $Var(Y_{ij}) = Var(\alpha_i) + Var(E_{ij}) + 2Cov(\alpha_i, E_{ij}) = \sigma^2 + \delta^2$, assuming α_i and E_{ij} are statistically independent and $Cov(Y_{ij}, Y_{ij'}) = \sigma^2$. The variable ID, which is used to identify each unique person in the sample, is also used here to identify α_i. Now μ, σ^2, and δ^2 can be estimated from (5), using the following PROC MIXED statements:

```
PROC MIXED DATA = dataname;
CLASS ID;
MODEL OUES= /S;
RANDOM ID/;
RUN;
```

Littell et al. propose an alternative model to (5) in the form of

$$Y_{ij} = \mu + E_{ij} \qquad (6)$$

in which the $Cov(Y_{ij}, Y_{ij'}) = \sigma^2$, the covariance pattern occurring between observations for a particular *i*, in the absence of the random effect α_i, is shifted to the error term E_{ij}. This eliminates the need for a random effect term in the model. The E_{ij}'s are no longer statistically independent. Now assume $Var(E_{ij}) = \sigma^2 + \delta^2$ and $Cov(E_{ij}, E_{ij'}) = \sigma^2$. Use "TYPE = CS" to evaluate the compound-symmetry structure of the variance for the errors in estimating σ^2 and δ^2. The same estimates μ, σ^2, and δ^2 from (5) can be obtained from (6) with the following:

```
PROC MIXED DATA = dataname;
CLASS ID;
MODEL OUES= /S;
REPEATED /SUBJECT=ID TYPE=CS;
RUN;
```

Both procedures use REML (restricted maximum likelihood) estimation which provides unbiased estimates of the two variance components (Diggle et al.). Maximum likelihood (ML) estimation is an available alternative in PROC MIXED for estimating the parameters, but it will not yield unbiased estimates. The reason for this is that it does not penalize for the number of parameters being estimated, whereas REML does (Diggle et al.).

The results of quantifying λ revealed a slight effect due to RTM in both the males (0.1778) and females (0.1998). At this point, since a technique was available to correct for RTM, we proceeded to use it. However, it is uncertain what steps should be taken in judging what is a sufficiently large value of RTM to justify correcting for it.

Table 1. Mean OUES, Follow-Up, Exercise and Estimates Obtained from PROC MIXED by Sex

	Female (n=419)	Male (n= 306)
Mean Oues ±SD		
-Initial	1626 ± 316	2427 ± 493
-Follow-Up	1620± 303	2410 ± 491
Mean Follow-Up Time t_2 (years) ±SD	1.99 ± 0.43	1.98 ± 0.42
Mean Exercise (minutes) ±SD		
-Initial	11.20 ± 3.91	14.10 ± 4.39
-Follow-Up	11.13 ± 4.02	14.00 ± 4.39
Estimated μ ±SE	1623 ± 14	2419 ± 27
Est. σ^2	76832	198849
Est. δ^2	19181	43289
Est. G	0.8002	0.8212
Est. 1-G or λ	0.1998	0.1788

Adjusting for Initial Value while Correcting for RTM

Using the estimate of RTM (Est. λ) from Table 1, one can apply techniques such as Vollmer's to correct for the effect of RTM when estimating the expected change adjusted for initial value. Vollmer's primary motivation for developing this method was to compare the change and initial value, before and after correcting for RTM, where the initial value was normalized for a set of covariates known to affect it. Here we do not normalize the initial value for any particular set of covariates, but simply standardize it.

Let $Z_i = (U_{i1} - \eta_1)/\gamma^2_1$ represent the true standardized initial value of $OUES_{i1}$. Assuming U_{ij} changes linearly with time t_{ij}, then

$$U_{i2} = U_{i1} + \beta_i t_{i2} \qquad (7)$$

and

$$\beta_i = \upsilon + \theta Z_i \qquad (8)$$

where β_i, the rate of change, is a linear function of the standardized initial value Z_i for the ith subject. β_i and U_{i1} can be estimated for the ith subject from the slope and intercept LS estimates of Y_{ij} on t_{ij} where t_{i2}represents the follow-up time for the ith subject and t_{i1} is set to 0. υ^* and θ^* can then be estimated from the LS estimates of the est. β_i regressed on est. Z_i in (8). υ^* and θ^* represent the intercept and slope estimates of the expected rate of change given the standardized initial value, but υ^* and θ^* are biased estimates due to the effect of RTM (λ), so consistent, unbiased estimates, υ' and θ', were derived (see Appendix) by correcting for est. λ. With the 2 sets of estimates, (υ^*,θ^*) and (υ',θ'), β^* and β' were obtained using (8). These represent the predicted rates of change before and after correcting for RTM.

Let C_i^* represent $(U_{i2} - U_{i1})^* = \beta_i^* t_{i2}$ and C_i' represent $(U_{i2} - U_{i1})' = \beta_i' t_{i2}$, or the predicted values of change in OUES based on β^* and β', respectivley. Then let

$$d_i = \begin{cases} C_i^* - C_i' & C_i^* > C_i' \\ C_i' - C_i^* & C_i^* < C_i' \\ 0 & C_i^* = C_i' \end{cases} \qquad (9)$$

where d_i represents the absolute difference between these predicted values of change. The C_i^* have a comparably steeper slope than the C_i' (not shown) when compared with the initial value since these represent the predicted change based on the predicted rate of change β^* uncorrected for λ. We use the difference d_i from these predicted values of change to transform, or adjust, the observed change in Figure 1 and Figure 2 using the following set of equations:

$$(Y_{i2} - Y_{i1})' = \begin{cases} (Y_{i2} - Y_{i1}) - d_i & C_i^* > C_i' \\ (Y_{i2} - Y_{i1}) + d_i & C_i^* < C_i' \\ 0 & C_i^* = C_i' \end{cases} \qquad (10)$$

so that $(Y_{i2} - Y_{i1})'$ equals the corrected change in OUES after the change adjusted for initial value is corrected for RTM. This corrected, or adjusted, change was compared to the observed initial value for each sex (Figure 3, Figure 4). Using this new set of data, slope parameter LS estimates of the adjusted change on initial value were obtained β_{adj} (Table 2) and concurred with the relatively flatter smoothed lines in the figures.

These results revealed a substantially diminished relationship between the change in OUES and initial OUES after correcting for RTM. They confirmed our initial suspicion that the relationship between the change in OUES and initial OUES (Figure 1, Figure 2) was partly an effect of RTM, although at that time we were not aware to what extent the relationship was distorted by RTM. Although we cannot claim that the true underlying relationship of the change in OUES and initial OUES is represented by Figure 3, Figure 4, and the est. β_{adj} coefficients, we can report, based on the evidence gathered here, that these latter representations of the relationship are less biased than the former (Figure1, Figure 2, and est. β_{raw}) as a result of correcting for RTM.

Table 2. Results of Evaluating Change in OUES Adjusted for Standardized Initial Value by Sex

	Female (n=419)	Male (n=306)
Mean Est. U_1 ±SD	1626 ± 316	2427 ± 493
Mean Est. Z ±SD	0 ± 1	0 ± 1
Mean Est. β ±SD	1.80 ± 103	-6.38 ± 158
Est. υ^* ± SE	1.80 ± 4.67	-6.38 ± 8.70
Est. θ^* ± SE	-39.645 ± 4.68	-44.47 ± 8.72
Est. υ'	1.80	-6.38
Est. θ'	-8.78	2.53
Mean Est. β' ±SD	1.80 ± 8.78	-6.38 ± 2.53
Mean Obs. Change OUES ($t_2 - t_1$) ±SD	-5.88 ± 196	-16.75 ± 294
Mean Adj. Change OUES ($t_2 - t_1$) ±SD	-5.41 ± 182.04	-18.46 ± 279.41
Est. β_{raw} ±SE	-0.233 ± 0.028	-0.182 ± 0.033
Est. β_{adj} ±SE	-0.044 ± 0.028	0.008 ± 0.033

DISCUSSION

Vollmer provides a thorough and intuitive discussion of the RTM phenomenon. He describes how measurement error applied to a set of measurements can result in unexpected patterns of change. We plotted observed change in OUES on initial OUES to illustrate the characteristic negative correlation between change and initial value in relationships affected by RTM. Although we may not know the true underlying relationship of change in lung function and initial lung function, we would not expect to see a relationship that indicates that lung function declines with increased baseline levels. Thus the true underlying relationship must be obscured to some extent by bias and we speculated that RTM was a partial source of that bias.

Investigators can apply methods that can diminish if not eliminate bias as a result of RTM. One of these methods is to obtain multiple replicates of the outcome measurement at the different timepoints at which the outcome is measured, and use the means of these replicates at the different timepoints as measurements for subsequent analysis. This diminishes the amount of within-subject variability, so that bias due to RTM is controlled. However in cases where obtaining multiple replicates is not possible, techniques such as Vollmer's may be a useful alternative if the effect due to RTM is known.

After applying Vollmer's method and finding θ', which represents the unbiased estimate of the expected rate of change given the standardized initial value, we decided to transform the observed change to reflect the relationship

of the change and initial value corrected for RTM (Figure3, Figure4). We achieved this by working backwards through Vollmer's method until we were able to find C_i^* and C_i, based on the biased θ^* and unbiased θ'. Using a graph that compared the slopes for C_i^*, C_i' and the initial value, we speculated that by treating the C_i' on the initial value as the H_o slope and C_i^* on the initial value as the H_a slope, we could find change values that would reject the H_a slope. This required transforming the observed change using the absolute differences between the C_i^* and C_i' as shown in (9) and (10).

CONCLUSIONS

In an analysis of change adjusted for initial value, it was shown that RTM can distort the true underlying relationship of the change and initial value. A method was applied for quantifying RTM using the between- and within-subject variability from a set of repeated measures data. These variance components were estimated using PROC MIXED, and RTM was directly calculated from these estimates. A method was used to correct for RTM when the change is adjusted for initial value. As an extension of this method, we showed transformed data comparing the change and initial value with RTM removed.

Acknoledgements – The authors wish to express their appreciation to Patti LeBlanc, Jan Hansen, Janet Thiessen, Linda Barr, Gail Traylor, and Dee Cumberland, who recruited the subjects and administered the interviews; to Paula DiCamillo and Donna Turner, who supervised the exercise testing; to Katya Gomez for administrative support; to Elizabeth McDonald for data entry; and to Bao Nguyen for data management. The authors also thank the physicians of Sonoma, California, the Sonoma Valley Hospital, and the consulting services of Dr. Milton Hollenberg without whose cooperation the study would not have been possible.
This work was supported by grant R01 AG09389 from the National Institute of Aging.

All correpsondence should be directed to:

Thaddeus J. Haight
School of Public Health-CFCH
U.C.-Berkeley
140 Warren Hall #7360
Berkeley, CA 94720-7360

REFERENCES

Brand, Richard "Regression to the Mean in Studies of Change", Lecture from Clinical Trials, Department of Biostatistics, University of California at Berkeley, 1996, pp. 1-20. (Handout Available Upon Request)

Littell, Ramon, Milliken, G., Stroup,W., Wollfinger, R. "The SAS System for Mixed Models", SAS Institute, Cary, NC, 1996, pp. 87-97.

Diggle, Peter, Liang, K., Zeger, S. "Analysis of Longitudinal Data", Oxford University Press Inc., New York, New York, 1994, pp. 64-68.

Vollmer, William "Comparing Change in Longitudinal Studies: Adjusting for Initial Value", Clinical Epidemiology, Vol 41, No.7, 1988, pp.651-657.

APPENDIX

In Brand's development of the coefficient of reliability, G, he defines a general formula for multivariate normally distributed variables X and Y

$$\begin{array}{ll} X & \mu_x \quad \sigma^2_x \quad \rho\sigma_x \\ Y & \mu_y \quad \rho\sigma_x\sigma_y \quad \sigma^2_y \end{array}$$

and defines

$$E[Y \mid X=x] = \mu_y + (\rho\sigma_y/\sigma_x)(X-\mu_x) \qquad (A1)$$

To find $E[Y_2 \mid Y_1 = y_1]$, we assume $Var(Y_{ij}) = \sigma^2 + \delta^2$ as before and replace the components of the general formula above with the following:

$$\begin{aligned} \sigma_x &= (\sigma^2+\delta^2)^{1/2} \\ \sigma_y &= (\sigma^2+\delta^2)^{1/2} \\ \rho &= \sigma^2/(\sigma^2+\delta^2) \\ \mu_y &= \mu \\ \mu_x &= \mu \end{aligned}$$

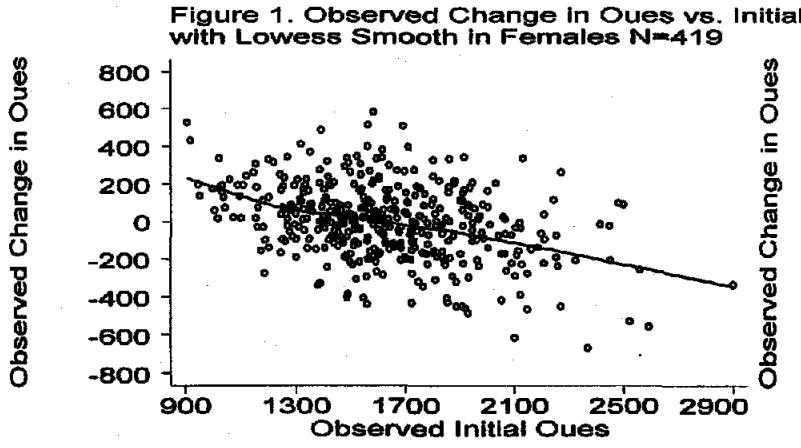
so $$E[Y_2 \mid Y_1 = y_1] = \mu + (\sigma^2/\sigma^2+\delta^2)(Y_1-\mu) \quad (A2)$$

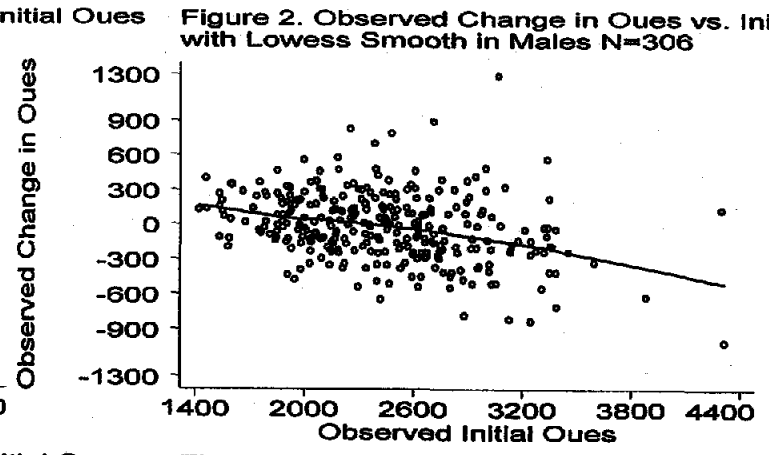
Vollmer develops estimates of υ and θ in the specific case of assuming fixed follow-up intervals for each subject and in the general case, for j=2, that relaxes the fixed interval assumption. We applied the latter case in this paper. It is assumed that measurement error is independent of time.

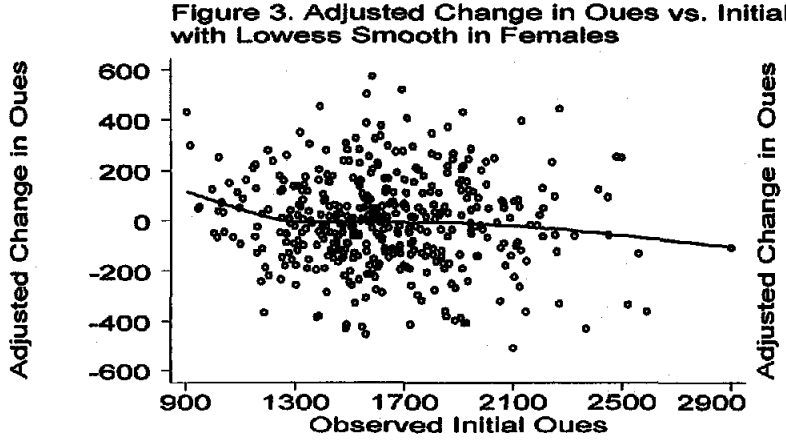
Vollmer solves for a consistent estimator of θ using the following

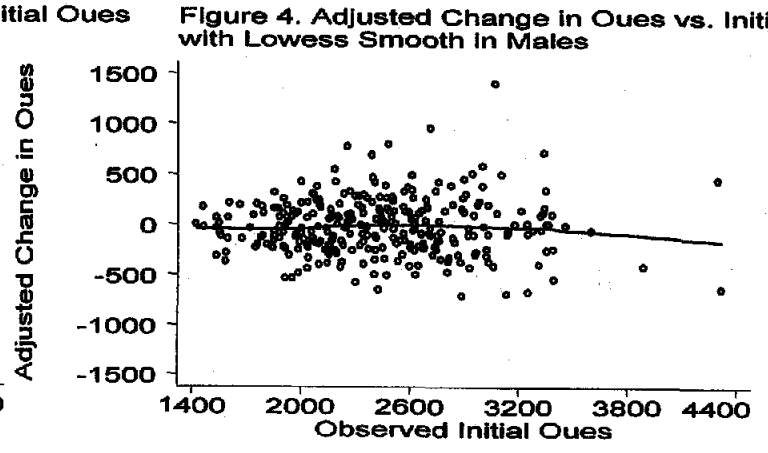
$$\theta' = (\theta^* + \lambda s\tau)/(1-\lambda) \qquad (A3)$$

where $\lambda = (\delta^2/s^2/\sigma^2_z)$ is the effect of RTM, s^2 is the total variability associated with the initial value, and $\tau = E(t_2^{-1})$. Vollmer estimates λ in a slightly different manner than our estimation which uses $1 - G = (\delta^2/\sigma^2+\delta^2) \approx \lambda$. In (A3) we substituted $(\sigma^2+\delta^2)^{1/2}$ for s to keep the pattern consistent with the estimate of RTM of 1-G. It is assumedthat $\upsilon' = \text{mean}(\beta) - \theta'\text{mean}(Z)$, as in the specific case. Vollmer does not include variance estimates of υ' and θ' for the general case.









Comparison of Change in Oues and Initial Oues

Uncorrected <above> and Corrected for RTM