THE EFFECT OF UNMEASURED CONFOUNDERS ON THE ANALYSIS OF LONGITUDINAL DATA

Mari Palta
and
Tzy-Jyun Yao

April 1989

Technical Report # 47

Department of Human Oncology
and
Department of Statistics

UNIVERSITY OF WISCONSIN-MADISON
THE EFFECT OF UNMEASURED CONFOUNDERS ON THE ANALYSIS OF LONGITUDINAL DATA

Mari Palta and Tzy-Jyun Yao

Biostatistics Center, University of Wisconsin
Madison, Wisconsin 53706

Biostatistics Technical Report No. 47
April 1989
SUMMARY

In the regression analysis of longitudinal (or clustered) data it is often implicitly assumed that the regression coefficients of covariates are identical across and within individuals (or clusters). Situations to the contrary arise from confounders not included in the model, biologically different within and between cluster mechanisms and from erroneous specification of the shape of the regression curve. A well known example of confounding is the presence of cohort and period effects in models of aging in epidemiologic research. We formulate a model with confounding and find that different approaches to longitudinal data analysis give different estimates of the regression coefficient when the confounder is excluded from the fitted model. Identifying differences in between and within individual effects can serve as a useful tool for finding confounders and for data interpretation.

Key words:    Longitudinal; Confounding; Robustness; Unmeasured Covariates.
1. Introduction

We consider data that are either longitudinal or clustered. A response of interest, y_i,j, and a covariate, x_i,j, are measured at points j for individual (or cluster) i, j = 1,...,k_i and i = 1,...,n. It is generally a feature of such data that the observations are correlated within individuals, necessitating special methods of analysis. Longitudinal data arise when j corresponds to different time points, whereas clustered data appear naturally when, for example, considering individuals within families or eyes within individual. The issues addressed in this paper apply equally to both situations.

In the analysis of longitudinal data, it is sometimes found that the estimate of the covariate effect differs in between and within individual regression analyses. The situations that give rise to such findings may be viewed as special cases of model misspecification. It is important to investigate how such misspecification affects methods for longitudinal data analysis, as well as to realize the opportunity for detecting lack of model fit that the longitudinal nature of the data affords.

Particular situations that may give rise to differing estimates of covariate effects between and within individuals are 1) biologically different mechanisms for the influence of x on y within and between individuals, 2) omitted confounders (as defined below), and 3) misspecification of the shape of the curve describing the relationship between x and y. We focus here on the first two situations. We adapt the definition of confounding of Kleinbaum and Kupper (1982, page 244), where a confounder is defined as a factor, the control of which changes the relationship between the primary factor under study and the outcome. In a multiple regression situation this
would clearly occur when the potential confounder is correlated with both the factor under investigation \( x \) and the outcome conditional on \( x \).

Case 3) was addressed by Louis et al. (1986) and by Vollmer et al. (1988) in the context of the effect of aging on lung function. Louis et al. investigated the interrelationship of nonlinearity and subject selection in producing different estimates of aging effects in longitudinal and cross-sectional analyses. As an example where situation 1) above may occur, consider the effect of pulse rate on blood pressure. There may exist a long term increase in blood pressure from a increase in an individual's 'true' underlying pulse rate, as well as an additional temporary effect hour to hour. In the same example some aspect of personality may tend to increase both blood pressure and pulse rate, but vary only between individuals. In both situations there would be a different effect of pulse rate on blood pressure between and within individuals.

Other well known examples of omitted confounders are so called cohort and period effects in longitudinal epidemiologic studies. The former arise from the fact that cross-sectional analyses estimate a combination of the effects of true aging and covariates associated with birth cohort. The latter arise in longitudinal data collection as the progression of time leads to changes in measurement technique or general health status of the population. Holford (1985) has addressed the analysis of cohort and period effects in the context of log-linear models.

We model here, a general situation of an omitted confounder. Circumstance 1) above can be considered a special case of between individual confounding in this model. We further investigate the effect of confounders on commonly used analysis techniques for longitudinal data. Comparing
regression coefficients from within and between individual regressions may provide important clues for improving model fit and for further data collection. We make some suggestions for how such comparisons can be made.

2. Model

2.1 Covariate Structure

Let \( \mu_{x_i} \) be the mean of the covariate \( x \) for subject \( i, i = 1, \ldots, n \) for each \( j = 1, \ldots, k_i \) where we limit consideration to \( k_i = k \) for all \( i \). Also assume that \( \mu_{x_i} \mid i.d. \sim N(0, \sigma^2_1) \) and

\[
x_{ij} = \mu_{x_i} + e_{x_{ij}} \quad \text{where} \quad e_{x_{ij}} \mid i.d. \sim N(0, \sigma^2_x)
\] (1)

and \( e_{x_{ij}} \) is assumed independent of \( \mu_{x_i} \). Assume that \( z_{ij} \) is an unmeasured confounder such that

\[
z_{ij} = \mu_{z_i} + e_{z_{ij}} \quad \text{where} \quad e_{z_{ij}} \mid i.d. \sim N(0, \sigma^2_z)
\] (2)

and \( e_{z_{ij}} \) is independent of \( \mu_{z_i} \). Further assume that \( x_{ij} \) and \( z_{ij} \) are related as described by the equations:

\[
e_{x_{ij}} = e_{z_{ij}} + e_{ij} \quad \text{where} \quad e_{ij} \mid i.d. \sim N(0, \sigma^2_e)
\] (3)

\[
\mu_{x_i} = \mu_{z_i} + d_i \quad \text{where} \quad d_i \mid i.d. \sim N(0, \sigma^2_d)
\] (4)

and \( e_{z_{ij}} \) is assumed independent of \( e_{ij} \), \( d_i \) independent of \( e_{ij} \) as well as \( \mu_{z_i} \). These assumptions are general in that \( x \) and \( z \) can be rescaled to make their means 0 and to make the coefficient in (4) equal 1. There is an implied assumption, however, that the relationship between \( x \) and \( z \) is not confounded.
by a third variable, so that the coefficients are the same between and within individuals, i.e., in equations (3) and (4). Denote the square of the correlation between $\mu_{x_i}$ and $\mu_{z_i}$ by $R_1$ so that $\sigma_d^2 = (1-R_1)\sigma_1^2$, and the square of the correlation between $e_{x_{ij}}$ and $e_{z_{ij}}$ by $R_2$ so that $\sigma_e^2 = (1-R_2)\sigma_2^2$.

It follows that the within individual squared correlation between $x_{ij}$ and $z_{ij}$ is $R_2$ also. We refer to $R_1$ as describing the between individual and $R_2$ as describing the within individual relationship between $x$ and $z$.

The situation of between individual confounding (cohort effect) corresponds to $\sigma_z^2 = 0$, thus $R_2 = 0$. Particularly, a differing biological mechanism between and within individuals can be modelled by taking $z_{ij} = \mu_{x_i}$. In general, differences between $R_1$ and $R_2$ can arise either from the nature of $z$ (for example, $z$ may be person specific as for a cohort effect) or by design and sampling procedures that differentially affect the variances $\sigma_e^2$ and $\sigma_d^2$.

### 2.2 Regression Model

We specify a random effects-conditional independence model similar to that of Laird and Ware (1982):

$$
y_{ij} \mid (x_{i1}, z_{i1}, a_i, b_i) \sim \text{iid } N(\alpha + \beta x_{ij} + Bz_{ij} + a_i + b_i x_{ij}, \sigma_0^2)$$

(5)

where $i = 1, \ldots, n; j = 1, \ldots, k$. We make the additional modelling assumptions:

$$
\begin{bmatrix}
a_i \\
b_i
\end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\
0
\end{bmatrix}, \begin{bmatrix} \sigma_3^2 & 0 \\
0 & \sigma_4^2
\end{bmatrix} \right)
$$

and $z_{ij}$ is independent of $a_i$ and $b_i$. In the above model $\alpha$, $\beta$ and $B$ are fixed effects whereas $a_i$ and $b_i$ are random effects. Since (5) defines $z$ as having
an effect on y while (3) and (4) define z to be related to x, we have implied that z is a confounder rather than merely an omitted covariate affecting the precision of equation (5).

3. Properties of Model, Conditional on Observed Covariate

In order to investigate the effect of unknown confounders, the variable z is assumed unmeasured in practice. We investigate the properties of model (5) conditional only on the observed covariate values \( \{x_{ij}\} \). The effect of the covariate \( z_{ij} \) can be divide into two independent components

\[
Bz_{ij} = B_{\mu z} + B_{\sigma z}
\]  

(6)

Since the distribution of \( \mu_{z_i} | \{x_{ij}\} \) depends on \( \{x_{ij}\} \) only through the sufficient statistic \( \tilde{x}_i \), and

\[
\tilde{x}_i ~ \sim N \left( \mu_{\tilde{x}_i}, (1-R_1) \frac{\sigma_1^2 + \sigma_2^2}{k} \right)
\]

we denote \( \frac{\sigma_1^2 + \sigma_2^2}{(1-R_1)\frac{\sigma_1^2 + \sigma_2^2}{k}} \) by \( r \) and obtain

\[
\mu_{\tilde{x}_i} | \{x_{ij}\} ~ \sim N \left( R_1 \tilde{x}_i, R_1 (1-R_1) \frac{\sigma_1^2}{k} + R_1 \frac{\sigma_2^2}{k} \right)
\]  

(7)

For the second term of (6), we note that the distribution of \( e_{z_{ij}} \) depends only on \( \{x_{ij}, \mu_{\tilde{x}_i} \} \) (see appendix 1). We obtain

\[
e_{z_{ij}} | \{x_{ij}\} \sim N \left( R_2 [x_{ij} - \mu_{\tilde{x}_i}], R_2 (1-R_2) \frac{\sigma_2^2}{k} \right)
\]  

(8)

and

\[
e_{z_{ij}} | \{x_{ij}\} ~ \sim N \left( R_2 [x_{ij} - \tilde{x}_i], R_2 (1-R_2) \frac{\sigma_2^2}{k} + R_2 \frac{\sigma_2^2}{k} \right)
\]  

(9)
Inserting these results in model (5) we obtain

\begin{align*}
y_{ij} \mid a_i, b_i, (x_{ij}) & \sim N \left( \alpha + (\beta + B R) x_{ij} + \\
& Br(R_1 - R_2) x_{i} + a_i + b_i x_{ij}, \\
& \sigma_0^2 + B^2 (R_2 (1 - R_2) \sigma_2^2 + r [R_1 (1 - R_1) \sigma_1^2 + \\
& (R_1 + R_2^2) \sigma_2^2 / k]) \right) \right)
\end{align*}

(10)

In addition

\[ \text{Cov}(y_{ij}, y_{ij'} \mid a_i, b_i, (x_{ij})) = B^2 r (R_1 (1 - R_1) \sigma_1^2 + R_1 \sigma_2^2 / k) \text{ for } j \neq j' \]

Conditional on \((x_{ij})\), therefore, the unmeasured confounder not only has the effect of changing the regression equation (the dependence of the mean on \(x\)) but also brings the covariance structure away from a conditional independence model. In contrast, the effect of an unmeasured covariate which is not a confounder would have been only to increase the residual variance in the random effects model. This follows since \(R_1 = R_2 = 0\) implies

\[ z_{ij} \mid (x_{ij}) \sim N(0, \sigma_Z^2) \] where \(\sigma_Z^2\) is the total variance of \(z_{ij}\) so that

\[ y_{ij} \mid a_i, b_i, (x_{ij}) \sim N(\alpha + \beta x_{ij} + a_i + b_i x_{ij}, \sigma_0^2 + B^2 \sigma_Z^2) \]

Proceeding by integrating out the random effects from model (10), we obtain

\begin{align*}
y_{ij} \mid (x_{ij}) & \sim N \left( \alpha + (\beta + B R) x_{ij} + Br(R_1 - R_2) x_{i} + \sigma_0^2 + B^2 (R_2 (1 - R_2) \sigma_2^2 + \\
& r [R_1 (1 - R_1) \sigma_1^2 + (R_1 + R_2^2) \sigma_2^2 / k]) + x_{ij}^2 \sigma_4^2 + \sigma_3^2 \right) \right)
\end{align*}

(11)

with
\[
\text{Cov}(y_{ij}, y_{ij'}) = B^2 x[R_1(1-R_1)\sigma_1^2 + R_1 \sigma_2^2/k] + x^i_j x^i_{j'} \sigma_4^2 + \sigma_3^2 \text{ for } j \neq j' \quad (12)
\]

It also follows that:

\[
\bar{y}_{i.} \mid (x_{i.}) \sim N \left\{ \alpha + [\beta + BR_2 + Br(R_1 - R_2)] \bar{x}_{i.}, \frac{\sigma_0^2}{k} + \frac{B^2}{k} [R_2(1-R_2)\sigma_2^2 + \right.

\left. rR_2^2 \sigma_2^2] + B^2 r[R_1(1-R_1)\sigma_1^2 + R_1 \frac{\sigma_2^2}{k}] + \sigma_3^2 + \bar{x}_{i.}^2 \sigma_4^2 \right\} \quad (13)
\]

\[
\bar{y}_{..} \mid (x_{i.}) \sim N \left\{ \alpha + [\beta + BR_2 + Br(R_1 - R_2)] \bar{x}_{..}, \frac{\sigma_0^2}{n} + \frac{B^2}{k} [R_2(1-R_2)\sigma_2^2 + \right.

\left. + rR_2^2 \frac{\sigma_2^2}{k}] + B^2 r \left[ R_1(1-R_1) \sigma_1^2 + R_1 \frac{\sigma_2^2}{k} \right] + \sigma_3^2 + \right.

\left. \frac{\sigma_4^2}{n} \sum_{i=1}^{n} \bar{x}_{i.}^2 \right\} \quad (14)
\]

4. Results of Applying Common Longitudinal Methods to Model (5)

4.1 Methods That First Fit Individual Slopes and Then Average Over Subjects

A classical technique for the analysis of longitudinal data consists of first fitting slopes \( \hat{\beta}_i \) to all individuals, \( i = 1, \ldots, n \). These are then either combined into an unweighted average or weighted as in the method of Hui and Berger (1983). These approaches are also described and compared in Palta and Cook (1987).

It follows from (11) and (13) that the expected value of the within individual slope

\[
\hat{\beta}_i = \frac{\sum (x_{ij} - \bar{x}_{i.})(y_{ij} - \bar{y}_{i.})}{\sum (x_{ij} - \bar{x}_{i.})^2}
\]

is \( (\beta + BR_2) \). If \( z \) is a between individual confounder so that \( R_2 = 0 \),
the within individual method simply yields $\beta$ as the estimate. This method, therefore, has the ability to adjust for unmeasured between individual confounders.

4.2 Methods that Reduce the Data to Cross-Sectional

Another common approach, at least among analysts not familiar with longitudinal data analysis, is to reduce the information to a cross-sectional data set. This is accomplished by averaging $x_{ij}$ and $y_{ij}$ across $k$ measurements for each individual (where $k$ may be chosen = 1 to correspond to using only one measurement per subject). From (13) and (14), it follows that the expected mean of the slope

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (\bar{x}_{i} - \bar{x})(\bar{y}_{i} - \bar{y})}{\sum (\bar{x}_{i} - \bar{x})^2}$$

is $\beta + BR_2 + Br(R_{1}-R_{2})$.

From (13) we note that the variance increases with $\bar{x}_{i}$. so that unweighted least squares will not be optimal for estimating the slope.

We note that under the assumption $R_{1} = R_{2} = R$, the between individual method gives ($\beta + BR$) as the estimate of the slope. Under this condition, therefore the between and within regressions give the same estimates of the relationship between $x$ and $y$. For any other situation, such as for a between individual confounder, the two estimates will differ (unless $B = 0$).

4.3 Multivariate Methods

Multivariate approaches to longitudinal data analysis depend on the development of estimates for the covariance matrix of the data. Recent work (Liang and Zeger, 1986) shows that when the mean structure is correctly specified, multivariate methods are asymptotically robust to misspecification
of the covariance matrix. Although this work was aimed at generalized linear models, the situation investigated here is included as a special case (the identity link function). As a consequence of the robustness, unweighted regression coefficients are consistent in the case investigated by Liang and Zeger. These coefficients are obtained by ignoring the correlation between observations on the same individual. The unweighted slope estimate is:

\[ \hat{b} = \frac{\sum_{ij} (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_j)}{\sum_{ij} (x_{ij} - \bar{x}_i)^2} = \frac{k\Sigma(\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})}{k\Sigma(\bar{x}_i - \bar{x})^2 + \Sigma(x_{ij} - \bar{x}_i)^2} \]

which may be rewritten

\[ \frac{(n-1)k}{n} \frac{\Sigma(\bar{x}_i - \bar{x})^2}{n-1} \frac{\Sigma(\bar{x}_i - \bar{x})y_i}{\Sigma(\bar{x}_i - \bar{x})^2} + \frac{(k-1)}{n} \sum_{i} \frac{\Sigma(x_{ij} - \bar{x}_i)^2}{k-1} \frac{\Sigma(\bar{x}_i - \bar{x}_i)y_{ij}}{\Sigma(x_{ij} - \bar{x}_i)^2} \]

Taking the limit of individual terms, invoking consistency of variance and regression estimates and the law of large numbers for the second term of the numerator and denominator, we find that as \( n \to \infty \)

\[ \hat{b} \to \beta + B(R_1\sigma_1^2 + R_2\sigma_2^2)/(\sigma_1^2 + \sigma_2^2) \] (15)

By comparison with the method in 4.2, we see that for a large enough number of subjects, the unweighted estimate will approach the one obtained by using a single measurement per subject (a cross-sectional study). We also note that
if \( R_1 = R_2 = R \), (15) becomes \( \beta + BR \), the slope obtained by the methods in 4.1 and 4.2.

Weighting the estimator involves making assumptions regarding the correlation structure. The assumption of exchangeability, i.e.,

\[
\text{corr}(y_{ij}, y_{ij'}) - \rho \text{ for all } i, j \neq j' \text{ leads to the slope }
\]

\[
\hat{b}(\rho) = \frac{(1-\rho) \sum_{i,j} (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_j) + \rho \sum_{i} k \left[ \sum_{j} (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i) \right]}{(1-\rho) \sum_{i,j} (x_{ij} - \bar{x}_i)^2 + \rho \sum_{i} k \left[ \sum_{j} (x_{ij} - \bar{x}_i)^2 \right]}
\]

Using an approach similar to that above, we find that as \( n \to \infty \)

\[
\hat{b}(\rho) \to (\beta + BR_2) + B(R_1 - R_2) \frac{(1-\rho) \sigma_1^2}{(1-\rho) \sigma_1^2 + (\rho k - 2\rho + 1) \sigma_2^2}
\]  

(16)

If \( R_1 = R_2 = R \), again the estimate approaches \( \beta + BR \), as with the other methods.

For large \( k \) (16) approaches the estimate obtained by the individual slope approach in Section 4.1. As demonstrated above (model given by (11) and (12)), exchangeability is not a correct assumption on the correlation structure of model (4). Formulas (15) and (16) demonstrate that when the mean structure is incorrectly specified, the estimates can be asymptotically dependent on the assumed correlation structure. The random effects-conditional independence approach of Laird and Ware (1982) is investigated for some numeric examples below. All multivariate models will give expected and asymptotic value of the slope as \( \beta + BR_2 + Br(R_1 - R_2) \) in the special case \( k = 1 \). With \( k = 2 \) and \( x_{i1} = x_1 \), \( x_{i2} = x_2 \) for all \( i \), the asymptotic result \( \beta + BR_2 \) is obtained. Clearly, then the results generally depend on the design matrix and \( k \).
4.4 Monte Carlo Results for Some Special Cases

Results (15) and (16) were verified by Monte Carlo studies for the
situation $\sigma^2_1 = 1$, $\sigma^2_2 = 9$, $\beta = 1$, $R_2 = 0$ and $z_{ij} = \mu x_{i1}$ for $i = 1, \ldots, n,$
j = 1, \ldots, k with $B = 0,1$ and $k = 2,4$. In this situation $R_1 = 1$.

[Place Figure 1-3 here]

Figures 1-4 show results for some multivariate models as $n \to \infty$ incrementally
with Figure 1 corresponding to unweighted regression as in (15). Figure 2
corresponds to the assumption of exchangeability for the correlation structure
and Figure 3 to an assumption of random effects and conditional independence.
The situations generated with $B=0$ represent no confounding. As expected, all
assumptions yield estimates of the slope that converge to $\beta = 1$ in this
situation. Situations where $B = 1$, are confounded. The first two models
give estimates that converge to the values in formulas (15) and (16), as
indicated by the horizontal lines. No explicit expression was derived for the
asymptote of the Laird and Ware model. Figure 3 shows that the value
depends on $k$ when $B = 1$. This result follows from comparison of the graphs
for $B = 0$ and $B = 1$. Since both were generated with the same seed random
fluctuations relative to the asymptote are parallel in the two graphs.
The graph with $B = 1$ represents different shifts from the line $\beta = 1$ for $k = 2$
and $k = 4$. Parallel to the exchangeability based model, $k = 4$ gives results
closer to $\beta$.

[Place Figure 4 here]

Finally Figure 4 shows results of including the term $x_{i1}$, in the
conditional independence model of Laird and Ware. In this case, since
the mean structure is correct conditionally on the observed covariate vector,
the results are asymptotically consistent. However, variance estimates
obtained from the conditional independence assumption will not be correct.

5. **Comparison of Between and Within Individual Regression Coefficients**

Comparison of between and within individual slopes can be utilized as a tool for detecting the presence of unmeasured confounders. It is useful, therefore, to obtain a variance against which the differences can be judged. We derive the variance of \((\bar{\beta} - \hat{\beta})\) where \(\bar{\beta}\) is the unweighted mean of within individual slopes and \(\hat{\beta}\) is the unweighted estimator of the between individual slope. A more detailed derivation is given in Appendix 2. The comparison of the slopes is to be viewed as an exploratory technique, since we have not yet investigated the behavior of the variance estimator under modelling assumptions other than those in this paper. A further problem is a tendency for outliers to be present among the individual slopes \(\beta_i\) in real life applications, especially with small \(k\). These strongly influence \(\bar{\beta}\) and its variance. It may, therefore, be preferable to compute a trimmed mean or use the median.

Briefly,

\[
\text{Var} (\bar{\beta} - \hat{\beta}) = \text{Var}(\bar{\beta}) + \text{Var}(\hat{\beta}) - 2\text{Cov}(\bar{\beta}, \hat{\beta})
\]

We show in Appendix 2 that \(\text{Cov}(\bar{\beta}, \hat{\beta}) = \sigma^2_u/n\) and can be directly estimated as

\[
\text{Cov}(\bar{\beta}, \hat{\beta}) = \frac{\sum \hat{\beta}_i \bar{\beta}_i (\bar{y}_i - \bar{\bar{y}})}{[n(n-1) \Sigma \bar{x}_i^2]}
\]

(17)

Furthermore, \(\text{Var}(\bar{\beta})\) can be estimated as the usual variance of the individual slopes, divided by \(n\), and the variance of \(\hat{\beta}\) as
\[ \text{Var}(\hat{\beta}) = \frac{\Sigma(y_{i}.-\bar{y}_{.})^2}{(n-1)\Sigma(x_{i}.-\bar{x}_{.})^2} + \left[ \frac{\Sigma(x_{i}.-\bar{x}_{.})^2}{\Sigma(x_{i}.-\bar{x}_{.})^2} \right] \hat{\sigma}_4^2 \]

where \( \hat{\sigma}_4^2 \) is an unbiased estimate of \( \sigma_4^2 \), for example the estimator (17) above multiplied by \( n \). The expression in brackets can be set = 0 for a rough approximation.

6. Discussion

We have demonstrated that several types of model misspecification in longitudinal data analysis lead to differences in the regression coefficients found with within and between individual regression. Additionally, the results of different multivariate models vary under these misspecifications.

The results we have derived above clearly depend on the specific modelling assumptions made. For example, the distribution of the confounder \( z_{ij} \) was assumed normal. Under this assumption, the omission of the covariate maintains the linear nature of the relationship between \( y_{ij} \) and \( x_{ij} \). Other distributional assumptions lead to more complex results. We believe the derivations are general enough, however, to demonstrate the types of perturbations that are possible in longitudinal data with omitted confounders.

It seems important to take advantage of the special opportunity to examine model fit and the presence of confounders that longitudinal data provide. We recommend the following steps: 1) Compute between and within regression coefficients and compare using the standard error derived above as a rough guideline; 2) Plot individual slopes \( \hat{\beta}_i \) against \( \bar{x}_{i.} \). In addition
to detecting outliers, this is a sensitive technique for detecting nonlinearity (provided, of course, that \( \hat{x}_i \) has a reasonable spread). For further testing of nonlinearity, the method of Hui and Berger (1983) can be utilized. If the above steps indicates good model fit, analysis can proceed with the fitting of multivariate models. If, on the other hand, Step 1) results in substantially different regression coefficients in between and within individual regression, and Step 2) does not reveal outliers or non-linearity, a search for non-included confounders should be undertaken.

Acknowledgement

This work was supported by NIH Grant No. R01 CA-18332.
References


Appendix 1: The distribution of $e_{z_{ij}} \mid (x_{ij})$

We first note that if $V \sim N(\mu_v, \sigma_v^2)$ and $U \mid V \sim N(\mu_{u|v}, \sigma_{u|v}^2)$, then

$$V \mid U \sim N(\frac{\sigma_v^2}{\sigma_{u|v}^2} \mu_u \mid v + \frac{\sigma_{u|v}^2}{\sigma_v^2 + \sigma_{u|v}^2} \mu_v, \frac{\sigma_{u|v}^2 \sigma_v^2}{\sigma_{u|v}^2 + \sigma_v^2}) - \frac{\sigma_v^2}{\sigma_{u|v}^2} \mu_u \mid v + \frac{\sigma_{u|v}^2}{\sigma_v^2 + \sigma_{u|v}^2} \mu_v, \frac{\sigma_{u|v}^2 \sigma_v^2}{\sigma_{u|v}^2 + \sigma_v^2}).$$

We also establish below that the distribution of $e_{z_{ij}} \mid (x_{ij})$, $\mu_{x_i}$ depends only on $[x_{ij} - \mu_{x_i}] = e_{x_{ij}}$. The principle above can then be applied to obtain the parameters of the distribution of $e_{z_{ij}} \mid e_{x_{ij}}$, and to integrate out $\mu_{x_i}$.

The distribution of $e_{z_{ij}} \mid (x_{ij})$, $\mu_{x_i}$ is equivalent to $e_{z_{ij}} \mid x_{ij}$, $\mu_{x_i}$, and

can be obtained from Bayes' rule and the following densities:

$$f(x_{ij} \mid e_{z_{ij}}, \mu_{x_i}) = (2\pi \sigma_e^2)^{-1/2} \exp\left[-\frac{(x_{ij} - \mu_{x_i} - e_{z_{ij}})^2}{2\sigma_e^2}\right]$$

$$g(x_{ij} \mid \mu_{x_i}) = (2\pi \sigma_2^2)^{-1/2} \exp\left[-\frac{(x_{ij} - \mu_{x_i})^2}{2\sigma_2^2}\right]$$

$$h(e_{z_{ij}} \mid \mu_{x_i}) = h(e_{z_{ij}})$$

We conclude that the density of $e_{z_{ij}} \mid (x_{ij})$, $\mu_{x_i}$ depends only on $e_{z_{ij}}$.

Appendix 2: Derivation of variance estimator.

We derive estimators for the individual terms in formula (1B) conditional on

$(x_{ij})$ and the random effects $(a_i, b_i)$.

$$\hat{\text{Var}}(\hat{\beta} - \bar{\beta}) = \hat{\text{Var}}(\bar{\beta}) + \hat{\text{Var}}(\hat{\beta}) - 2 \hat{\text{Cov}}(\hat{\beta}, \bar{\beta}) \quad (1B)$$

(a) $\hat{\text{Var}}(\bar{\beta})$:

For each individual $i$, we can write

$$\hat{\beta}_i = \beta_i + e_{\beta_i} = \beta + b_i + e_{\beta_i} \quad (2B)$$

where $\hat{\beta}_i$ is the usual least squares estimator for the within individual regression coefficient $\beta_i$, and $e_{\beta_i}$ are independently distributed errors.
It follows that

$$\hat{\text{Var}}(\hat{\beta}) = [n(n-1)]^{-1} \sum_{i} (\hat{\beta}_i - \bar{\beta})^2$$  (3B)

is an unbiased estimator of the variance of $\bar{\beta}$.

(b) $\hat{\text{Var}}(\hat{\beta})$:

We have

$$\hat{\text{Var}}(\hat{\beta} | (x_{ij})) = [\Sigma(\hat{x}_i, \bar{x}_i)^2]^{-2} \Sigma(\hat{x}_i, \bar{x}_i)^2 \text{Var}(\bar{y}_i) =$$

$$[\Sigma(\hat{x}_i, \bar{x}_i)^2]^{-1} \left[ (k^{-1}[\sigma_0^2 + B^2 R_2(1-R_2) \sigma_2^2 + r R_2^2 \sigma_2^2 / k] + B^2 R_1(1-R_1) \sigma_1^2 + R_1 \sigma_2^2 / k) + [\Sigma(\hat{x}_i, \bar{x}_i)^2]^{-2} \Sigma(\hat{x}_i, \bar{x}_i)^2 \sigma_4^2 \right]$$

In order to estimate this quantity, we note that

$$E \left[ \frac{\Sigma(\bar{y}_i, \bar{y}_i)^2}{(n-1) \Sigma(\hat{x}_i, \bar{x}_i)^2} \right] = \frac{1}{(n-1) \Sigma(\hat{x}_i, \bar{x}_i)^2} \sum_i E(\hat{y}_i, \bar{y}_i)^2 =$$

$$[n \Sigma \Sigma(\hat{x}_i, \bar{x}_i)^2]^{-1} \Sigma \text{Var}(\bar{y}_i) =$$

$$\hat{\text{Var}}(\hat{\beta}) = \left[ \frac{\Sigma(\hat{x}_i, \bar{x}_i)^2 \bar{x}_i^2}{\Sigma(\hat{x}_i, \bar{x}_i)^2} - \frac{\Sigma \hat{x}_i^2 n \Sigma(\hat{x}_i, \bar{x}_i)^2}{\Sigma \Sigma(\hat{x}_i, \bar{x}_i)^2} \right] \sigma_4^2$$

so an unbiased estimator of $\hat{\text{Var}}(\hat{\beta})$ is

$$\hat{\text{Var}}(\hat{\beta}) = \frac{\Sigma(\bar{y}_i, \bar{y}_i)^2}{(n-1) \Sigma(\hat{x}_i, \bar{x}_i)^2} + \left[ \frac{\Sigma(\hat{x}_i, \bar{x}_i)^2 \bar{x}_i^2}{\Sigma(\hat{x}_i, \bar{x}_i)^2} - \frac{\Sigma \hat{x}_i^2 n \Sigma(\hat{x}_i, \bar{x}_i)^2}{\Sigma \Sigma(\hat{x}_i, \bar{x}_i)^2} \right] \sigma_4^2$$  (4B)

where $\sigma_4^2$ is an unbiased estimator of $\sigma_4^2$. 

18
(c) \( \hat{\text{Cov}}(\hat{\beta}, \hat{\beta}) \):

From (2B) we obtain

\[
\text{Cov}(\hat{\beta}_1, \hat{y}_1) = \text{Cov}(\beta_1 + e_{\beta_1}, \alpha_1 + \beta_1 \hat{x}_1) = \hat{x}_1 \cdot \text{Var}(\beta_1) = \hat{x}_1 \cdot \sigma^2_4
\]

It follows that \( \text{Cov}(\hat{\beta}, \hat{\beta}) = \text{Cov}(n^{-1} \Sigma_{i=1}^{n} \hat{\beta}_i, [\Sigma(\hat{x}_i, \hat{x}_i)^2]^{-1} \Sigma(\hat{x}_i, \hat{x}_i) \hat{y}_i) \)

\[
= \frac{\Sigma(\hat{x}_i, \hat{x}_i) \hat{x}_1 \cdot \sigma^2_4}{n} = \frac{\sigma^2_4}{n}
\]

Several options exist for the estimation of \( \sigma^2_4 \). One easily computed estimator is based on the equation

\[
E\left( (\hat{\beta}_1 - \bar{\beta}) (\hat{y}_i - \bar{y}_i) \right) = \text{Cov}(\hat{\beta}_1 - \bar{\beta}, \hat{y}_i - \bar{y}_i) + E(\hat{\beta}_1 - \bar{\beta}) E(\hat{y}_i - \bar{y}_i)
\]

\[
= (1 - \frac{1}{n})^2 \text{Cov}(\hat{\beta}_1, \hat{y}_i) + \frac{1}{n^2} \sum_{i \neq 1} \text{Cov}(\hat{\beta}_1, \hat{y}_i)
\]

\[
= (1 - \frac{2}{n}) \hat{x}_1 \cdot \sigma^2_4 + \frac{1}{n^2} \sum_{i \neq 1} \hat{x}_i \cdot \sigma^2_4
\]

so that \( E[\Sigma(\hat{\beta}_1 - \bar{\beta}) (\hat{y}_i - \bar{y}_i)] = (1 - \frac{2}{n}) \Sigma_{i=1}^{n} \hat{x}_1 \cdot \sigma^2_4 + \frac{1}{n} \Sigma_{i=1}^{n} \hat{x}_i \cdot \sigma^2_4
\]

\[
= \frac{n-1}{n} \Sigma_{i=1}^{n} \hat{x}_i \cdot \sigma^2_4 - (n-1) \Sigma_{i=1}^{n} \text{Cov}(\hat{\beta}, \hat{\beta})
\]

leading to the unbiased estimator

\[
\hat{\text{Cov}}(\hat{\beta}, \hat{\beta}) = \Sigma(\hat{\beta}_1 - \bar{\beta})(\hat{y}_i - \bar{y}_i) / [(n-1) \Sigma_{i=1}^{n} \hat{x}_i]
\]  

(5B)
Figure 1. Simple regression

Fitted model: $E(Y_{ij}) = \alpha + \beta X_{ij}$
Figure 2. Weighted regression

Fitted model: \( \mathbb{E}(Y_{ij}) = \alpha + \beta X_{ij} \)
Figure 3. Laird and Ware’s EM algorithm

Fitted model: \( \mathbb{E}(Y_{ij}) = \alpha + \beta X_{ij} \)
Figure 4. Laird and Ware's EM algorithm

Fitted model: $E(Y_{ij}) = \alpha + \beta X_{ij} + rB\bar{X}_i$. 