Applied Mathematical Sciences, Vol. 4, 2010, no. 5, 225 - 236

Regression, Model Misspecification and Causation, with Pedagogical Demonstration

Gregory L. Light

Department of Management, Providence College
Providence, Rhode Island, 02918 USA
glight@providence.edu

Abstract

This paper shows, by a proposition and a numerical example, how a classic simple or multiple normal regression can achieve with 0.99 probability a near perfect fit to a random sample of any size but due to the omission of an independent variable the signs of the estimated coefficients are all wrong, thus distinguishing prediction from causation.

Mathematics Subject Classification: 62J05, 62H20, 62J10, 62H12, 62H15

Keywords: Regression variable omission, model incomplete bias, perfect regression results, limitations of regression, non-causal predictors

1 Introduction

Model misspecification in regression has long been a well-recognized research problem (for standard textbook expositions on this topic, see, e.g., [4]); the estimation biases resulting from a misspecified model can be very serious (cf., e.g., [5]). Depending on the applications, a misidentification of a variable X as a (or even *the*) *cause* of Y may result in severe consequences. For example, careless correlation reports in health-related matters mislead the public at the minimum, and yet all too often one is provided with such information (which is not to say that there lacks rigorous research methodology; see, e.g., [9]). We are thus motivated to show in this paper how X can be a highly reliable positive predictor of Y due to a population coefficient of correlation close to 1 and yet as a deterministic *cause* $\frac{\partial Y}{\partial X} < 0$.

Section 2 below will highlight the issue on hand by the model

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon,\ \beta_2 < 0,\ \beta_3 > 0, \tag{1}$$

$$X_3 = \gamma_1 + \gamma_2 X_2 + u,\ \gamma_2 > 0, \tag{2}$$

with the random terms ϵ and u satisfying all the standard assumptions, and will also provide a detailed numerical example by a simulation of ϵ and u, resulting in two sample regression equations:

$$\hat{Y}_i = 776.4 - 554.8 X_{i2} + 71.4 X_{i3},\ \text{with } R^2 = 0.99996; \tag{3}$$

$$\hat{Y}_i = 1476.5 + 885.4 X_{i2},\ \text{with } R^2 = 0.97823. \tag{4}$$

In either equation all the coefficients are significant at the two-tailed $p < 0.01$.

Finally Section 3 will conclude with a summary.

2 Analysis

Proposition 1 *Let the population regression equation be*

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon, \tag{5}$$

where:

(1) $X_1 \equiv 1$ and X_2 is nonstochastic,

(2)

$$X_3 = \gamma_1 + \gamma_2 X_2 + u, \tag{6}$$

(3)

$$\epsilon \sim N\left(0, \sigma_\epsilon^2\right),\ E\left(\epsilon_i \epsilon_j\right) = 0,\ \forall i \neq j, \tag{7}$$

$$u \sim N\left(0, \sigma_u^2\right),\ E\left(u_k u_l\right) = 0,\ \forall k \neq l, \tag{8}$$

$$\text{with } \epsilon \text{ and } u \text{ being independent}, \tag{9}$$

and

(4) $\beta_2 < 0$, $\{\beta_3, \gamma_2, \beta_2 + \beta_3\gamma_2\} \subset (0, \infty)$, with σ_ϵ and σ_u sufficiently small relative to the absolute values of β_1, β_2, β_3, and γ_2, then a regression on a

random sample of size n as based on the ordinary least squares estimation of the form

$$\hat{Y}_i = A_1 + A_2 X_{i2},\ i = 1, \cdots, n, \tag{10}$$

is such that

$$\lim_{\sigma_\epsilon,\ \sigma_u \to 0} R^2 = 1, \tag{11}$$

$$\lim_{\sigma_\epsilon,\ \sigma_u \to 0} p_{A_j} = 0,\ j = 1, 2, \text{ with} \tag{12}$$

$$A_2 > 0. \tag{13}$$

Proof. By assumptions (1), (2) and (3), we have

$$\begin{aligned} Y &= \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \\ &= (\beta_1 + \beta_3\gamma_1) + (\beta_2 + \beta_3\gamma_2) X_2 + (\beta_3 u + \epsilon) \\ &\equiv \alpha_1 + \alpha_2 X_2 + \eta \end{aligned} \tag{14}$$

satisfying all the classical normal linear regression hypotheses. Assumption (4) implies that as $\sigma_\epsilon,\ \sigma_u \to 0$, one has $Y_i - \hat{Y}_i \to 0\ \forall i \in \{1, \cdots, n\}$, i.e., approaching a perfect fit through the sample $\{(X_i, Y_i) \mid 1 \le i \le n\}$, so that $R^2 \to 1$ and $p_{A_j} \to 0\ \forall j = 1, 2$; further, since $E(A_2) = \alpha_2 \equiv \beta_2 + \beta_3\gamma_2 > 0$, we have $A_2 > 0$. ■

Remark 1 *It is true that one may estimate $\alpha_2 \equiv \beta_2 + \beta_3\gamma_2$ from the above reduced equation* (14) *for predicting Y by X_2, with the regression satisfying all the standard assumptions thus to defy even the most sophisticated residual analyses (see, e.g.,* [6, 10]*) in detecting the specification error. However, prediction based on correlation is not causation; in fact, from the original full equation* (5) *one can argue that X_2 by itself is a negative factor of Y; consider for example: $X_2 = 1$ represents the male gender, which performs a certain task as measured by Y less well than the female gender $X_2 = 0$, but $X_3 \equiv$ heights is a strong positive factor of Y so that males perform the task better not because of the gender but because of the taller heights. As such, a correct regression model is to come from a theoretical mathematical deduction (for an emphasis on this point and how best to estimate regression parameters under model uncertainty, cf., e.g.,* [2, 8]*); if not, a regression equation in itself is only an extension of correlation, and correlation is not causation - - a common textbook caution, which incidentally, however, may lend itself to the erroneous notion that regression, being more sophisticated, must be about causal-effect; in this regard, even in the research literature one can find the identification of predictor with cause (see, e.g.,* [1]*).*

Remark 2 *We also note that in the above Proposition 1 the fact that X_3 is stochastic does not affect any of the desirable properties of the least squares estimation, since by assumption ϵ and u are independent. Nor is the apparent multicollinearity of X_2 and X_3 a problem, since*

$$Var\left(b_j\right) = \frac{\sigma_\epsilon^2}{\sum_{i=1}^n \left(X_{ij}-\bar{X}_j\right)^2\left(1-r_{23}^2\right)}, \forall j=2,3, \tag{15}$$

$$in\ \hat{Y}_i = b_1 + b_2 X_{i2} + b_3 X_{i3}, \tag{16}$$

so that $\forall r_{23}^2 < 1$ one has

$$\lim_{\sigma_\epsilon^2 \to 0} Var\left(b_j\right) = 0; \tag{17}$$

this can be seen from the following example.

Example 1 *Given $n = 20$, $(X_{1,2},\cdots,X_{10,2},X_{11,2},\cdots,X_{20,2}) = (0,\cdots,0,1,\cdots,1)$,*

$$X_3 = 10 + 20X_2 + u, \quad u \sim N\left(0, \sigma_u^2 = 1\right), \tag{18}$$

$$and\ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon, \quad \epsilon \sim N\left(0, \sigma_\epsilon^2 = 4\right), \tag{19}$$

with ϵ independent of u,

find $\beta_1 \in \mathbb{R}$, $\beta_2 < 0$, and $\beta_3 > 0$ such that with 0.99 probability:

(1) a regression of Y_i against (X_{i2}, X_{i3}) on a random sample of size n will yield $R^2 \geq 0.99$, with the two-tailed $p_{b_j} \leq 0.01$ $\forall j = 1, 2, 3$, and

(2) a simple regression of Y_i against X_{i2} will yield $R^2 \geq 0.95$, $p_{A_j} \leq 0.01$ $\forall j = 1, 2$, and $A_2 > 0$.

Solution 1 *Since*

$$\sigma_\epsilon^{-2} \sum_{i=1}^{20} \left(Y_i - b_1 - b_2 X_{i2} - b_3 X_{i3}\right)^2 \sim \chi_{17}^2, \tag{20}$$

we determine the maximum error sum of squares with 0.99 probability to be

$$SSE_{\max,0.99} \equiv \chi_{0.01,17}^2 \sigma_\epsilon^2 = 33.409 \times 4 = 133.636; \tag{21}$$

then

$$s_{b_2,\max,0.99}^2 = \frac{133.636}{\sum_{i=1}^{20}\left(X_{i2}-\bar{X}_2\right)^2 \cdot \left(1-r_{23,\max,0.99}^2\right)}, \tag{22}$$

where

$$\sum_{i=1}^{20}\left(X_{i2}-\bar{X}_2\right)^2 = 5 \tag{23}$$

and

$$\begin{aligned}
\left(1-r^2_{23,\max,0.99}\right) &= \frac{\left(\sum_{i=1}^{20}\left(X_{i3}-\widehat{10}-\widehat{20}X_{i2}\right)^2\right)_{\min,0.99}}{\left(\sum_{i=1}^{20}\left(X_{i3}-\bar{X}_3\right)^2\right)_{\max,0.99}} &(24)\\
&= \frac{\chi^2_{0.99,18}\sigma_u^2}{20Var\left(X_{i3}\right)_{\max,0.99}} &(25)\\
&= \frac{7.015}{20\times\left[400Var\left(X_{i2}\right)+\widehat{Var}\left(u\right)_{\max,0.99}\right]} &(26)\\
&= \frac{7.015}{2038.67} = 0.003, &(27)
\end{aligned}$$

with

$$\begin{aligned}
Var\left(X_{i2}\right) &= \frac{5}{20} \quad \textit{and} &(28)\\
\widehat{Var}\left(u\right)_{\max,0.99} &= \frac{\chi^2_{0.01,18}}{18} = \frac{34.805}{18}, &(29)
\end{aligned}$$

so that

$$\begin{aligned}
s^2_{b_2,\max,0.99} &= \frac{133.636}{5\times 0.003} = 8909 &(30)\\
\textit{and } s_{b_2,\max,0.99} &= 94.4. &(31)
\end{aligned}$$

Similarly we calculate $s^2_{b_3,\max,0.99}$ *by replacing* $\sum_{i=1}^{20}\left(X_{i2}-\bar{X}_2\right)^2$ *in Equation* (22) *with*

$$\begin{aligned}
&\left(\sum_{i=1}^{20}\left(X_{i3}-\bar{X}_3\right)^2\right)_{\min} &(32)\\
&= 20Var\left(X_{i3}\right)_{\min} &(33)\\
&= 20\times 20^2 Var\left(X_{i2}\right) \ \textit{(by dropping } Var\left(u_i\right)\textit{)} &(34)\\
&= 2000, &(35)
\end{aligned}$$

to arrive at

$$\begin{aligned}
s^2_{b_3,\max,0.99} &= \frac{133.636}{2000\times 0.003} = 22.3 &(36)\\
\textit{and } s_{b_3,\max,0.99} &= 4.7. &(37)
\end{aligned}$$

Now since

$$Cov\left(b_2, b_3\right) = \frac{-\sigma_\epsilon^2 r_{23}}{\sqrt{\sum_{i=1}^{20}\left(X_{i2}-\bar{X}_2\right)^2 \cdot \sum_{i=1}^{20}\left(X_{i3}-\bar{X}_3\right)^2} \cdot \left(1-r_{23}^2\right)} < 0, \quad (38)$$

we have

$$\begin{aligned} Var\left(b_1\right) &= \bar{X}_2^2 Var\left(b_2\right) + \bar{X}_3^2 Var\left(b_3\right) + 2\bar{X}_2\bar{X}_3 Cov\left(b_2, b_3\right) + \frac{\sigma_\epsilon^2}{n} && (39) \\ &< \bar{X}_2^2 Var\left(b_2\right) + \bar{X}_3^2 Var\left(b_3\right) + \frac{\sigma_\epsilon^2}{n}; && (40) \end{aligned}$$

thus, we set

$$\begin{aligned} s_{b_1,\max,0.99}^2 &= 0.25 \cdot s_{b_2,\max,0.99}^2 + \bar{X}_{3,\max,0.99}^2 \cdot s_{b_3,\max,0.99}^2 \\ &\quad + \frac{s_{\max,0.99}^2}{20} && (41) \\ \text{(by Eq. (21))} &= 0.25 \times 8909 + \bar{X}_{3,\max,0.99}^2 \times 22.3 + \frac{133.636/17}{20} \; . && (42) \end{aligned}$$

Since

$$Var\left(X_{i3}\right) = 400 Var\left(X_{i2}\right) + Var\left(u_i\right) = 400 \times 0.25 + 1 = 101, \quad (43)$$

we have

$$Var\left(\bar{X}_3\right) = \frac{1}{20^2} \cdot (20 \times 101) \approx 5 \quad (44)$$

so that

$$\begin{aligned} \bar{X}_{3,\max,0.99} &= \left(10 + 20\bar{X}_2\right) + 3\sqrt{5}, && (45) \\ & \quad \text{three standard deviations above the mean;} && (46) \end{aligned}$$

hence,

$$\bar{X}_{3,\max,0.99}^2 = 26.7^2 \quad (47)$$

and substituting it into Equation (42)*, we have*

$$\begin{aligned} s_{b_1,\max,0.99}^2 &= 18127.5 && (48) \\ \text{and } s_{b_1,\max,0.99} &= 134.6. && (49) \end{aligned}$$

Next, without loss of generality, consider the case of $\beta_1 > 0$; we wish to identify the unique value β_1^* that has a 0.01 probability to yield a $b_1 \in (0, \beta_1)$

with b_1 greater than the null-hypothesis claimed $\beta_1 = 0$ by $\left(t_{17,0.005} \cdot s_{b_1,\max,0.99}\right)$ so as to produce a two-tailed $p \leq 0.01$; i.e.,

$$\begin{aligned} b_1 &\equiv \beta_1 - t_{17,0.01} \cdot s_{b_1,\max,0.99} && (50)\\ \text{and } \frac{b_1}{s_{b_1,\max,0.99}} &= t_{17,0.005}; && (51)\\ \text{i.e., } \beta_1 &= \left(t_{17,0.005} + t_{17,0.01}\right) \cdot s_{b_1,\max,0.99} && (52)\\ &\lessapprox 2 \times t_{17,0.005} \times 134.6 && (53)\\ &\equiv \beta_1^* = 2 \times 2.898 \times 134.6. && (54)\end{aligned}$$

Thus,

$$\beta_1^* = 780.5. \tag{55}$$

Similarly,

$$\beta_2^* \equiv -2 \times 2.898 \cdot s_{b_2,\max,0.99} = -5.8 \times 94.4 = -547.1, \tag{56}$$

and

$$\beta_3^* \equiv \max\left\{2 \times 2.898 \cdot s_{b_3,\max,0.99} = 27.4, \quad \beta_3^{**}\right\}, \tag{57}$$

where β_3^{**} is determined from the requirement of $R^2 \geq 0.99$; to that end, we consider

$$\frac{SSE_{\max,0.99}}{SST_{\min}} \equiv 1 - R^2 = 0.01, \tag{58}$$

where the minimal total sum of squares as defined by $\sigma_u = \sigma_\epsilon = 0$ is

$$\begin{aligned} SST_{\min} &\equiv nVar\left(Y\right)_{\min} \quad (\text{cf. Equation } (19)) && (59)\\ &= n\left[\left(\beta_2^* + 20\beta_3\right)^2 Var\left(X_2\right) + \beta_3^2\sigma_u^2 + \sigma_\epsilon^2\right]_{\sigma_u=\sigma_\epsilon=0} && (60)\\ &\equiv 20\left(\beta_2^* + 20\beta_3^{**}\right)^2 \times 0.25, && (61)\end{aligned}$$

so that (recalling Equation (21)) $100 \cdot SSE_{\max,0.99} = 13363.6 = SST_{\min} = 5\left(\beta_2^* + 20\beta_3^{**}\right)^2$, i.e., $\beta_2^* + 20\beta_3^{**} \approx \sqrt{2672}$, and since by Equation (56) $\beta_2^* = -547.1$, we have

$$\beta_3^{**} \approx \frac{\sqrt{2672} + 547.1}{20} = 29.9 \equiv \beta_3^* \text{ (cf. Equation (57))}. \tag{62}$$

To sum up, we have obtained

$$\begin{aligned} \beta_1^* &\equiv 780.5, && (63)\\ \beta_2^* &\equiv -547.1, \text{ and} && (64)\\ \beta_3^* &\equiv 29.9. && (65)\end{aligned}$$

However, the above $\beta_3^* \equiv 29.9$ is yet to be adjusted upward to provide, with 0.99 probability, that

$$\hat{Y}_i = A_1 + A_2 X_{i2}, \qquad R^2 \geq 0.95, \tag{66}$$
$$p_{A_1} \leq 0.01 \text{ and } p_{A_2} \leq 0.01. \tag{67}$$

Here in analogy with the above multiple regression, we have:

$$SSE_{\max,0.99} \equiv \chi^2_{0.01,18} \sigma^2_{(\beta_3 u+\epsilon)} = 34.805 \times (\beta_3^2 \times 1 + 4), \text{ (cf. Eq. (21))} \tag{68}$$

and (cf. Eq. (60))

$$\begin{aligned} SST_{\min,0.99} &= n\left[(\beta_2^* + 20\beta_3)^2 \, Var\,(X_2) + \chi^2_{0.99,18}\left(\beta_3^2\sigma_u^2 + \sigma_\epsilon^2\right)\right] && (69)\\ &= 20\left[(-547.1 + 20\beta_3)^2 \times 0.25 + 7.015\left(\beta_3^2 + 4\right)\right]. && (70)\end{aligned}$$

We next solve for β_3 in

$$\begin{aligned} 0.05 &= \frac{34.805\,(\beta_3+2)^2}{5\,(-547.1+20\beta_3)^2} && (71)\\ &> \frac{SSE_{\max,0.99}}{SST_{\min,0.99}}, && (72)\end{aligned}$$

and we obtain

$$\check{\beta}_3 = 71, \tag{73}$$

which is sufficient (but not necessary) for $p_{A_j} \leq 0.01 \ \forall j = 1, 2$ with 0.99 probability, as shown below:

For $p_{A_2} \leq 0.01$ we solve for β_3 in

$$\frac{\alpha_2\,(\equiv \beta_2^* + \beta_3\gamma_2)}{s_{A_2,\max,0.99}} = 2t_{18,0.005}, \text{ (recall Eq. (53))} \tag{74}$$

where $\beta_2^* = -547.1$, $\gamma_2 = 20$, $t_{18,0.005} = 2.878$, and

$$\begin{aligned} s_{A_2,\max,0.99} &= \sqrt{\left(\frac{SSE_{\max,0.99}}{18}\right)\left(\sum_{i=1}^{20}\left(X_{i2} - \bar{X}_2\right)^2\right)^{-1}} && (75)\\ &< \sqrt{\left(\frac{34.805\,(\beta_3+2)^2}{18}\right)} \cdot \frac{1}{5} \text{ (as in Eq. (72))} && (76)\\ &= 0.62\,(\beta_3+2), && (77)\end{aligned}$$

so that Equation (74) yields

$$\begin{aligned} 20\beta_3 - 547.1 &= 2 \times 2.878 \times 0.62\,(\beta_3+2) = 3.57\,(\beta_3+2), && (78)\\ \text{and thus, } \beta_3 &= 33.7 < \check{\beta}_3 = 71. && (79)\end{aligned}$$

For p_{A_1} we calculate

$$\frac{\alpha_1\left(\equiv \beta_1^* + \beta_3\gamma_1\right)}{s_{A_1,\max,0.99}} \tag{80}$$

by substituting $\beta_1^* \equiv 780.5$, $\check{\beta}_3 = 71$, $\gamma_1 = 10$, and $s_{A_1,\max,0.99}$

$$= \sqrt{\left(\frac{SSE_{\max,0.99}}{18}\right)\cdot\left(\frac{1}{n} + \frac{\bar{X}_2^2}{\sum_{i=1}^{20}\left(X_{i2} - \bar{X}_2\right)^2}\right)} \tag{81}$$

$$= \sqrt{\left(\frac{34.805\left(71^2 + 4\right)}{18}\right)} \times 0.1 = 31.2 \text{ (by Eq. (68), (73))}, \tag{82}$$

and we find

$$\frac{\alpha_1}{s_{A_1,\max,0.99}} = 47.8, \tag{83}$$

which clearly yields a $p_{A_1} << 0.01$.

We thus have established

$$Y_i = 780.5 - 547.1X_{i2} + 71X_{i3} + \epsilon_i, \;\; \epsilon_i \sim N\left(0,4\right). \tag{84}$$

A simulation of Equation (18) yielded

$(X_{1,3},\cdots,X_{20,3})$ = (9.2, 10.6, 10.9, 9.7, 7.5, 10.0, 10.2, 9.6, 9.5, 10.8, 31.9, 31.3, 29.9, 29.6, 28.9, 29.3, 29.0, 29.7, 29.8, 30.3),

substituting which into Equation (84) with a simulation of ϵ_i then yielded

$(Y_1, \cdots, Y_{20})$ = (1431.8, 1536.1, 1553.5, 1466.5, 1311.9, 1491.7, 1504.2, 1463.4, 1456.0, 1549.4, 2499.7, 2456.3, 2352.0, 2339.4, 2293.7, 2312.3, 2294.8, 2334.0, 2349.7, 2386.6),

and a regression of Y_i against (X_{i2}, X_{i3}) yielded

$$\hat{Y}_i = 776.4 - 554.8X_{i2} + 71.4X_{i3},\; R^2 = 0.99996,\; S.E. = 2.93, \tag{85}$$

$$p_1 = 9.3\times 10^{-26},\; p_2 = 5.1\times 10^{-18}, \text{ and } p_3 = 4.7\times 10^{-25}, \tag{86}$$

but the simple regression of Y_i against X_{i2} resulted in

$$\hat{Y}_i = 1476.5 + 885.4X_{i2},\; R^2 = 0.97823,\; S.E. = 69.62, \tag{87}$$

$$p_1 = 4.7\times 10^{-23}, \text{ and } p_2 = 2.1\times 10^{-16}. \tag{88}$$

Remark 3 *A comparison between the above* $R^2_{simple} = 0.97823$ *and* $R^2_{multi} = 0.99996$ *attests the validity of applying* $R^2 \approx 1$ *as a criterion for correct model specification (cf., e.g.,* [3, 11]*, for other methods of testing models).*

Remark 4 *The above Example 1 highlights the basic fact that with* $\beta_1, \beta_2, \cdots, \beta_K, \beta_{K+1}$ *sufficiently large relative to* σ_ϵ *in*

$$Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \beta_{K+1} X_{K+1} + \epsilon, \quad K \geq 2, \tag{89}$$

one can always achieve a sample regression with all the desirable statistics; under such conditions, if

$$X_{K+1} = \sum_{j=1}^{K} \gamma_j X_j \tag{90}$$

$$\text{with } (\beta_{K+1}\gamma_j + \beta_j)\,\beta_j < < 0 \text{ for some } j, \tag{91}$$

then a sample regression with X_{K+1} *excluded is to produce* b_j *carrying the opposite sign to that with* X_{K+1} *included. Here one is also reminded that the above Equation* (90) *can be nonlinear (cf., e.g.,* [7]*, for estimation of multivariable polynomial regression equations).*

3 Summary Remark

The above analysis has shown that simple regression with low R^2 achieves little purpose and multiple regression with $R^2 \approx 1$ is a criterion for correct model specification, but even a multiple regression with the best inferential statistics is no guarantee for being a correct model. Thus, correct regression models must come theoretical mathematical deduction; for example, in economics the aim of regression is mostly about estimation of the parameters of a theoretically derived equation, rather than an empirical hypothesis testing; likewise, universal physical constants, such as Planck h has been estimated from known functional forms. To conclude, either for intrinsic aesthetic value or for extrinsic utilitarian consideration, prediction is better served by cause-effect than by correlation.

References

[1] E. Boros, P.L. Hammer and J.N. Hooker, Predicting cause-effect relationships from incomplete discrete observations, *SIAM J. Disc. Math.* 7(4) (1994), 531-543.

[2] P.J. Kempthorne, Admissible variable-selection procedures when fitting regression models by least squares for prediction, *Biometrika*, 71 (1984), 593–597.

[3] J.-S. Kim and E. Frees, Omitted variables in multilevel models, *Psychometrika*, 71(4) (2006), 659-690.

[4] J. Kmenta, *Elements of Econometrics*, Macmillan, New York, 1971.

[5] H.J. Larson and T.A. Bancroft, Biases in prediction by regression for certain incompletely specified models, *Biometrika*, 50 (1963), 391–402.

[6] D.Y. Lin, L.J. Wei and Z. Ying, Model-checking techniques based on cumulative residuals, *Biometrics*, 58(1) (2002), 1-12.

[7] J.G. Lin, Modeling test responses by multivariable polynomials of higher degrees, *SIAM J. Sci. Comput.*, 28(3) (2006), 832–867.

[8] J.R. Magnus, The traditional pretest estimator, *Theory Probab. Appl.*, 44(2) (2000), 293-308.

[9] M. Palta and C. Seplaki, Causes, problems and benefits of different between and within effects in the analysis of clustered data, *Health Serv. and Outcomes Res. Meth.*, 3(3-4) (2002), 177-193.

[10] Z. Pan and D.Y. Lin, Goodness-of-fit methods for generalized linear mixed models, *Biometrics*, 61(4) (2005), 1000-1009.

[11] Y. Xia, Model checking in regression via dimension reduction, *Biometrika*, 96(1) (2009), 133-148.

Received: July, 2009