A Procedure for Weight and Model Selection in Assay Development

HUEY L. JU AND SHEIN-CHUNG CHOW*

Department of Statistics, National Chengchi University, R.O.C.
** Biostatistics Department, Bristol-Myers Squibb Company, U.S.A.*

ABSTRACT

The problem of weight and model selection for standard curve in the calibration of an instrument for assay development is considered. An instrument is usually calibrated by studying the relationship between a number of known standards and their corresponding responses. The relationship between known standards and their responses is usually described by a standard curve. Based on the established standard curve, an unknown sample can be determined. In practice, since the responses may have different variabilities at different known standards and the relatioship between the standards and their corresponding responses may not be linear, it is important to select an appropriate model with an appropriate weight in establishing a standard curve for obtaining an accurate and reliable assay result. In this paper, criteria for weight and model selection are proposed to choose an appropriate statistical model among five commonly used models with three different possible weights. An example concerning the validation of a plasma assay of a pharmaceutical compound is presented to illustrate the use of the proposed criteria.

Key words: **Calibration, standard curve, assay validation, weight selection, model selection.**

INTRODUCTION

In the pharmaceutical industry, when a new pharmaceutical compound is discovered, assay and test procedures are necessarily developed for determining the active ingredient of the compound in compliance with the United States Pharmacopeia and National Formulary[1] standards of the identity, strength, quality, and purity of the compound. An assay method is usually developed based on some instruments such as gas chromatography (GC) and high performance liquid chromatography (HPLC). The Current Good Manufacturing Practice (CGMP) codified in 21 Part 211 of Code of Federal Regulations (CFR) indicates that an instrument shall be suitable for its intended purposes and shall be capable of producing valid results. The instrument shall be routinely calibrated, inspected, and checked according to written procedures. The CGMP requires that an assay method be validated to ensure its accuracy and reliability before it can be used for

Correspondence to: Huey L. Ju

Accepted for Publication: Dec. 21, 1995

product testing. The assay method usually involves the calibration of an instrument which relies on the selected standard curve (or calibration curve) [2,3].

In this paper, we focus on some statistical issues that commonly encountered in the development of an assay method for a pharmaceutical compound. We propose a procedure to select the most appropriate model and weight from the five commonly used models for assay development.

CALIBRATION AND STANDARD CURVE

For the development of an assay method, the CGMP (see, e.g., 21 CFR 211.194(a)) requires that an assay method for assessing compliance of pharmaceutical products with established specifications must meet proper standards of accuracy and reliability. An assay method which does not meet established specifications shall not be used. To meet the established specifications, the calibration of an instrument is essential.

A common approach for the calibration of an instrument is to have a number of known standard concentration preparations put through the instrument to obtain the corresponding responses. On the basis of these standards and their corr-esponding responses, a calibration curve can be obtained by fitting an appropriate statistical model between these standards and their corresponding responses. The calibration curve is ususlly referred to as the *standard curve*. For a given unknown sample, the concentration can be determined based on the standard curve by replacing the dependent variable with its response.

For the calibration of an instrument, a linear regression model is often employed to determine the standard curve. The standard curve is then used to determine the unknown sample. The standard linear calibration involves two commonly used methods, namely the *classical* method and the *inverse* method[4,5]. The method described above is usually referred to as the classical method. For the inverse method, a similar standard curve can be obtained by interchanging the dependent variable (response) and the independent variable (standard) in the classical method. The concentration of the given sample can be determined similarly.

STATISTICAL MODELS FOR STANDARD CURVE

The accuracy and precision of an assay depends on the estimate of the unknown sample[6]. The determination of the unknown sample is based on the standard curve which relies on the selection of an appropriate statistical model. Therefore, it is important to select an appropriate statistical model for obtaining reliable assay results. In practice, the following statistical models are commonly used for establishing standard curve in calibration for assay development. These models are currently acceptable to the FDA.

Let X*i* and Y*i* be the known standard concentration preparations and the corresponding responses, where $i = 1,_{...},$n. The five commonly used models are given below:

$$\text{Model 1: } Yi = \alpha + \beta X_i + \varepsilon_i \ ;$$
$$\text{Model 2: } Yi = \beta X_i + \varepsilon_i \ ;$$
$$\text{Model 3: } Yi = \alpha + \beta_1 X_i + \beta_2 X_i^{\ 2} + \varepsilon_i \ ;$$
$$\text{Model 4: } Yi = \alpha X_i^{\ \beta} \cdot \varepsilon_i \ ;$$
$$\text{Model 5: } Yi = \alpha e^{\beta X_i} \cdot \varepsilon_i \ ;$$

Let x_0 and y_0 be the unknown sample and its corresponding response, respectively. The primary objective is to estimate (or determine) the unknown sample based on the select model.

Model 1 is the most commonly used statistical model for standard curve. Based on model 1, the unknown sample at y_0 can be determined as follows:

$$x_0 = \frac{y_o - a}{b},$$

where a and b are the ordinary least squares estimators of α and β, respectively.

If we assume that the standard curve passes through the origin, i.e., there is a zero intercept, then model 1 reduces to model 2. In this case, the unknown sample can be obtained as

$$x_0 = \frac{y_o}{b},$$

where b is the LS estimator of β under model 2.

In some situations, the relationship between X_i and Y_i may be quadratic such as model 3. In this case, the unknown sample x_0 can be determined by solving the quadratic equation

$$b_2 x^2 + b_1 x + (a - y_0) = 0$$

This leads to

$$x_0 = [2b_2]^{-1}\left[-b_1 \pm \sqrt{b_1^2 - 4b_2(a - y_0)}\right]$$

Models 4 and 5 are sometimes of particular interest when the relationship between X_i and Y_i is believed to be nonlinear. In fact, model 4 is equivalent to a linear regression model after a logarithmic transformation, i.e.,

$$\log(Y_i) = \log(\alpha) + \beta \log(X_i) + \varepsilon_i',$$
$$or$$
$$Y_i' = \alpha' + \beta X_i' + \varepsilon_i',$$

where $\varepsilon_i' = \log(\varepsilon_i)$. Thus, the unknown sample can be obtained as

$$x_0 = exp\left[\frac{log(y_0) - log(a)}{b}\right],$$

where a and b are estimates of α and β obtained under the above model. Similarly, model 5 can be linearized by taking a logarithmic transformation, i.e.,

$$\log(Y_i) = \log(\alpha) + \beta X_i + \varepsilon_i',$$

where $\varepsilon_i' = \log(\varepsilon_i)$. Therefore, the unknown sample can be determined as

$$x_0 = \frac{log(y_0) - log(a)}{b},$$

where a and b are estimates of α and β obtained under the above model. It can be seen that under each model, the standard curve can be obtained by fitting an ordinary linear regression.

CRITERION FOR MODEL SELECTION

The question of particular interest to researchers is "How to select an optimum statistical model for determining the standard curve based on the observed calibration data?" To address this question, we proposed the following ad hoc criterion for selecting the most appropriate statistical model among the above five models.

Since models 1-4 are polynomials and model 5 can be approximated by a polynomial, we first fit a model with X at higher orders (say, X^3, X^4, ...). The recommended selection procedure is described below:

Step 1: Starts with the following linear model

$$Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \beta_4 X_i^4 + \varepsilon_i.$$

Let p_{34} be the p-value for testing H_{034} : $\beta_3 = \beta_4 = 0$. If p_{34} is greater than a predetermined level of significance, then go to Step 2, otherwise go to Step 4.

Step 2: Since β_3 and β_4 are not significantly different from zero, the above model reduces to model 3. That is,

$$Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i.$$

We then consider a model among models 1-3. Let p_2 be the p-value for testing H_{02}: $\beta_2 = 0$. If p_2 is less than a predetermined level of significance, then model 3 is chosen; otherwise, go to the next step.

Step 3: If β_2 is not significantly different from 0, model 3 reduces to:

$$Y_i = \alpha + \beta X_i + \varepsilon_i.$$

In this case, we choose between model 1 and model 2 by testing H_0: $\alpha = 0$. If the p-value for testing H_0 is smaller than the predetermined level of significance, then model 1 is chosen; otherwise, model 2 is selected. Note that we assume that $\beta \neq 0$.

Step 4: We select model 4 or model 5. Since model 4 and model 5 have the same number of parameters, we would select the model with a smaller residual sum of squares. The residual sum of squares may be obtained from the PROC NLIN procedure of SAS[7], or any other softwares for nonlinear regression.

A flow chart for the procedure is given in Figure 1.

CRITERION FOR WEIGHT SELECTION

For the calibration of an instrument, it is often observed that the response of a higher standard concentration preparation usually has a larger variability. Therefore, the ordinary least squares approach may not be appropriate. In this case, a weighted least squares method is then considered to remove the heterogeneity of the variability. The weight is selected so that the variance of the response at each standard concentration preparation is stabilized, i.e. the variance of the responses at each standard concentration preparation remains a constant. The selection of an appropriate weight depends on the pattern of the standard deviation of the responses at each standard concentration preparations. For example, if the standard deviation of the response at each standard concentration preparation X is proportion to X, an appropriate choice for the weight is $1/X^2$. Table 1 lists suggested weights for three situations. These weights are commonly adopted in practice.

Table 1. Suggested weights for three situations

SD (Y) is proportional to	Weight
Constant	1*
$\sqrt{X}$	1/X
X	$1/X^2$

*No weights.

For the selection of weights, if there are replicates at each standard concentration preparation X, we suggest to fit the following linear regression model to study the relationship between the standard deviation of Y and the standard concentration preparation X for variance stabilization.

$$SD(Yi) = \beta_0 + \beta_1 \sqrt{X_i} + \beta_2 X_i + \varepsilon_i,$$

where $SD(Y_i)$ is the standard deviation of the response Y_i at the standard concentration preparation X_i.

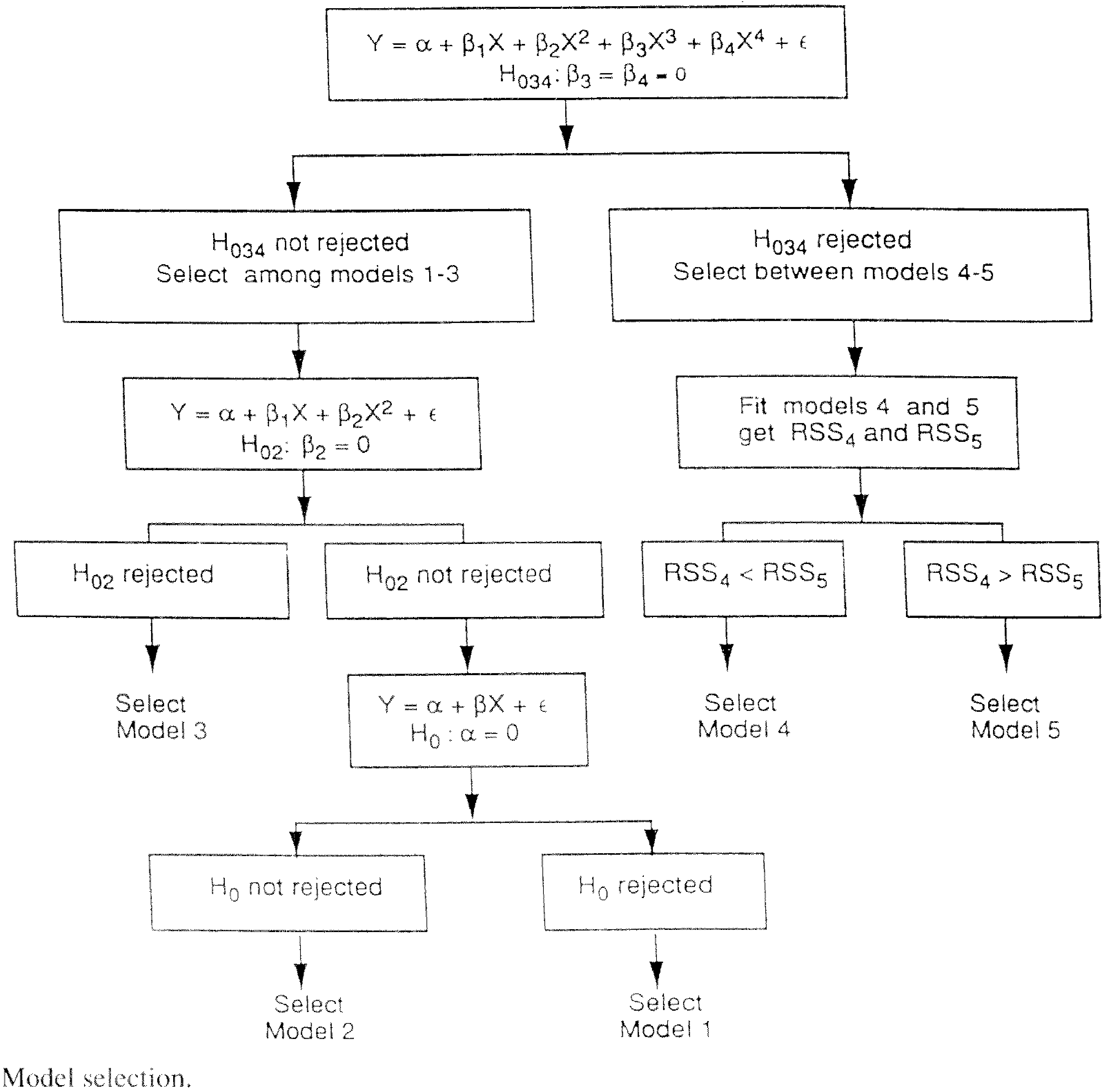


Figure 1. Model selection.

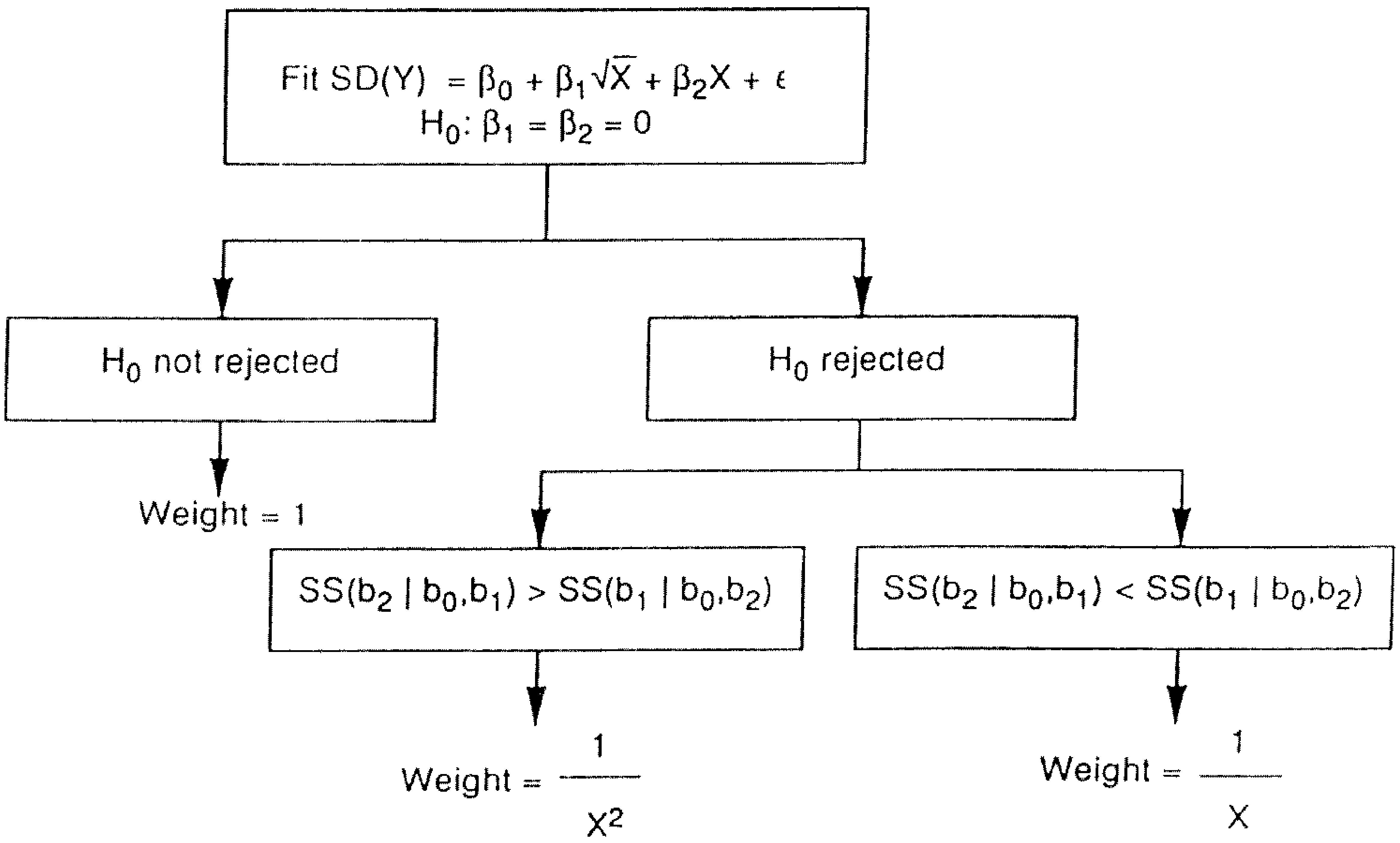

Fit SD(Y) = β0 + β1√X + β2X + ε
H0: β1 = β2 = 0
H0 not rejected
H0 rejected
Weight = 1
SS(b2 | b0,b1) > SS(b1 | b0,b2)
SS(b2 | b0,b1) < SS(b1 | b0,b2)
Weight = 1/X2
Weight = 1/X

Figure 2. Weight selection.

Let p_{12} be the p-value for testing the null hypothesis H_0: $\beta_1=\beta_2=0$. Also, let △ be the level of significance for weight selection. The criterion for weight selection can be summarized as follows:

(1) If $p_{12}>$ △, then no weighting is necessary;

(2) If $p_{12}<$ △ and $SS(b_2 \mid b_0, b_1)>SS(b_1 \mid b_0, b_2)$, then weights$=1/X^2$;

(3) If $p_{12}<$ △ and $SS(b_2 \mid b_0, b_1)<SS(b_1 \mid b_0, b_2)$, then weights$=1/X$,

where $SS(b_2 \mid b_0,b_1)$ represents the contribution of the sum of squares due to the inclusion of $\beta_2 X_i$ when β_0 and $\beta_1\sqrt{X_i}$ are already in the model.

Similarly, the extra sum of squares due to $\beta_1\sqrt{X_i}$ can be expressed by $SS(b_1 \mid b_0,b_2)$. Figure 2 summarizes the procedure for weight selection.

AN EXAMPLE

Consider the calibration of an instrument for plasma concentration of a pharmaceutical compound. The calibration was done on three separate days. Nine standard concentration preparations (x=0.0, 0.5, 1.0, 2.0, 5.0, 10.0, 15.0, 20.0, and 30.0) were chosen. The response of interest is peak response. For each level of standard concentration preparation, three responses (replicates) were obtained on each day. Table 2 lists the responses of these standard concentration preparations. The standard deviations of the responses at each level of standard concentration preparation are given in Table 3. It can be seen from Table 3 that the standard deviation of the response at higher level of standard concentration preparation tends to be larger. Therefore, a weighted least squares method is necessary to stabilize the variance.

For weitght selection, we will use the data from the three days. The p-value for testing the null hypothesis

$$H_0: \beta_1=\beta_2=0$$

is less than 0.0001. This implies that the standard deviation of the peak response is highly correlated with the concentration. Therefore, a weight which is function of X is needed to stabilize the variance of the peak response. Since

$$SS(b_1 \mid b_0,b_2)=0.00009<SS(b_2 \mid b_0,b_1)=0.0088,$$

the $1/X^2$ is therefore chosen.

Following the procedure for model selection, we first fit the model:

$$Y_i=\alpha+\beta_1 X_i+\beta_2 X_i^2+\beta_3 X_i^3+\beta_4 X_i^4+\varepsilon_i.$$

Table 4 summarizes the results from the model selection procedure. Three hypotheses are performed for days 1-3. The first hypothesis is to

Table 2. Calibration data for weight and model selection

Day	Standard	Peak Response		
	Concentration	Replicate 1	Replicate 2	Replicate 3
1	0.0	0.086	0.110	0.110
	0.5	0.160	0.170	0.160
	1.0	0.240	0.220	0.200
	2.0	0.340	0.350	0.340
	5.0	0.650	0.630	0.770
	10.0	1.400	1.360	1.290
	15.0	2.030	1.920	2.030
	20.0	3.020	2.830	2.360
	30.0	3.730	3.770	3.960
2	0.0	0.000	0.039	0.000
	0.5	0.074	0.088	0.069
	1.0	0.130	0.110	0.120
	2.0	0.210	0.210	0.250
	5.0	0.530	0.500	0.470
	10.0	1.100	1.060	1.000
	15.0	1.690	1.480	1.310
	20.0	2.290	2.170	2.160
	30.0	3.250	3.410	3.030
3	0.0	0.032	0.030	0.000
	0.5	0.073	0.081	0.083
	1.0	0.110	0.100	0.100
	2.0	0.190	0.210	0.170
	5.0	0.430	0.430	0.400
	10.0	0.920	0.920	0.870
	15.0	1.380	1.280	1.280
	20.0	1.840	1.800	1.950
	30.0	2.820	2.690	2.380

Table 3. Standard deviations at each standard concentration preparation

Standard	SD (Y)		
Concentration	Day 1	Day 2	Day 3
0.0	0.01	0.02	0.02
0.5	0.01	0.01	0.01
1.0	0.02	0.01	0.01
2.0	0.01	0.02	0.02
5.0	0.08	0.03	0.02
10.0	0.06	0.05	0.03
15.0	0.06	0.19	0.06
20.0	0.34	0.07	0.08
30.0	0.12	0.19	0.23

test if higher order terms (X^3 and X^4) are needed in the model. The second hypothesis is to test if the second order term of X is needed. And the third hypothesis is to test for the term of intercept. The test are conducted sequentially. For day 1 and day 2, model 1 is selected since both hypotheses H_{034} and H_{02} are not rejected, which implies no evidence that including higher order terms will improve the model. Although the hypothesis H_{034} : $\beta_3=\beta_4=0$ is rejected at the 5% level for day 3 which suggests that either model 4 or 5 may be appropriate, it should be noted that the adjusted R^2 for the model with higher order terms included (approximation to models 4 or 5) and the first order term (model 1) included are 0.9888 and 0.9816, respectively. The small difference of the adjusted R^2 of the two models implies that the variation of the peak response could be explained reasonably well by model 1. The model with the higher order terms of the concentration included does not improve much more in explaining the variation of the peak response than model 1. Therefore, model 1 is considered adequate for day 3. The model selected for days 1-3 is confirmed by the results of the 3-day combined data, where model 1 is shown to be the most appropriate model.

After a tentative model (model 1) is identified, the residual plots of the peak response versus the concentration are shown from Figures 3-10. Figures 3-6 give the plots with no weight incorporated in the model. It can be seen that the variance of the residual of the peak response increases when the concentration gets larger. Therefore, a weighted least squares should be applied. A weight of $1/x$ is first employed. However, the heterogeneity of the variability still exists. This problem is improved after the weight $1/x^2$ is included in the model (see Figures 7-10). Note that the standard deviation of Y does not clearly exhibit the patterns for potential weights of $1/x$ and $1/x^2$. Other weights may be more appropriate. However, the purpose of this example is to demonstrate the proposed selection criterion for weight selection among the three possible weights.

Table 4. P-Values for model selection

	$Y=\beta_0+\beta_1X+\beta_2X^2+\beta_3X^3+\beta_4X^4+\varepsilon$	$Y=\beta_0+\beta_1X+\beta_2X^2+\varepsilon$	$Y=\beta_0+\beta_1X+\varepsilon$
Day	$H_{034}:\beta_3=\beta_4=0$	$H_{02}:\beta_2=0$	$H_0:\beta_0=0$
1	0.39	0.29	<0.01
2	0.84	0.08	<0.01
3	0.03	0.02	<0.01
3-day	0.86	0.46	<0.01

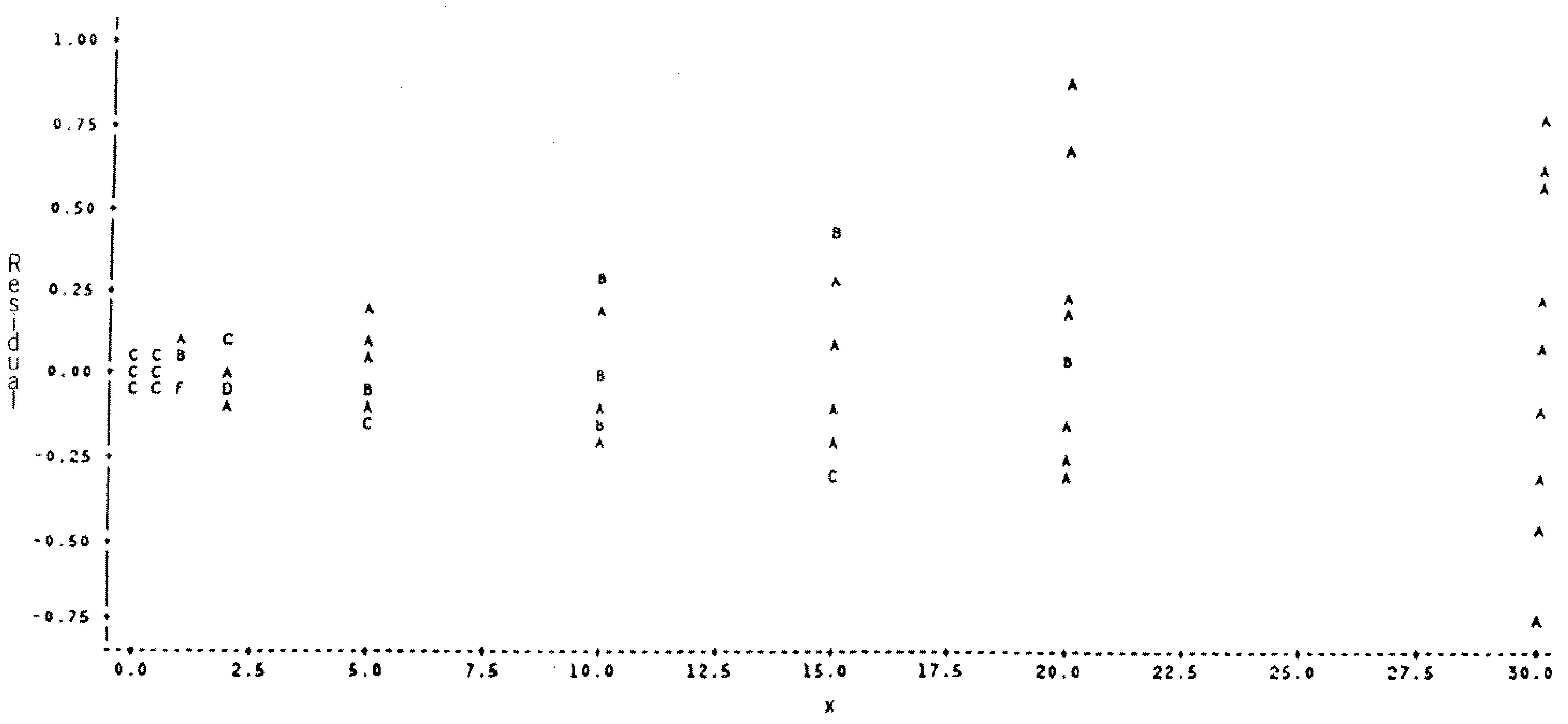


Figure 3. Plot of residual peak response versus concentration from model 1, no weight transformation, 3-day combined.(Legend: A=1 obs, B=2 obs, etc.)

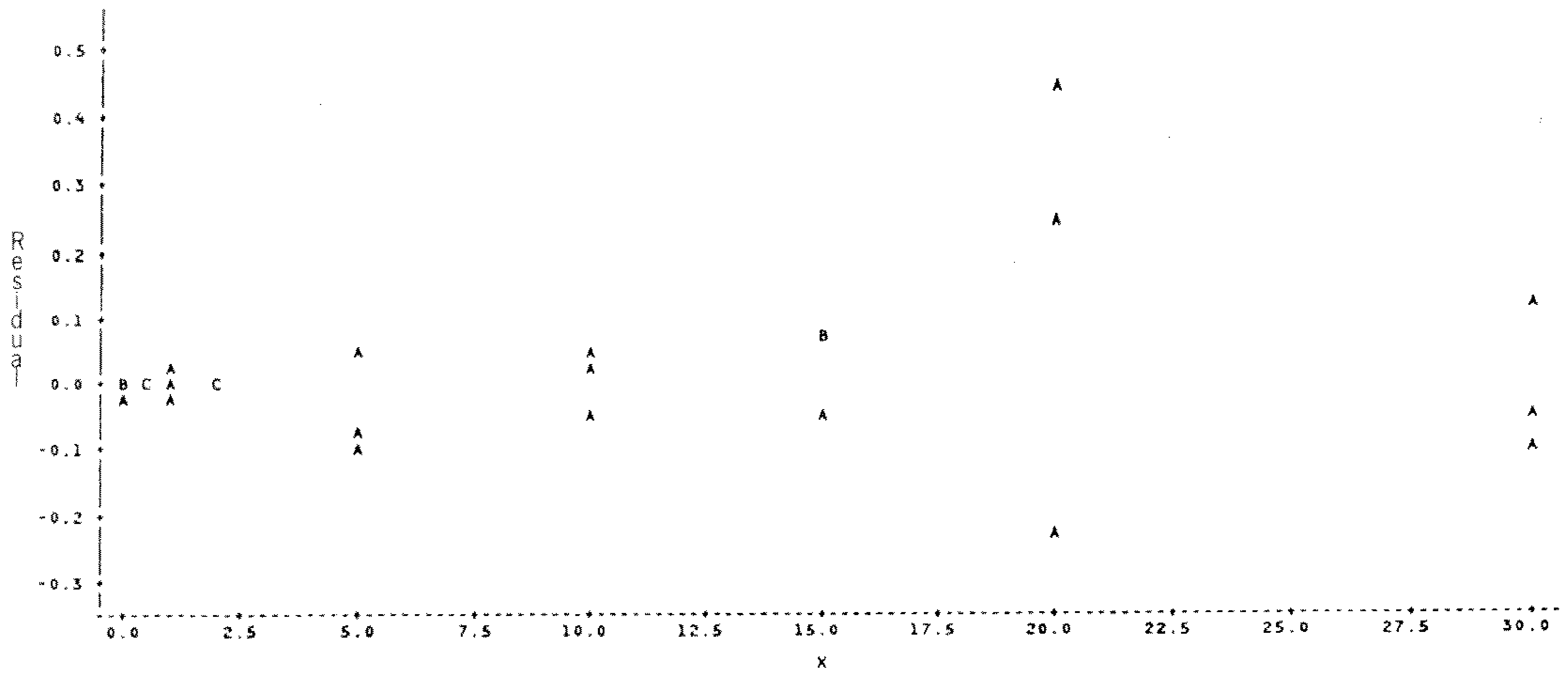


Figure 4. Plot of residual peak response versus concentration from model 1, no weight transformation, day 1. (Legend: A=1 obs, B=2 obs, etc.)

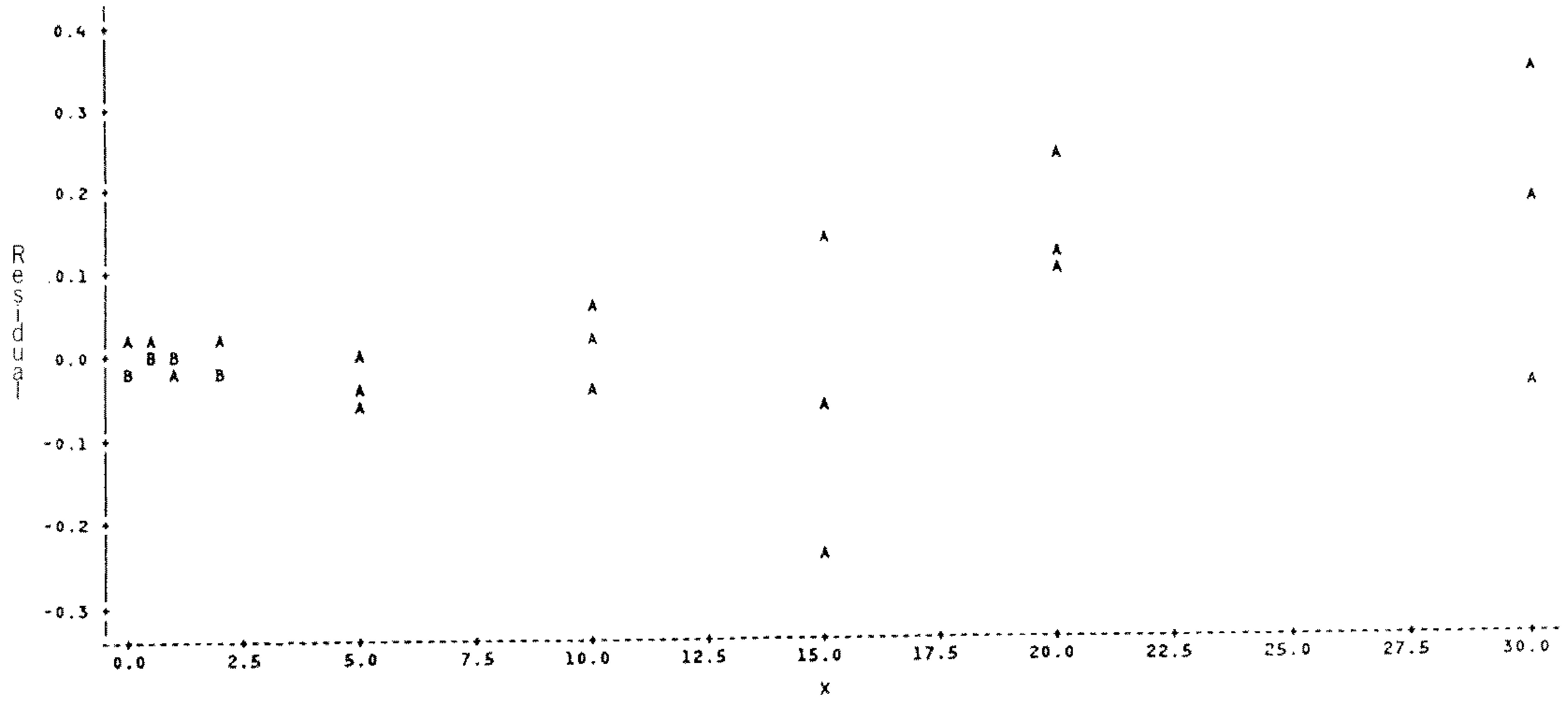


Figure 5. Plot of residual peak response versus concentration from model 1, no weight transformation, day 2. (Legend: A=1 obs, B=2 obs, etc.)

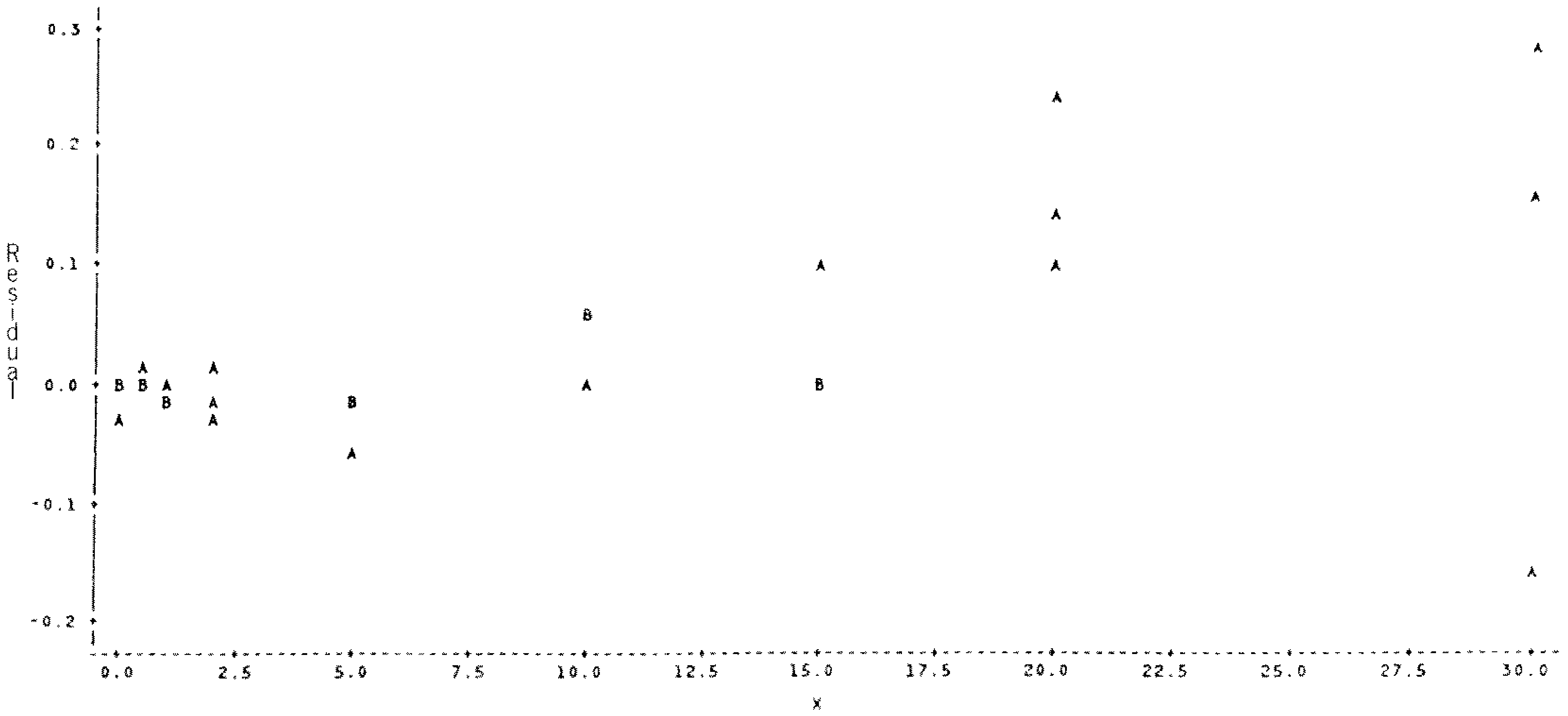


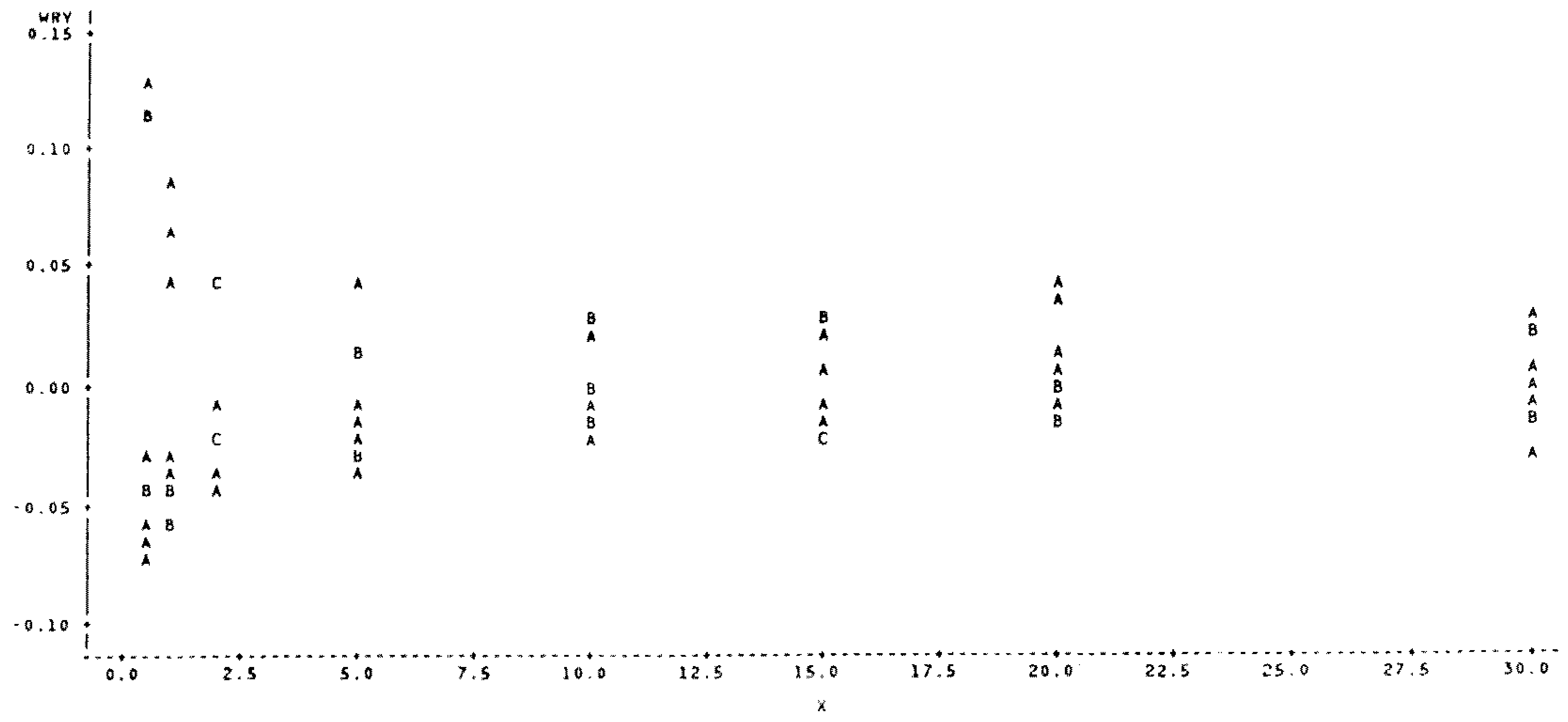
Figure 6. Plot of residual peak response versus concentration from model 1, no weight transformation, day 3. (Legend: A=1 obs, B=2 obs, etc.)

This example illustrates the procedures proposed in the paper. For the model selection, the hypothesis testing results show that model 1 is the best choice of model. However, the p-value for testing whether wights are needed to remove the non-constant variance problem shows significant results. This implies that a weight of a function of X is needed to be incorporated in the model. The residual plots from model 1 confirms that a weighted least squares should be applied to resolve the problem of non-constant variance. And the problem is improved after the weight $1/x^2$ is employed.

Note that in our example, for illustration purpose, we simply ignore the lack-of-fit test in the model selection. In practice, however, lack-of-fit test is an important issue in the determination of standard curve when there are replicates. Hence, it should be taken into account in the model selection when appropriate.

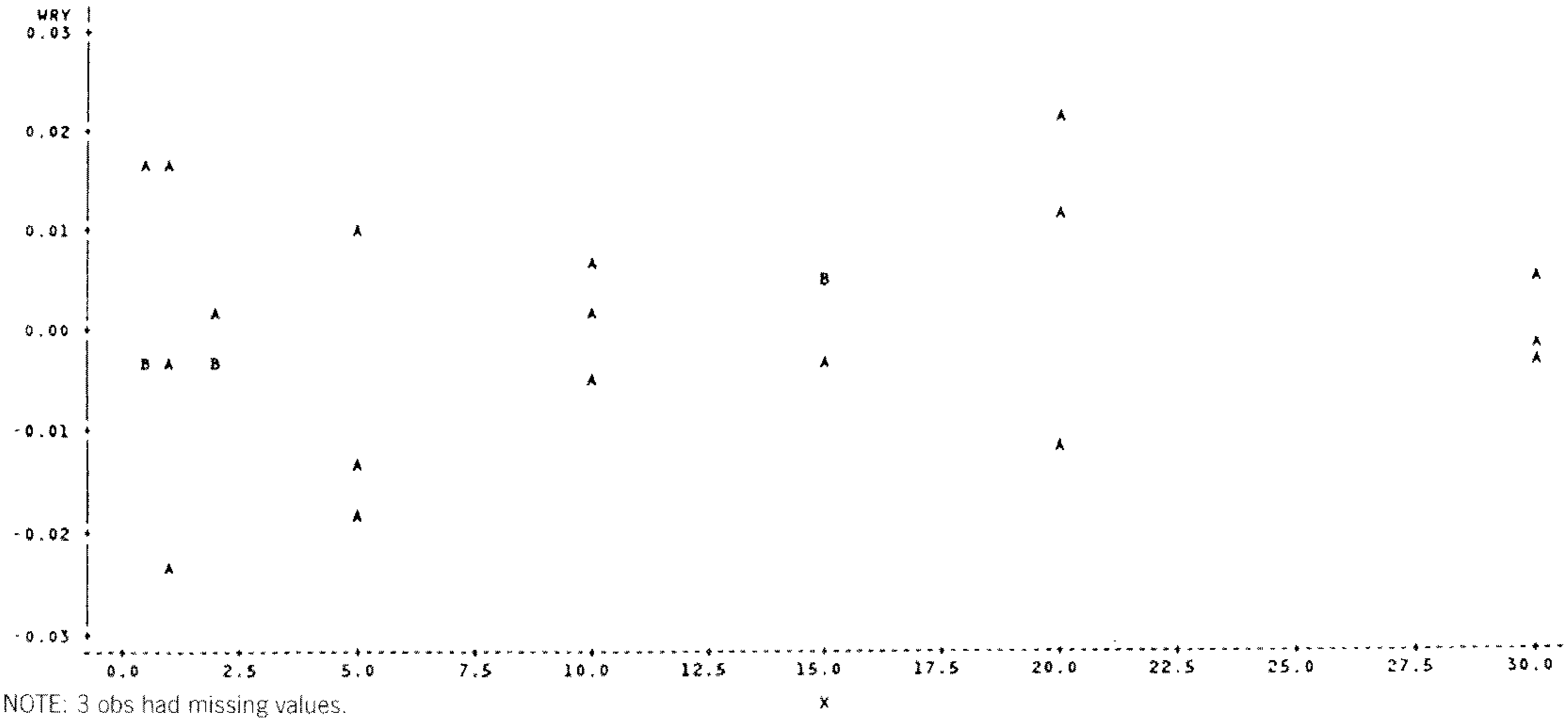
DISCUSSION

As it can be seen, the calibration of an instrument involves the selection of a set of standards (or standard preparations). The CGMP indicates



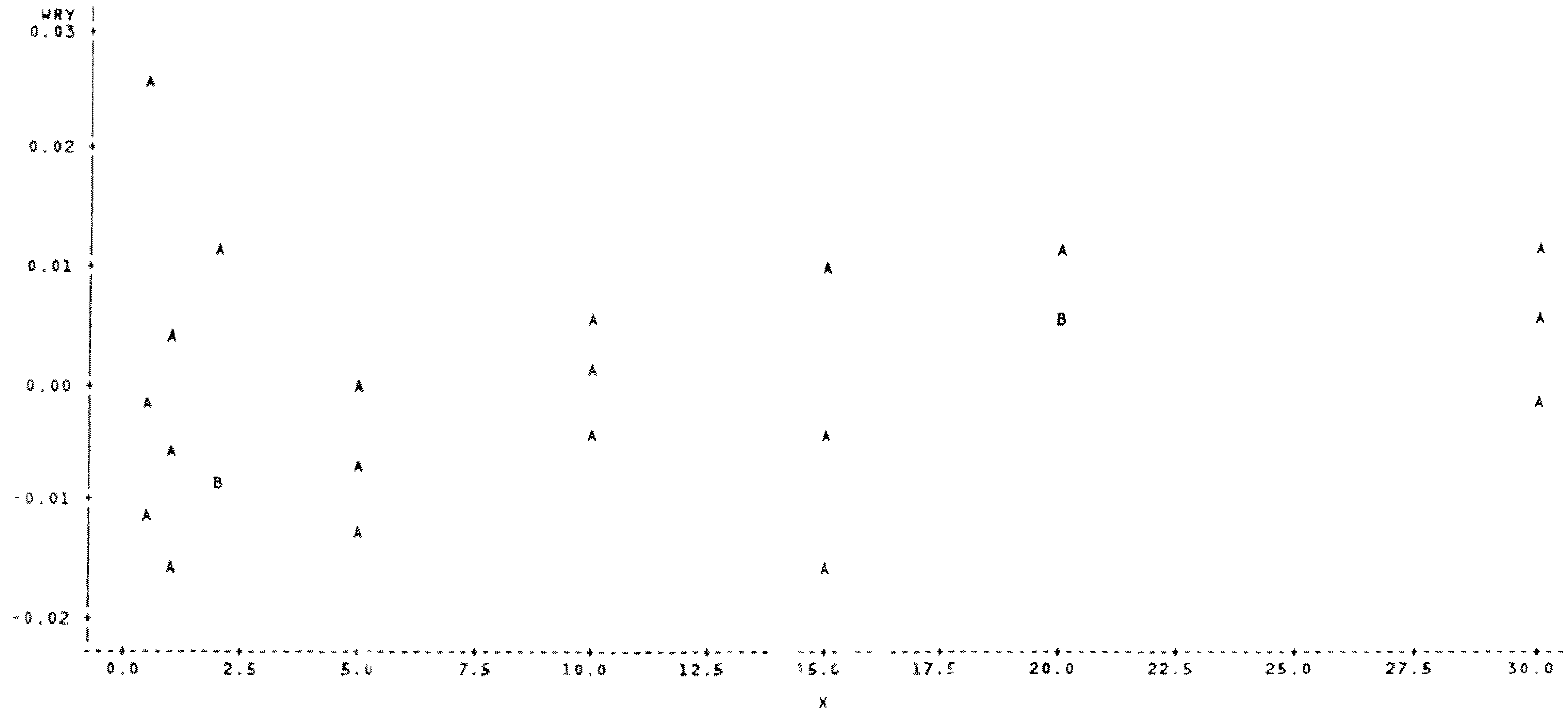
NOTE: 9 obs had missing values.

Figure 7. Plot of residual peak response versus concentration from model 1, weight=$1/x^2$, 3-day combined. (Legend: A=1 obs, B=2 obs, etc.)



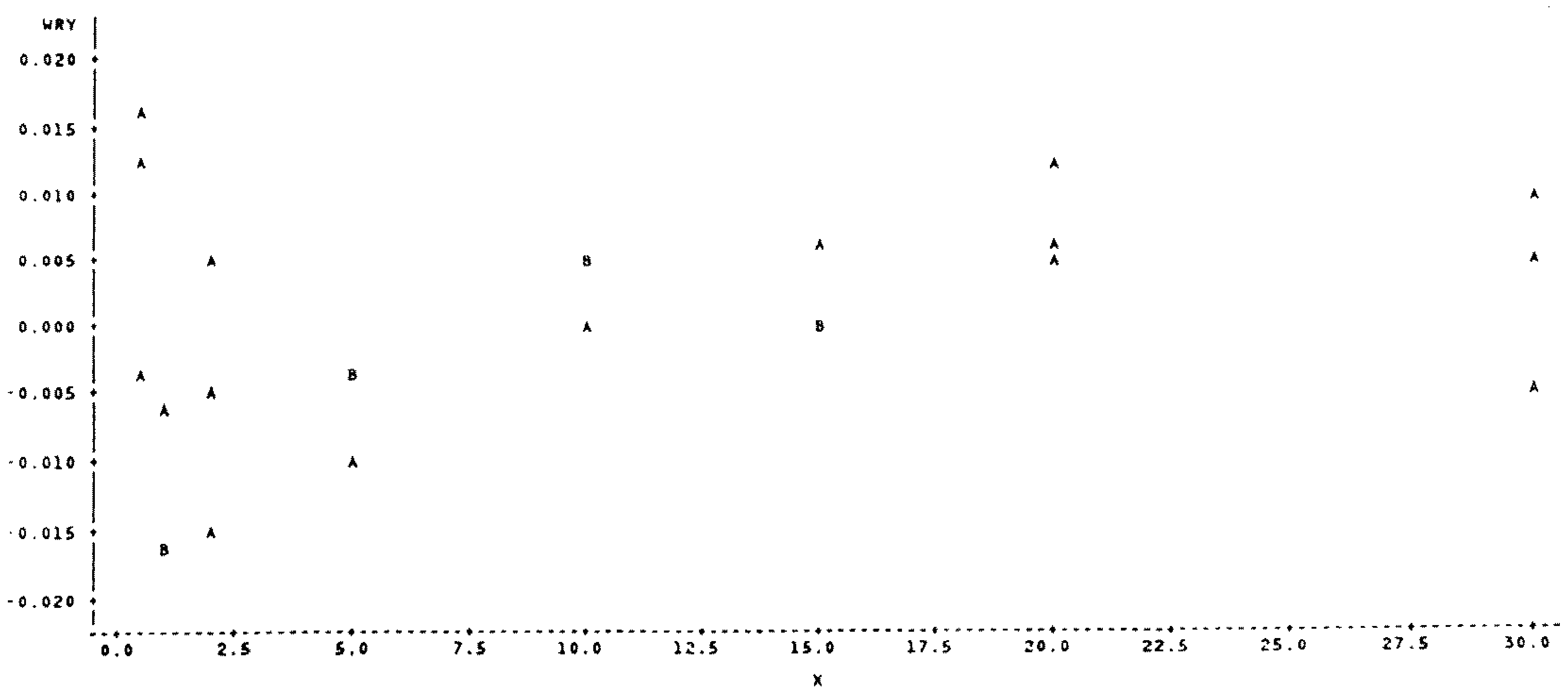
NOTE: 3 obs had missing values.

Figure 8. Plot of residual peak response versus concentration from model 1, weight=$1/x^2$, day 1. (Legend: A=1 obs, B=2 obs, etc.)



NOTE: 3 obs had missing values.

Figure 9. Plot of residual peak response versus concentration from model 1, weight=$1/x^2$, day 2. (Legend: A=1 obs, B=2 obs, etc.)



NOTE: 3 obs had missing values.

Figure 10. Plot of residual peak response versus concentration from model 1, weight=$1/x^2$, day 3. (Legend: A=1 obs, B=2 obs, etc.)

that where practical, the calibration standards used for assay development shall be in compliance with the USP / NF standards. If the USP / NF standards are not practical for the parameter being measured, an independent reproducible standard shall be used. If no applicable standards exist, an in-house standard shall be developed and used.

The procedure proposed in this paper allows one to select an appropriate model for establishing the standard curve for calibration in the development of an assay. The proposed method is a hypothesis testing procedure, which will choose the best model from five models commonly used and accepted by the FDA. The algorithm proposed in this paper is, in fact, consisted of multiple testing. It should be noted that the repeated testing will inflate the overall type I error. Appropriate multiple comparison procedures might be used to adjust the overall type I error.

Although it is not stated in this paper, note that an exploratory analysis including examination of scatter plots is usually recommended in helping to establish the model. The plots will give preliminary insights into the relationship between the response and the standard. After fitting the model, the residual plots should be examined. The choice of different models should depend on the distribution of the random error. In addition to examine the residual plots, a test for lack of fit may be applied to confirm the model selected. A lack of fit test is essential to check if the model selected is appropriate. If the test results do not support the model, then an alternative model should be considered.

If different models are selected from the proposed procedure on different days, it is an indication that the assay method might not be valid. For example, if different weights are chosen for different days, then it implies that day to day variation exists. In this case, one should examine the instrument and the validation procedure carefully. It is important to ensure the validity of the assay method in order to obtain a reliable estimate of the unknown standard from the model selected.

REFERENCES

1.USP / NF. 1990. The United States Pharmacopeia XXII and National Formulary XVII. The United States Pharmacopeial Convention, Inc., Rockville, Maryland.

2.Chow, S.C. and Shao, J. 1990. On the Difference between the Classical and Inverse Methods of Calibration. J. Roy. Stat. Soc., C, 39: 219-228.

3.Chow, S.C. and Liu, J.P. 1995. Statistical Design and Analysis in Pharmaceutical

Science: Validation, Process Controls and Stability Ch2. Marcel Dekker, New York.

4.Krutchkoff, R.G. 1967. Classical and Inverse Methods of Calibration. Technometrics. 9: 525-539.

5.Krutchkoff, R.G. 1969. Classical and Inverse Regression Methods of Calibration in Extrapolation. Technometrics. 11: 605-608.

6.Chow, S.C. and Tse, S.K. 1991. On Variance Estimation in Assay Validation. Statistics in Medicine. 10: 1543-1553.

7.SAS. 1993. SAS / STAT User's Guide, Version 6.2, 2nd Edition, Vol.2. pp.1135-1194. SAS Institute, Cary, N.C.

化驗方法發展中模型與權數之選取

林　慧　周賢忠

國立政治大學統計系
美國必治妥－施貴寶藥廠

摘　要

本文考慮了在化驗開發時（assay development），儀器刻劃（calibration）的過程中，對標準曲線（standard curve）模型與權數（weight）的選取。通常儀器的刻劃是藉由研究已知標準與其反應值（response）的關係。而這兩者間的關係又常藉由標準曲線來描述。經由標準曲線，我們可以決定未知樣本。由於在不同已知標準的反應值可能有不同的變異（variability），且標準與其應對之反應值可能呈現非線性的關係，因此如何選擇建立標準曲線合適的模型及權數，以獲得準確的化驗結果實為重要之課題。在本文中，我們考慮了五種最常用的模型及三種可能的權數。我們提出了一個選擇模型及權數的準則。一個確認(validate)製藥化合物中之血漿化驗的實例將用來說明所提出的準則。

關鍵詞：儀器刻劃，標準曲線，化驗確認，變異。