

Inverse Prediction Using SAS® Software: A Clinical Application

Jay N. Mandrekar, PhD, Cristine Allmer, BS

Division of Biostatistics, Mayo Clinic, Rochester, MN

ABSTRACT

An important application of regression methodology is in the area of prediction. Oftentimes investigators are interested in predicting a value of a response variable (Y) based on the known value of the predictor variable (X). However, sometimes there is a need to predict a value of the predictor variable (X) based on the known value of the response variable (Y). In such situations, it is improper to simply switch the roles of the response and predictor variables to get the desired predictions i.e., regress X on Y. A method that accounts for the underlying assumptions while estimating or predicting X from known Y is known as inverse prediction. This approach will be illustrated using the PROC REG, and PROC GPLOT procedures in SAS®. The calculations for the 95% confidence limits for a predicted X from a known Y will also be presented. The macro and its application will be demonstrated using data from clinical / laboratory studies.

INTRODUCTION

Analysis of data from a study where outcome of interest (Y) is measured as a continuous variable and independent or predictor variables (X) can be measured as either continuous or categorical variables is oftentimes performed using regression methods. In general, one of the several goals of linear regression is to find the line that best predicts Y from X. Linear regression analyzes the relationship between two variables, X and Y. For each subject (or experimental unit), we collect data on both X and Y and we want to find the best straight line through the data. Linear regression does this by finding the line that minimizes the sum of the squares of the vertical distances of the points from the line. In some situations, the slope and/or intercept have a scientific meaning. In other cases, we use the linear regression line as a standard curve to find new values of X from Y, or Y from X. Note that linear regression does not test whether collected data are linear. It assumes that the data are linear, and finds the slope and intercept that make a straight line best fit the data. An extension of simple linear regression is multiple regression where the goal is to learn more about the relationship between several independent or predictor variables and a dependent or outcome variable.

Applications of regression analysis exist in almost every field. In economics, the dependent variable might be a family's consumption expenditure and the independent variables might be the family's income, number of children in the family, and other factors that would affect the family's consumption patterns. In political science, the dependent variable might be a state's level of welfare spending and the independent variables measures of public opinion and institutional variables that would cause the state to have higher or lower levels of welfare spending. In sociology, the dependent variable might be a measure of the social status of various occupations and the independent variables characteristics of the occupations (pay, qualifications, etc.). In psychology, the dependent variable might be individual's racial tolerance as measured on a standard scale and with indicators of social background as independent variables. In education, the dependent variable might be a student's score on an achievement test and the independent variables characteristics of the student's family, teachers, or school. In this paper we will however, focus our attention to the application of regression in a clinical or laboratory setting.

USE OF REGRESSION MODELING

Regression approach is commonly used in two different settings, where the goals could be estimation or prediction. The goal of estimation is to determine the value of the regression function (i.e., the average value of the response variable), for a particular combination of the values of the predictor variables. Regression function values can be estimated for any combination of predictor variable values, including values for which no data have been measured or observed. The goal of prediction is to determine either the value of a new observation of the response variable, or the values of a specified proportion of all future observations of the response variable for a particular combination of the values of the predictor variables. Predictions can be made for any combination of predictor variable values, including values for which no data have been measured or observed. Predictions made outside the observed space of predictor variable values are referred as extrapolations which are sometimes necessary, but require caution while interpreting the results.

Regression approach is also used in calibration where the goal is to quantitatively relate measurements made using one measurement system to those of another measurement system. This is done so that measurements can be compared in common units or to tie results from a relative measurement method to absolute units.

WHAT IS NEW HERE?

Oftentimes investigators are interested in predicting a value of a response variable (Y) based on the known value of the predictor variable (X). However, sometimes there is a need to predict a value of the predictor variable (X) based on the known value of the response variable (Y). In such situations, it is improper to simply switch the roles of the response and predictor variables to get the desired predictions i.e., regress X on Y. This is because the primary assumption that X is measured without error and Y is a dependent, random and normally distributed variable is violated. A method that accounts for the underlying assumptions while estimating or predicting X from known Y is known as inverse prediction. Here we illustrate this approach using the REG and GPLOT procedures from SAS®. The calculations for the 95% confidence limits for a predicted X from a known Y will also be presented. The macro with ODS commands and its application will be demonstrated using relevant data from clinical / laboratory studies.

A WORD ON REGRESSION AND PLOTTING PROCEDURES

The REG procedure is one of many regression procedures in the SAS System. It is a general-purpose procedure for regression, while other SAS regression procedures provide more specialized applications. REG procedure is very flexible: performs linear regression with many diagnostic capabilities, selects models using one of nine methods, produces scatter plots of raw data and statistics, highlights scatter plots to identify particular observations, and allows interactive changes in both the regression model and the data used to fit the model.

The GPLOT procedure plots the values of two or more variables on a set of coordinate axes (X and Y). The coordinates of each point on the plot correspond to two variable values in an observation of the input data set. The procedure can also generate a separate plot for each value of a third (classification) variable. It can also generate bubble plots in which circles of varying proportions representing the values of a third variable are drawn at the data points.

STATISTICAL NOTATIONS AND FORMULATION

In this section we present the notations and formulation required for the calculations involved in prediction of an estimate and the 95% confidence limits. In an inverse prediction scenario goal is to predict X from known Y. These equations are incorporated in the SAS macro developed as a part of this work and we refer reader to Sokal (1995) for details on the theoretical concept behind these formulae.

$$\hat{X}_i = \text{Predicted value of } X_i \text{ for a given value of } Y_i = \frac{(Y_i - a)}{b_{Y,X}}$$

$$\text{LCL} = 95\% \text{ Lower Confidence Limit} = \bar{X} + \frac{b_{Y,X}(Y_i - \bar{Y})}{D} - H$$

$$\text{UCL} = 95\% \text{ Upper Confidence Limit} = \bar{X} + \frac{b_{Y,X}(Y_i - \bar{Y})}{D} + H$$

where,

$$D = b_{Y,X}^2 - t_{0.05, n-2}^2 s_b^2$$

$$H = \frac{t_{(0.05, n-2)}}{D} \sqrt{s_{Y,X}^2 \left[D \left(1 + \frac{1}{n} \right) + \frac{(Y_i - \bar{Y})^2}{\sum x^2} \right]}$$

$$b_{Y,X} = \text{Regression Coefficient} = \frac{\sum xy}{\sum x^2}$$

$$s_b^2 = \text{Variance of Regression Coefficient} = \sqrt{\frac{s_{Y,X}^2}{\sum x^2}}$$

$$s_{Y,X}^2 = \text{Error Mean Square}$$

Using these formulae, we can also calculate 90%, 99% etc. confidence limits by appropriate choice of t-value with respective degrees of freedom. One thing that needs to be noted here is that the confidence interval is not symmetric about the predicted value of X but is symmetric about $\bar{X} + \frac{b_{Y,X}(Y_i - \bar{Y})}{D}$.

ILLUSTRATION

We will illustrate the SAS commands and the relevant output with a Cytomegalovirus (CMV) dataset described in the next subsection.

DATA DISCRIPTION

This study involved 250 known concentrations of the CMV virus. The goal was to assess the confidence level in the new technology (PCR Light Cycler) developed by one of the Diagnostic tool development company. There were 50 samples of each of the 5 known concentration level (10, 100, 1000, 10000, 100000) run through the PCR Light cycler machine. In real life scenario when a patient's blood is analyzed through this tool all we get is a number which is expected to be an estimate of the concentration of virus. We wanted to assess how good this machine was and develop an algorithm to predict the actual level of virus that would have been in the blood of the patient based on the number provided by the Light cycler machine. Thus in this case we are trying to predict X which is actual amount of virus in patient's blood using Y which is the number given out by the PCR Lightcycler. Given the skewed nature of the data we have used logarithm transformation and all the analysis is then done on a log scale (i.e. log to the base 2). Few data points from the CMV study are presented in Table 1 below.

Table 1: Sample Data: CMV

Dose (X)	Copies (Y)
10000	11800
10000	10610
100000	95850
100000	101550
1000000	966000
1000000	885000
10000000	9860000
10000000	10305000
100000000	100750000
100000000	103700000

A scatter plot of known concentration put in the Lightcycler (Dose) versus the value given as an output from the lightcycler machine (Copies) is plotted as a first step in the analysis (refer Figure 1). This figure shows that the data needs to be transformed before fitting a linear regression model. It is common to use logarithmic transformation prior to running linear regression models in clinical microbiology settings. The result of this transformation is displayed in Figure 2.

Figure 1: Scatter Plot of Known Concentration (Dose) versus Concentration Given by Lightcycler (Copies)

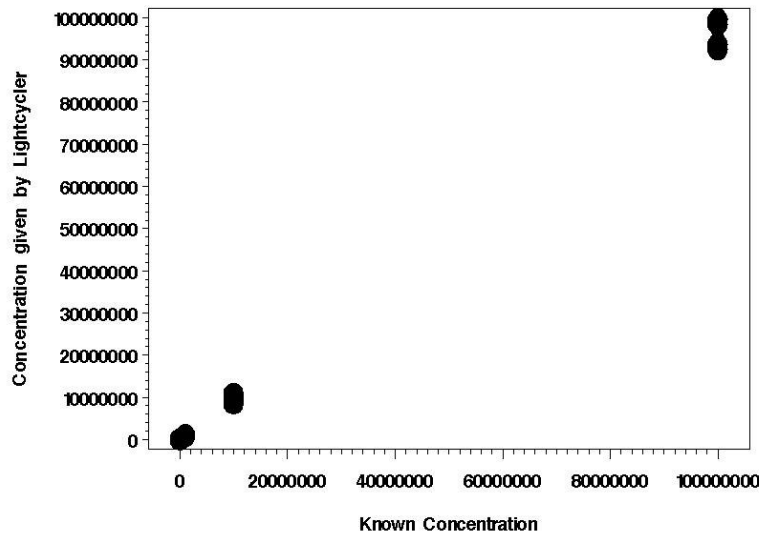
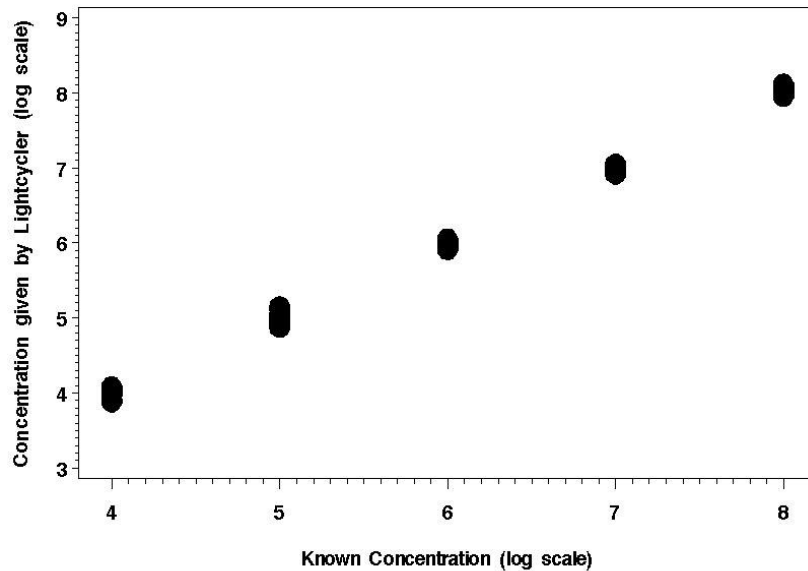


Figure 2: Scatter Plot of Known Concentration (Dose) versus Concentration Given by Lightcycler (Copies) Using Log Transformed Data



REGRESSION ANALYSIS

Output 1: Regression Model with Concentration given by Lightcycler (Y) as a Dependent Variable

The REG Procedure
Dependent Variable: lcopies

Number of Observations Read	250
Number of Observations Used	250

n

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	500.69774	500.69774	542600	<.0001
Error	248	0.22885	0.00092278		
Corrected Total	249	500.92659			

S^2_{yx}

Root MSE	0.03038	R-Square	0.9995
Dependent Mean	6.00055	Adj R-Sq	0.9995
Coeff Var	0.50624		

\bar{Y}

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	1	-0.00364	0.00837	-0.43	0.6645	-0.02013	0.01286
ldose	1	1.00070	0.00136	736.61	<.0001	0.99802	1.00337

a

B_{yx}

S_b

As a second step in the analysis, we fit a linear regression model using following SAS commands that utilize REG procedure and ODS from SAS. Here &yvar is a concentration given by Lightcycler and &xvar is known concentration.

```
PROC REG DATA = cmv CORR;
  MODEL &yvar = &xvar / CLB;

ODS OUTPUT ParameterEstimates = ParEst1
  ANOVA = Anova1 NObs = NumObs1;
```

The output produced by the above commands is presented in Output 1. The results from the linear regression analysis allow us to get the estimates of $S_{Y,X}^2$, \bar{Y} (given by Ybar), a, $b_{Y,X}$, s_b and sample size n.

Output 2: Regression Model with Known Concentration Level (X) as a Dependent Variable

The REG Procedure
Dependent Variable: Idoss

Number of Observations Read	250
Number of Observations Used	250

Analysis of Variance					
Source	DF	Sum of Squares	Mcan Square	F Value	Pr > F
Model	1	499.77157	500.77157	542600	<.0001
Error	248	0.22843	0.00092127		
Corrected Total	249	500.0000			

SS_{XX}

Root MSE	0.03035	R-Square	0.9995
Dependent Mean	6.00000	Adj R-Sq	0.9995
Coeff Var	0.50582		

Xbar

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	1	0.00637	0.00836	0.76	0.4466	-0.01009	0.02284
lcoptes	1	0.99885	0.00136	736.61	<.0001	0.99618	1.00152

As a third step in the analysis, we fit a linear regression model using following SAS commands that utilize REG procedure and ODS from SAS. As in previous step &yvar is a concentration given by Lightcycler and &xvar is known concentration. Note here that we are fitting a regression of X on Y to get additional 2 pieces required in the computation of an estimate and 95% confidence limit using inverse prediction methodology.

```
PROC REG DATA = &dsname ALL;
  MODEL &xvar = &yvar / CLB;

ODS OUTPUT ANOVA = Anova1 Simple Statistics = Stats2;
```

The output produced by the above commands is presented in Output 1. The results from the linear regression analysis allow us to get the estimates of $SS_{X,X}$ (i.e. $\sum x^2$) and \bar{X} (given by Xbar).

COMPUTATION OF PREDICTED VALUE AND 95% CONFIDENCE LIMITS

Based on various necessary estimated obtained from steps 2 and 3 above we can then use following SAS commands to calculate the predicted value of the concentration and it's 95% confidence limits. First we calculate t-value with n-2 degrees of freedom. In the following SAS commands we have used 0.975 to get an appropriate t-value required in the computation of 95% confidence limit.

```
tvalue = abs(tinv(0.975, (NObsUsed-2)));
```

Next set of commands allow us to import all statistics into the inverse prediction equation for X and for the 95% CI.

Note that up until this point we are dealing with log transformed data. Thus all the estimates and respective confidence limits are still on the log scale.

```
Xih = (Yi-a) / Byx;  
D = Byx**2 - (tvalue**2 * Sb**2);  
H = (tvalue / D) * sqrt ( S2yx * (D * (1+(1/NObsUsed)) + ((Yi - Ybar)**2) / Sxx));  
  
LL = Xbar + ((Byx * (Yi - Ybar)) / D) - H;  
UL = Xbar + ((Byx * (Yi - Ybar)) / D) + H;
```

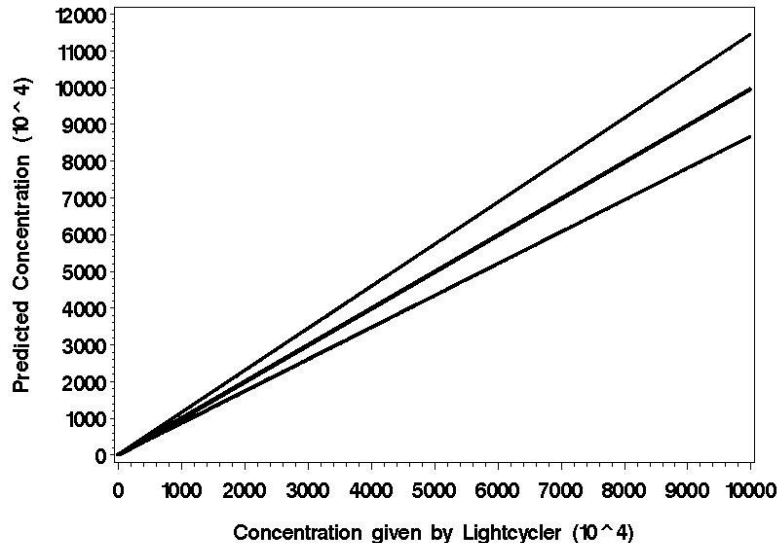
As a final step, we then convert the predicted value and 95% confidence interval to the original scale as follows.

```
Xihat = 10**Xih;  
LCL = 10**LL;  
UCL = 10**UL;
```

A plot of the inverse prediction estimates along with a 95% confidence limits on the original scale is generated using following commands and the plot is given in Figure 3.

```
PROC GPLOT;  
    PLOT (Xihat LCL UCL)* dose;  
RUN;
```

Figure 3: Inverse Prediction Estimates along with 95% Confidence limits



CONCLUSION

In this manuscript, we have provided an overview of the theory of inverse prediction, along with calculations for the 95% confidence limits for a predicted X from a known Y. The inverse prediction approach applied to the CMV dataset can be used to get an estimate of a true concentration level of CMV or other viruses in the blood of a respective patient, along with 95% confidence limits that can be used by other labs as a reference. As with any prediction model, however, a larger sample size covering a wider range of viral concentrations on a log scale is necessary to avoid extrapolation.

REFERENCES

1. Sokal R.R. and Rohlf F.J. *Biometry*, 3rd edition, 1995, W.H. Freeman and Company.
2. Smith TF, Espy MJ, Mandrekar J, Jones MF, Cockerill FR, Patel R. Quantitative real-time polymerase chain reaction for evaluating DNAemia due to cytomegalovirus, Epstein-Barr virus, and BK virus in solid-organ transplant recipients. *Clin Infect Dis*. 2007 Oct 15;45(8):1056-61.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Address all correspondences to:

Jay Mandrekar, Ph.D.

Mayo Clinic, Division of Biostatistics

200 First Street SW

Harwick 7

Rochester MN 55905

Phone: (507) 266 0573

Fax: (507) 284 9542

Email: mandrekar.jay@mayo.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.