Chapter (2)

The Nature of Calibration

2.1Introduction

Statistical calibration analysis provide a way to predict a quantity from the observation of another one by using a dose-response type relationship. The problem occurs in biological sciences when the quantity to be calibratied is hard or expensive to measure or is not observable.

It is not important in any topic of calibration to distinguish between absolute and comparative calibration .These two activities are both called calibration, they are conceptually different and lead to different issues in statistical modeling.

In absolute calibration a quick or non –standard measurement is either known or made with negligible error. With comparative calibration one instrument or measurement technique is calibrated against another with neither one being inherently a standard so that there is no standard measurement X.we discuss here absolute calibration.

2.2Mathematical Formulation of the Univariate Calibration problem

Let the true values associate with the standard and test method be designated by ξ and η respectively .We assume $\eta = f(\xi)$ and $f(\xi) = \beta_0 + \beta_1\xi$, where β_0 and β_1 are the intercept and slope parameter respectively.

In the first stage of the calibration process, the calibration experiment , n pairs of observations (X_i, Y_i) are obtained where X_i and Y_i are observed values of $\xi_i$and η_i respectivel

$$Y_i = \eta_i + \varepsilon_i \qquad i = 1,2,\dots,n$$
$$X_i = \xi_i + \delta_i \qquad i = 1,2,\dots,n \qquad (1)$$

Where εi and δi are experimental errors. In absolute calibration problem $\delta i = 0$ for all i. Produces the following model

$$Y_i = \eta_i + \varepsilon_i = f(X_i) + \varepsilon_i \qquad i = 1,2,\dots,n \qquad (2)$$

In the case of the linear calibration problem this becomes:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1,2,\dots,n \qquad (3)$$

The next assumption which is model is that they $\varepsilon i's$ are independent normal random variables with mean 0 and variance ${\sigma_1}^2$.

Having established the calibration curve /line we proceed to the second stage of the calibration process. A sample is presented with a

specific unknown value η and one or more measurements are made using the test method from which are obtained the

$$\grave{Y_j} = \eta_i + \grave{\varepsilon_j} = f(X_i) + \grave{\varepsilon_j} \qquad j = 1,2,\dots,m \quad \dots (4)$$

$$= \beta_0 + \beta_1 \xi + \grave{\varepsilon_j} \qquad j = 1,2,\dots,m \dots (5)$$

In the linear calibration problem,where $\grave{\varepsilon_1}, \grave{\varepsilon_2}, \dots, \grave{\varepsilon_m}$ are independent normal random variables with mean 0 and variance ${\sigma_2}^2$. Whrer ${\sigma_1}^2 = {\sigma_2}^2 = \sigma^2$.

Given the data from first and second stages. Inferences are now made about the unknown ξ that corresponds to η for the sample being measured. For the linear model ξ is given by:

$$\xi = \frac{(\eta - \beta_0)}{\beta_1} \dots (6)$$

2.3The Classical and Inverse Approaches to Calibration

2.3.1The Classical Estimator

Eisenhart (1939) set the stage for classical investigations of absolute calibration problems. His analysis and solution of the inverse estimation problem has come to be called classical. Eisenhart obtained his estimate of ξ by considering the regression of Yon X.

$$E(Y/X = x) = \beta_0 + \beta_1 x$$

The estimated regression line of Yon X is given by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$= \overline{Y} = \frac{S_{xy}}{S_{xx}}(X - \bar{x}) \quad \dots (7)$$

Where

$$S_{xy} = \sum_i (x_i - \bar{x})(Y_i - \bar{Y})$$

$$S_{xx} = \sum_i (x_i - \bar{x})^2$$

Eisenhart then inverted equation (2.6) to give an estimator of ξ, the unknown X, which has since become known as the classical estimator .Let it be denoted by ξ_c.Then

$$\xi_c = \bar{x} + \frac{S_{xx}}{S_{xy}}(\hat{\bar{Y}} - \bar{Y})$$

Where $\hat{\bar{Y}}$ is the mean of the m observations at the prediction stage.If one makes the assumption of normal errors in models (2.2)and(2.3) ,then ξ_c is the maximum likelihood estimator of ξ. Eisenhart also produced and

interval estimate for ξ based on the t_distribuation with $(n-2)degrees\ of\ freedom.$

Feiller (1954) produced interval estimates for ξ identical to those of Eisenhart using a fiducial argument. Fieller showed that the calibration problem could be reduced to considering the ratio of the means of two normally distributed random variables.

The classical approach to interval estimation has caused consternation over the years because if the slope parameter β_1 is not significantly different from zero the interval is either the whole real line or even two disjoint semi –infinite lines. As a result of this problem,Berkson (1969) and Shulka (1972) obtained asymptotic expressions for the bias and mean square error$(M.S.E)$ of ξ_c conditional on the event $|\hat{\beta}_1| > 0$.

2.3.2. Inverse Predictions

At times, a regression model of Y on X is used to make a prediction of the value of X which gave rise to a new observation Y. This is known as an inverse prediction .We illustrate inverse predictions by two examples:

1 .Abrade association analyst has regressed the selling price of a product (Y) on its cost (X) for the 15 member firms of the association .The selling price $Y_{h(new)}$ for another firm not belonging to the trade association is known ,and it is desired to estimated the cost $X_{h(new)}$ for this firm.

2. A regression analysis of the decrease in cholesterol level (Y) against dosage of a new drug (X) has been conducted, based on observation for 50 patients. A physician is treating a new patient for whom the cholesterol level should decrease by $Y_{h(new)}$.It is desired to estimate the appropriate dosage level decrease $Y_{h(new)}$.

The inverse prediction problem is also known as a calibration problem since it is applicable when expensive , and time-consuming measurements (X) based on n observations.

The resulting regression model is then used to estimate for a new approximate measurement $Y_{h(new)}$.

In inverse prediction model (3) is assumed as before:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1,2,\dots,n$$

The estimated regression function based on n observations is obtained as usual:

$$\hat{Y}_i = b_0 + b_1 X_i \quad (8)$$

A new observations $Y_{h(new)}$ becomes available, and it is desired to estimate the level $X_{h(new)}$ which gave rise to this new observation .A natural point estimator is obtained by solving (2.7) for X ,given $Y_{h(new)}$:

$$\hat{X}_{h(new)} = \frac{Y_{h(new)} - b_0}{b_1} \qquad b_1 \neq 0$$

Where $\hat{X}_{h(new)}$ denotes the point estimator of the new level $X_{h(new)}$.

$\hat{X}_{h(new)}$ is, indeed the maximum likelihood estimator of $X_{h(new)}$ for regression model(3).

It can be shown that approximate $1-\alpha$ confidence limits for $X_{h(new)}$ are:

$$\hat{X}_{h(new)} \pm t(1-\alpha/2; n-2)s(\hat{X}_{h(new)}) \qquad (9)$$

Where :

$$s^2(\hat{X}_{h(new)}) = \frac{MSE}{b_1^2}\left[1 + \frac{1}{n} + \frac{(\hat{X}_{h(new)} - \bar{X})^2}{\sum (X_i - \bar{X})^2}\right]$$

2.4Using Matlab program for linear regression calibration

2.4.1Introduction

The name MATLAB stands for MATrix LABorato ry. MATLAB was written originally to provide easy access to matrix software developed by the LINPACK (linear system package)and EISPACK (Eigen system package) projects.

MATLAB [1] is a high-performance language for technical computing. It integrates computation, visualization, and programming environment. Furthermore, MATLAB is amodern programming language environment: it has sophisticated data structures, contains built-in editing and debugging tools, and supports object-oriented programming. These factors make MATLAB an excellent tool for teaching and research.

MATLAB has many advantages compared to conventional computer languages (e.g. ,C, FORTRAN) for solving technical problems. MATLAB is an interactive system whose basic data element is an array that does not require dimensioning. The software package ehas been commercially available since 1984 and is now considered as a standard tool at most
universities and industries worldwide.

It has powerful built-in routines that enable a very wide variety of computations. It also has easy to use graphics commands that make the visualization of results immediately available. Septic applications are collected in packages referred to as toolbox. There are toolboxes for signal processing, symbolic computation, control theory, simulation, optimization, and several other ¯fields of applied science and statistic.

2.4.2The linear regression

Her regression problem belongs to the family of the most common practical questions. The goal is to get a model of the relationship between one variable Y and one or more variables X. The model gives the part of the variability of Y taken in account or explained by the variation of X. A function f represents the central part of the knowledge. The remaining part is dedicated to the residuals, which are similar to a noise. The model is $Y = f(X) + e$

2.4.3 Regression Models:

The simplest case is the linear regression Y = aX+b+e where the function f is affine. A case a little more complicated occurs when the function belongs to a family of parametrized functions as f(X) = cos (w X), the value of w being unknown. Statistics Toolbox™ software provides tools for the study of such models. When f is totally unknown, the problem of the nonlinear regression is said to be a nonparametric problem and can be solved either by using usual statistical window techniques or by wavelet based methods.

2.4.4 Regression Applications:

These regression questions occur in many domains. For example:

- Metallurgy, where you can try to explain the tensile strength by the carbon content
- Marketing, where the house price evolution is connected to an economical index
- Air-pollution studies, where you can explain the daily maximum of the ozone concentration by the daily maximum of the temperature

Two designs are distinguished: the fixed design and the stochastic design. The difference concerns the status of X.

2.4.5 Fixed-Design Regression:

When the X values are chosen by the designer using a predefined scheme, as the days of the week, the age of the product, or the degree of humidity, the design is a fixed design. Usually in this case, the resulting X values are equally spaced. When X represents time, the regression problem can be viewed as a de-noising problem.

2.4.6 Stochastic Design Regression:

When the X values result from a measurement process or are randomly chosen, the design is stochastic. The values are often not regularly spaced. This framework is more general since it includes the analysis of the relationship between a variable Y and a general variable X, as well as the analysis of the evolution of Y as a function of time X when X is randomized.

2.4.7Monte Carlo Simulation:

Monte Carlo simulation is a computerized mathematical technique that allows people to account for risk in quantitative analysis and decision making. The technique is used by professionals in such widely disparate fields as finance, project management, energy, manufacturing, engineering, research and development, insurance, oil & gas, transportation, and the environment.

Monte Carlo simulation furnishes the decision-maker with a range of possible outcomes and the probabilities they will occur for any choice of action.. It shows the extreme possibilities—the outcomes of going for broke and for the most conservative decision—along with all possible consequences for middle-of-the-road decisions.

The technique was first used by scientists working on the atom bomb; it was named for Monte Carlo, the Monaco resort town renowned for its casinos. Since its introduction in World War II, Monte Carlo simulation has been used to model a variety of physical and conceptual systems.

2.4.8 How Monte Carlo simulation works

Monte Carlo simulation performs risk analysis by building models of possible results by substituting a range of values—a probability distribution—for any factor that has inherent uncertainty. It then calculates results over and over, each time using a different set of random

values from the probability functions. Depending upon the number of uncertainties and the ranges specified for them, a Monte Carlo simulation could involve thousands or tens of thousands of recalculations before it is complete. Monte Carlo simulation produces distributions of possible outcome values.

By using probability distributions, variables can have different probabilities of different outcomes occurring. Probability distributions are a much more realistic way of describing uncertainty in variables of a risk analysis. Common probability distributions include:

Normal – Or "bell curve." The user simply defines the mean or expected value and a standard deviation to describe the variation about the mean. Values in the middle near the mean are most likely to occur. It is symmetric and describes many natural phenomena such as people's heights. Examples of variables described by normal distributions include inflation rates and energy prices.

Lognormal – Values are positively skewed, not symmetric like a normal distribution. It is used to represent values that don't go below zero but have unlimited positive potential. Examples of variables described by lognormal distributions include real estate property values, stock prices, and oil reserves.

Uniform – All values have an equal chance of occurring, and the user simply defines the minimum and maximum. Examples of variables that could be uniformly distributed include manufacturing costs or future sales revenues for a new product.

Triangular – The user defines the minimum, most likely, and maximum values. Values around the most likely are more likely to occur. Variables that could be described by a triangular distribution include past sales history per unit of time and inventory levels.

PERT- The user defines the minimum, most likely, and maximum values, just like the triangular distribution. Values around the most likely are more likely to occur. However values between the most likely and extremes are more likely to occur than the triangular; that is, the extremes are not as emphasized. An example of the use of a PERT distribution is to describe the duration of a task in a project management model.

Discrete – The user defines specific values that may occur and the likelihood of each. An example might be the results of a lawsuit: 20%

chance of positive verdict, 30% change of negative verdict, 40% chance of settlement, and 10% chance of mistrial.

During a Monte Carlo simulation, values are sampled at random from the input probability distributions. Each set of samples is called an iteration, and the resulting outcome from that sample is recorded. Monte Carlo simulation does this hundreds or thousands of times, and the result is a probability distribution of possible outcomes. In this way, Monte Carlo simulation provides a much more comprehensive view of what may happen. It tells you not only what could happen, but how likely it is to happen.

Monte Carlo simulation provides a number of advantages over deterministic, or "single-point estimate" analysis:

- Probabilistic Results. Results show not only what could happen, but how likely each outcome is.
- Graphical Results. Because of the data a Monte Carlo simulation generates, it's easy to create graphs of different outcomes and their chances of occurrence. This is important for communicating findings to other stakeholders.
- Sensitivity Analysis. With just a few cases, deterministic analysis makes it difficult to see which variables impact the outcome the most. In Monte Carlo simulation, it's easy to see which inputs had the biggest effect on bottom-line results.
- Scenario Analysis: In deterministic models, it's very difficult to model different combinations of values for different inputs to see the effects of truly different scenarios. Using Monte Carlo simulation, analysts can see exactly which inputs had which values together when certain outcomes occurred. This is invaluable for pursuing further analysis.
- Correlation of Inputs. In Monte Carlo simulation, it's possible to model interdependent relationships between input variables. It's important for accuracy to represent how, in reality, when some factors goes up, others go up or down accordingly.

An enhancement to Monte Carlo simulation is the use of Latin Hypercube sampling, which samples more accurately from the entire range of distribution functions.