Chapter 1
INTRODUCTION AND SUMMARY

1.1 The Calibration Problem

A statistical calibration problem is a kind of inverse prediction, or inverse regression. First we consider univariate calibration, and then extend this idea to multivariate case. Suppose one wishes to measure the amount of chemical, $x$, in a sample. Direct measurement is very difficult or expensive or time consuming. Hence we consider a related measurement, $y$, which is easy and inexpensive to obtain. For that we prepare $N$ samples with known amounts of the chemical $x_i$, $i = 1, 2, \ldots, N$. Measure the corresponding value of $y_i$, $i = 1, 2, \ldots, N$. A model is fitted to these data and this model is used in a future study to estimate the true characteristic $x$ using the less expensive or easily obtained measurement $y$. We deal with the situation where $x$ is fixed (i.e., non-random). The relationship between $y$ and $x$ in the above univariate calibration problem is

$$y = a_0 + a_1 x + e$$

(1.1.1)

Hence

$$y_i = a_0 + a_1 x_i + e_i \quad i = 1, 2, \ldots, N$$

where
e_i’s are independent and e_i \sim N (0, \sigma^2), y_i \sim N (a_0 + a_1x_i, \sigma^2). a_0, a_1, \sigma^2 may or may not be known. If any of these are unknown, the value of x_i say \theta for a particular value of y_i say y_0 can be obtained using (1.1.1).

The model is known as calibration curve and estimation process is known as calibration or inverse regression.

If x and y are vectors then the above distribution becomes a multivariate normal distribution. Then

\[ y_i \sim N (A x_i, \Sigma) \quad i = 1, 2, \ldots, N \tag{1.1.2} \]

where \( y_i \) is a \( p \times 1 \) vector, \( x_i \) is a \( q \times 1 \) vector, \( A \) is an unknown \( p \times q \) parameter matrix and the variance covariance matrix \( \Sigma \) is an unknown \( p \times p \) positive definite matrix. If \( N \) independent observations \( y_1, y_2, \ldots, y_N \) are available, corresponding to the values \( x_1, x_2, \ldots, x_N \) of the variable \( x \) then let

\[ Y = (y_1, y_2, \ldots, y_N) \quad \text{and} \quad X = (x_1, x_2, \ldots, x_N) \]

Each column of \( Y \) are independent multivariate normal random vectors with \( E(Y) = AX \) and \( \text{Cov}(Y) = I_N \otimes \Sigma \) \tag{1.1.3}

where \( Y \) denotes the \( Np \times 1 \) vector obtained by writing the columns of \( Y \) one below the other. We shall assume throughout that the \( q \times N \) matrix has rank \( q \). Now consider a particular \( p \times 1 \) normally distributed
random vector \( y \) corresponding to an unknown value \( \theta \) of \( x \), and independent of \( Y \) in (1.1.3). Assuming that the same multivariate linear model as in (1.1.3) holds, we get

\[
E(y) = A\theta \text{ and } \text{Cov}(y) = \Sigma
\]  

(1.1.4)

Note that since \( A \) is a \( p \times q \) matrix, for the identifiability of \( \theta \) in (1.1.4), we need the condition of \( p \geq q \). This will be assumed throughout this dissertation. This multivariate calibration becomes univariate when \( p = q = 1 \). The problem addressed in this dissertation is the construction of confidence regions for \( \theta \) in (1.1.4).

As a generalisation of the model (1.1.4), we can consider models where \( \theta \) in (1.1.4) is a non-linear function of lower unknown parameters, denoted by \( s \times 1 \) vector \( \eta \) (\( s \leq q \)). Then instead of (1.1.4), we have the model

\[
E(y) = A \, t(\eta) \text{ and } \text{Cov}(y) = \Sigma
\]  

(1.1.5)

where \( t(\eta) \) is a \( q \times 1 \) vector valued function of \( \eta \). The column of \( X \) in (1.1.3) will now be the values of \( t(.) \) evaluated at known design points. A polynomial regression is an example of the model (1.1.5). Now the problem is the construction of confidence regions for \( \eta \) in (1.1.5) based on \( Y \) in (1.1.3) and \( y \) in (1.1.5).
Because of the varied types of practical application problems, two types of confidence regions are constructed in calibration namely, Single use confidence regions, and Multiple use confidence regions

i) Single use confidence region: If the calibration curve (1.1.3) is used to construct a confidence region for $\theta$ corresponding to a single observation $y$ in (1.1.4).

ii) Multiple use confidence region: If the calibration curve (1.1.3) is used to construct confidence regions for a sequence of $y$’s following the model (1.1.4), corresponding to possibly different values of $\theta$.

In this work, we have considered only single use confidence regions and two possibilities for the variance covariance matrix $\Sigma$ in (1.1.3). The cases are

1) $\Sigma$ is completely unknown positive definite matrix.

2) $\Sigma = \sigma^2 I_p$, where $\sigma^2$ is an unknown positive scalar.

In each case, we consider a suitable canonical form of the models (1.1.3) and (1.1.4) or the models (1.1.3) and (1.1.5).
1.2 Some Applications

Most of the applications of calibration theory appear in chemometrics, especially on quantitative spectral analysis. A number of articles on calibration have appeared in the journals *Analytical Chemistry* and *Journal of Chemometrics*. Martens and Naes (1989) give a good account of calibration, as it applies to quantitative spectral analysis. An important article on this topic is the paper by Thomas and Haaland (1990), which appeared in *Analytical Chemistry*.

Numerous applications of calibration theory have appeared in the statistics literature. We shall describe some of them below.

1. Croarkin and Varner (1982) illustrated the use of a linear calibration function to calibrate optical imaging system used to measure features in the size range .5 to 12 micrometers (μm) on photomasks for integrated circuits. In particular, one laboratory measured with four replications each, the widths of 10 reference lines on a standard prepared by the National Institute of Standards and Technology (NIST) The response variable in the straight line regression is \( y \) (measured line width) and the independent variable is \( x \) (NIST certified line width).
2. Another interesting application of linear calibration described in Smith and Corbett (1987) and data are based on the 1984 Olympic games. The problem is the accurate measurement of marathon running courses. Accurate measurement is extremely important since any serious error in the measurement of the running course can have a significant impact on the outcome of the sporting events. The length of the running course is defined as the length of the shortest legal route that a runner can follow. The method of measurement is called bicycle method. The counter readings (i.e., the number of revolutions of the front wheel) can be noted and a model can be set up establishing the relationship between the counter reading and the true distance. This model can then be used to estimate the unknown true distance after obtaining the corresponding counter reading.

3. An application described in the paper by Brown (1982), and later in the paper by Brown and Sundberg (1987). This application is concerned with a controlled calibration experiment where $p = 2$ factors, pigmentation and viscosity of paint, were controlled each at three levels in a three by three completely balanced experiment. In this example, a bivariate observation vector, representing certain
optical properties of the paint samples, was related to the viscosity through a linear model similar to (1.1.3). We can estimate the parameters in the calibration curve, i.e., in the linear model (1.1.3), using available data on known viscosity values and then it is used to estimate the unknown viscosity of another paint sample.

4. Another important application of calibration described in Oman and Wax (1984) and Oman (1988), of estimating gestational age by ultrasound measurements at two fetal bone lengths: the femur length F and the biparietal diameter (BPD) (essentially the diameter of the skull along the axis between the ears). (F, BPD) measurements were obtained for n = 1114 women whose week of pregnancy W was accurately determined, with the object of forming a calibration curve to estimate the W₀ corresponding to a future pair (F₀, BPD₀) of measurements. In this example the model corresponds to the model (1.1.3).

5. A biological study reported in Lieftinck Koeijers (1988). The problem is to estimate the age of a dead hare. It was noted that the age is related to several criteria, like body weight, eye lens weight and length of the hind foot, and furthermore, a linear relationship exists between the various criteria and age (on a log scale). The
linear model parameters can be estimated based on a calibration experiment involving several hares of known ages. The model can be used to estimate the unknown age of a dead hare on the basis of the vector of measured criteria.

6. An example of data from a controlled calibration experiment is given in Aitchison and Dunsmore (1975). Enzyme concentration in blood plasma can be determined by a long and costly laboratory method where as an auto analyser method is quick and cheap. Here to calibrate the auto analyser, nine plasma samples selected to cover the range of enzyme concentrations have each been divided into four aliquots, one aliquot being assigned to the laboratory method and the other three to separate analyser determinations.

7. Another application is the estimation of blood alcohol concentration. This deals with the calibration of breath estimates of blood alcohol concentration, based on the results of a laboratory test. Here, the $y_j$’s are breath estimates of blood alcohol concentrations and the $x_j$’s are the actual concentration of alcohol in the blood, obtained by a laboratory test. The relevant data are based on a study conducted at Acadiana Criminalstics laboratory, New Iberia, Louisiana. Here we estimate unknown blood alcohol
concentration $\theta$, after obtaining the corresponding breath estimate say $y_0$ using an appropriate fitted model between $y_i$’s and $x_j$’s. Such an estimation is preferable to the actual laboratory determination of the blood alcohol concentration, since it is much easier and faster to obtain the breath estimates. This method can be used to examine whether a driver exceeds the allowable limit of alcohol or not.

8. Statistical calibration has been widely used at the National Institute of standards and Technology (NIST) in numerous applications dealing with the calibration of measuring instruments. Picard and Bryson (1992) used calibration theory for seismic verification of threshold test ban treaty.

9. The major crop of the farmers of Kerala is rubber. This is one of the main income yielding crops in Kerala. The usual method of making rubber sheets is through the collection of rubber latex through farmer societies. This way of collection reduces the hardship of the farmers, since they need not bother about the different stages of processing of rubber sheets. But one problem remains, that is, related to the immediate calculation of the weight of the sheet from 1kg rubber latex. It is very difficult to calculate the weight of the sheet from 1kg rubber latex and it consumes much time. This is an
area where we would be able to make use of statistical calibration theory which may be of interest to farmers in the locality.

For this we can proceed as follows.

Take samples of rubber latex from each farmer. Number them respectively, and then prepare dry rubber sheet of each sample and weigh them. Let $Y$ be the weight of rubber latex and $X$ be the weight of rubber sheet. We can fit a suitable relationship between $X$ and $Y$ using available data of the form $Y = a_0 + a_1 x$. Using this model we can estimate unknown value $\theta$ of $x$ for a particular value of $y$.

Hence through the application of calibration theory we are in a position to measure the weight of each rubber sheet from the rubber latex at the earliest and again it helps the farmer and buyer to fix the price on the spot accurately. Thus this theory would be a milestone in the development of rubber plantation and business in Kerala. Similar methods can be applied for grading of rubber sheets as RMS-1, RMS-2, RMS-3, RMS-4 etc. the same method can be used for grading of spices like cardamom, determining the quality of tea leaves etc.
1.3 Review of Literature

The theme of our dissertation is the construction of single use confidence regions in multivariate calibration problems. The pioneer article addressing this aspect is due to Eisenhart (1939) which was followed by Fieller (1954). The above two papers deal with univariate calibration. After this, many papers have appeared dealing with point and confidence set estimation in univariate calibration. An excellent review of calibration theory appears in Scheffe (1973), Osborne (1991) and Brown (1993). However, the developments in multivariate calibration is fairly recent and the literature in this area, especially on the construction of confidence regions, is quite limited. Brown’s (1982) paper is the first article dealing with the construction of confidence regions in multivariate calibration. For the models (1.1.3) and (1.1.4), a classical approach and a Bayesian approach are very well explained. Point estimation and construction of confidence regions for \( \theta \) are also described in the above paper. Generally classical region is obtained by taking an appropriate statistic with known distribution and based on this statistic we can easily construct the confidence region for \( \theta \) (unknown value of the explanatory variable).
But Bayesian approach in slightly different. Bayesian approach is briefly explained in William G. Hunter and Warren F. Lamboy (1981). In the above article a Bayesian solution to the problem (1.1.1) is obtained by deriving the posterior distribution of explanatory variable namely \( p(x \mid data) \). Jose and Jickey (2005) develop Bayesian methods for constructing single use confidence regions in multivariate calibration. We will not further elaborate upon this, because we consider only classical approach throughout this thesis.

Now we examine some of the confidence regions given by different statisticians. The construction of confidence regions is firstly due to Brown (1982). From the beginning to the current developing stage of this work, the construction is being improved by each expert in the field.

Brown’s (1982) classical region can be described as follows; consider the statistic,

\[
U_1(\theta_1) = \left( \frac{N - q - p + 1}{p} \right) \left( I + \theta_1 Y_1 \theta_1 \right)^{-1} (y - y_1 \theta_1)^T S^{-1} (y - y_1 \theta_1) \tag{1.3.1}
\]

which follows a central F – distribution with \((q, N - q - p + 1) \) degrees of freedom where

\[
\theta_1 = (XX^1)^{-\frac{1}{2}} \theta_1 \quad Y_1 = YX^1 (XX^1)^{-\frac{1}{2}} \tag{1.3.2}
\]
A 100 (1 – α)% confidence region for \( \theta_1 \) (and hence for \( \theta \)) can be easily obtained using the above statistic. Some drawbacks of Brown’s region have been pointed out by Wood (1982, 1986) and by Oman (1988), one of them being that the region can be empty. The fact that the region can be empty is easily seen by noting that \( U_1(\theta_1) \) in (1.3.1) is essentially a weighted least squares criterion and its minimum value with respect to \( \theta_1 \) will be non-zero and can be arbitrarily large. In particular, the minimum value of \( U_1(\theta_1) \) can be larger than the 100(1 – α)\(^{th}\) percentile of the central F-distribution with \((q, N – q – p + 1)\) degrees of freedom.

Oman (1988) constructed a non-empty conservative confidence region for the parameter \( \eta \) in the model (1.1.5). We shall briefly describe Oman’s (1988) procedure for the construction of confidence region of \( \theta_1 \) in (1.3.2), which is based on the model (1.1.3).

Let \( U_2(\theta_1) = (y – Y_1 \theta_1)’ G (GSG)^{-} G (y – Y_1 \theta_1) \) \hspace{1cm} (1.3.3)

where \( G = Y_1 (Y_1’ Y_1)^{-} Y_1’ \) and the superscript ‘-’ in (1.3.3) denotes a generalised inverse. Oman (1988) notes that the distribution of \( U_2(\theta_1) \) depends on the unknown parameters \( B, \Sigma \) and \( \theta_1 \) and shows that there exists \( c(\theta_1) \) such that,
\[ P[U_2(\theta_1) \leq c(\theta_1)] \geq 1 - \alpha \]  

(1.3.4)

for all B and \( \Sigma \). In other words, the set

\[ \{ \theta : U_2(\theta_1) \leq c(\theta_1) \} \]  

(1.3.5)

is a conservative confidence region for \( \theta_1 \), with coverage probability at least \( 1 - \alpha \). Oman (1988) also proves that \( c(\theta_1) \) depends on \( \theta_1 \) only through \( (\theta_1'\theta_1) \) and is an increasing function of \( \theta_1'\theta_1 \). In practical applications, an upper bound will be available for \( \theta_1'\theta_1 \) and in order to implement this confidence region, Oman (1988) recommends the use of the corresponding upper bound for \( c(\theta_1) \), instead of \( c(\theta_1) \), itself. The use of such an upper bound is for computational simplicity, since no analytic expression is available for \( c(\theta_1) \), and its values are to be numerically evaluated. A serious drawback of Oman’s (1988) procedure is that it is not invariant under the transformation,

\[ Y_1 \rightarrow CY_1, \; y \rightarrow Cy \; \text{and} \; S = CSC' \]  

(1.3.6)

where \( C \) is a \( p \times p \) non-singular matrix.

Brown and Sundberg (1987) construct a likelihood based confidence region for the parameter \( \theta_1 \) in (1.3.2). They have derived the profile likelihood for \( \theta_1 \) i.e., the maximum of the likelihood assuming that \( \theta_1 \) is fixed, normalised by the likelihood maximised over all the
parameters. The profile likelihood of $\theta_1$ is thus a function of only $\theta_1$. Brown and Sundberg (1987) constructed the confidence region for $\theta_1$, based on the statistic $U_3(\theta_1)$, which is minus twice the profile likelihood. The confidence region is of the form,

$$\{\theta : U_3(\theta) \leq l_{q,\alpha}\}$$

(1.3.7)

where $l_{q,\alpha}$ is the $100(1 - \alpha)^{th}$ percentile of the central chi square distribution with q degrees of freedom. Brown and Sundberg (1987) justifies the use of $l_{q,\alpha}$ based on small $\Sigma$ asymptotics in the special case of $q = 1$, i.e., they show that when $q = 1$, $U_3(\theta_1)$ in approximately distributed as central chi-square with one degree of freedom when $\Sigma \to 0$ and $N \to \infty$. However, such an approximation can be poor in general and this restricts the applicability of this procedure to construct confidence regions.

Fujikoshi and Nishii (1984) and Davis and Hayakawa (1987) construct confidence regions for the models (1.1.3) and (1.1.4) which is satisfactory than the previous one. Let,

$$U_4(\theta) = \frac{N - q - p + 1}{q(1 + \theta^t(X^tX)^{-1}\theta)}(\hat{\theta} - \theta)^t\hat{S}^{-1}\hat{B}(\hat{\theta} - \theta)$$

(1.3.8)
where $\hat{B}$ is the least squares estimator of Bin model (1.1.3) and
\[
\hat{\theta} = (\hat{B}'S^{-1}\hat{B})^{-1}\hat{B}'S^{-1}y.
\]

Fujikoshi and Nishii (1984) showed that $U_4(\theta)$ has an asymptotic
F-distribution with degrees of freedom $(q, N - q - p + 1)$. Their results
can obviously be used to construct confidence regions for $\theta$, which are
valid asymptotically. For the special case when $\Sigma = \sigma^2 I_p$, asymptotic
procedures have also been developed by Thomas (1991) using the
asymptotic distribution of the maximum likelihood estimator of $\theta$.

An exact confidence region is derived by Mathew and Kasala
(1994) for the models (1.1.3) and (1.1.4). Their region is always non-
empty and invariant under non singular transformations. Their region
can be described as follows.

Let $Y_{i_{0_i}} = (Y_1 + y\theta'_1)(I_m + \theta_1\theta'_1)^{-\frac{1}{2}}$ and

\[
y_{o_i} = (1 + \theta'_1\theta_1)^{-\frac{1}{2}} (y - Y_1 \theta_1)
\]

(1.3.9)

Define

\[
U_5(\theta_1) = \frac{y_{o_i}^\dagger S^{-1}Y_{i_{0_i}}} {(1 + y_{o_i}^\dagger S^{-1}y_{o_i})}
\]

(1.3.10)
Mathew and Kasala (1994) show that \( \frac{(N-q-p+1)U_5(\theta_1)}{[q(1-U_5(\theta_1))]} \) has a central F-distribution with degrees of freedom \((q, N-q-p+1)\). This leads to exact confidence regions for \( \theta_1 \). However the pivot statistic used here is a rather complicated function of \( \theta_1 \) and hence the shape of their region in general is unknown i.e., it is very difficult to study any properties of such a confidence region. But being non-empty and invariant, their region should certainly be of use in hypothesis testing.

A conservative confidence region is derived by Mathew and Zha (1996), which overcome the main drawback of the confidence region given by Mathew and Kasala (1994) i.e., their confidence regions are applicable to finite samples and this appear quite satisfactory. Their region can be described as follows.

Consider the statistic \( \left( \hat{\theta}_1 - \theta_1 \right)^{\prime} A^{\prime}S^{-1}\left( \hat{\theta}_1 - \theta_1 \right) \) (1.3.11)

\[ \hat{\theta}_1 = \left( A_1^{\prime}S^{-1}A_1 \right)^{-1} A_1^{\prime}S^{-1}y \]

where \( A_1 = B(X'X)^{\frac{1}{2}} \) and \( \theta_1 \) is given in (1.3.2) if \( A_1 \) is unknown we replace \( A_1 \) by its estimator namely \( Y_1 \), \( Y_1 \) is given in (1.3.2). The pivot used here is the modification of the above statistic which is denoted by \( U_6(\theta_1) \), where
\[ U_6(\theta_1) = \]
\[
\left( \frac{N-p-q+1}{q} \right) (y-Y_i\theta_1)' S^{-1} Y_i (Y_i'S^{-1} Y_i)^{-1} Y_i'S^{-1} (y-Y_i\theta_1) \\
1 + y' \left[ S^{-1} - S^{-1} Y_i (Y_i'S^{-1} Y_i)^{-1} Y_i'S^{-1} \right] y \]

(1.3.12)

The modification is done in order to make the distribution of \( U_6(\theta_1) \) tractable. They obtained an upper bound for the 100 \( 1 - \alpha \)th percentile of \( U(\theta_1) \) say \( c(\theta_1', \theta_1) \), then the set \( \{ \theta_1 : U(\theta_1) \leq c(\theta_1', \theta_1) \} \)

(1.3.13)

is a conservative confidence region for \( \theta_1 \) with coverage probability of at least \( 1 - \alpha \). However their region satisfy all the necessary conditions (non-empty, invariant, coverage probability) for a good confidence region but their computational work is very difficult.

### 1.4 Summary of the Thesis

From the review presented in section 1.3 it is clear that there are no satisfactory single use confidence regions available for multivariate calibration problem. Even though Mathew and Zha’s (1996) region is applicable to finite samples, the computational work involved is too much. For the models (1.1.3), (1.1.4) and (1.1.5) we have succeeded in obtaining a single use confidence region which satisfies all the necessary conditions for a good confidence region and it overcomes the
difficulty in the computation of confidence region. Here we develop suitable approximations for avoiding computational difficulty.

The thesis is arranged in five chapters. In chapter 1, we introduce the calibration theory and describe some examples and applications. A review of literature is done in section 1.3, and summary of the present work is given in 1.4.

Chapter 2 deals with the development of single use confidence region for $\theta$ in model (1.1.3) and (1.1.4), and $\eta$ in the model (1.1.5), where $\Sigma$ is completely unknown. The pivot used here is $K(\theta_1)$, which is given in chapter 2. Percentiles of $K(\theta_1)$ can be bounded from above using a function of only $\theta_1'\theta_1$ say $c(\theta_1'\theta_1)$. Consequently, if $c(\theta_1'\theta_1)$ is the upper bound for the 100 $(1 - \alpha)^{th}$ percentile of $K(\theta_1)$, then the set $\{\theta_1 : K(\theta_1) \leq c(\theta_1'\theta_1)\}$ is a conservative confidence region for $\theta_1$ with coverage probability of at least $1 - \alpha$. Our main work is the computation of $c(\theta_1'\theta_1)$ in an easier way using suitable approximations. The procedure is explained in section 2.5. In section 2.4, we have described the necessary modifications to the above approach in order to construct confidence regions for $\eta$ in the model (1.1.5). It should be noted that, the canonical form used for the computation of confidence region is that in Mathew and Zha (1996).
In chapter 3 we have applied our approximations to two examples. The first example is the paint finish data of Brown (1982), where the models (1.1.3) and (1.1.4) are applicable. Our second example is the gestational age data analysed in Oman and Wax (1984) where the model (1.1.5) is applicable. The fitness of our approximations are described in detail for both the examples.

Chapter 4 is in the same spirit as chapter 2 except that we are dealing with the case where $\Sigma = \sigma^2 I_p$. In order to construct conservative confidence regions for an unknown value of the explanatory variable $x$, we are following the same steps as described in chapter 2 which are given in detail in section 4.3. A suitable approximation which reduces the effort in computational work is introduced in section 4.4.

In chapter 5 we analyse our approximations (given in section 4.4 of Chapter 4) using an example. This example is based on the data analysed in Smith and Corbett (1987), dealing with the measurement of marathon running courses.

1.5 Conclusion

A detailed and comprehensive data analysis based on actual data is reported in this dissertation. The simulation results show the accuracy of our approximation. Using this approximation we get the required
confidence region in an easier way. This shows the applicability of our results, illustrates the practical implementation of our procedures and also provides comparison among competing procedures. It is quite surprising fact that our approximation method overcomes all the computational difficulties and also it possesses all the good characteristics of other existing methods. In our work we have considered only single confidence regions and our approximation is suitable for single use confidence regions. An important unsolved problem is the minimisation of difficulties in the computational work associated with joint and multiple use confidence regions. Bayesian methods can also be developed for constructing more efficient confidence regions. This is another area which needs further research. Some such works are currently under investigation.