Chapter 1
Introduction and Review of Literature

"Man can learn nothing unless he proceeds from known to the unknown"
- Claude Bernard

1.1 Introduction

Regression analysis is widely used for prediction and forecasting. This is generally explored in order to model the relationship between different variables. The data employed in a regression analysis may be observational or experimental. Moreover, it may be time series data or cross-sectional data. Linear models are one of the most useful and popular tools in data analysis. In simple words, a univariate response is modeled as the sum of a ‘linear predictor’ and a zero mean random error term. The dependent (response) variable is considered to be a linear combination of independent (explanatory) variables. The unknown parameters are usually estimated by least-squares, that is, by finding the values of parameters that minimize the residual sum of squares. If the dependence of mean of response variable on predictor is linear, then linear regression model is very useful. It is simple to explore and provides a concise description of the data.

In general, the amount of information obtained in a regression study is affected by the appropriateness of the regression model, the amount of data collected and the values of the employed independent variables. If the dependence is far from linear, then it is not possible to establish the relationship in terms of a straight line and we extend the model by adding simple terms like square of predictor variable or making transformation of variables, but it becomes difficult to guess (or add) the most appropriate functional form. So the basic point here is: let the data show us the appropriate functional form and that is the idea behind using a scatterplot smoother.

A smoother is a tool for summarizing the trend of response measurement as a function of one or more predictor measurements. Since it does not assume a rigid form of
dependence of response variable on the set of predictors, hence it is often referred to as a tool for non-parametric regression. Many non-parametric methods do not perform well when there are a large number of independent variables in the model. The sparseness of data in this setting inflates the variance of the estimates. The problem of rapidly increasing variance for increased dimensionality is sometimes referred to as the "curse of dimensionality".

To overcome these difficulties, Stone (1985) proposed additive models. These models estimate an additive approximation to the multivariate regression function. The benefits of an additive approximation are, at least, twofold. First, each of the individual additive term is estimated using a univariate smoother and the curse of dimensionality is avoided at the cost of not being able to approximate universally. Second, estimates of the individual terms explain how the dependent variable changes with the corresponding independent variables. Although, a lot of problems can be solved using additive models, however, the link function between expected value of the dependent variable and additive predictors is missing in this approach.

In order to overcome this, Hastie and Tibshirani (1990) proposed generalized additive models. These models assume that the mean of the dependent variable depends on an additive predictor through a nonlinear link function. Generalized additive models (GAMs) permit the response probability distribution to be a member of the exponential family of distributions. Many widely used statistical models belong to this general class, including additive models for Gaussian data, nonparametric logistic models for binary data and nonparametric log-linear models for Poisson data.

Generalized additive models find applications in analyzing

a) Air pollution data
b) Survival data
c) Binary response data
d) Case-control studies
e) Dose response curves
f) Sensitivity analysis
g) Clinical trials, and
h) Observational studies.

Generalized additive models are useful in finding predictor-response relationships in many kinds of data without using a specific model. These models are used to identify and characterize the effect of potential factors on an outcome variable. They can be applied in
any setting where a linear or generalized linear model is typically used. These settings include standard continuous response regression, categorical or ordered categorical response data, count data, survival data and time series.

1.2 Overview of Classical Regression Models

1.2.1 Linear Regression Model

Suppose that $Y$ is a response random variable and $X_1, ..., X_k$ is a set of predictor (explanatory) variables. The method for estimating how the value of $Y$ depends on the values of $X_1, ..., X_k$ is given by standard linear regression model, which assumes that the expected value of $Y$ has a linear form

$$E[Y|X_1, ..., X_k] = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k.$$ 

The estimates of $\beta_0, \beta_1, ..., \beta_k$ can be obtained by least square method. However, the standard linear model works under the following assumptions:

(i) **Constant Variance**: For any value $x_i$ of the independent variable $X$, the corresponding population of potential values of the dependent variable has a variance $\sigma^2$ that does not depend on $X$. That is, the different populations of potential values of the dependent variable, corresponding to different values of $X$, have same variance.

(ii) **Independence**: Any value of the dependent variable $Y$ is statistically independent of any other value of $Y$. Equivalently, any one value of the error term is statistically independent of any other value.

(iii) **Normality**: For any value $x_i$ of the independent variable $X$, the corresponding population of potential values of the dependent variable has a normal distribution.

When the assumptions of independence, normality and constant variance get violated, certain variations of the linear regression models are inevitable.

1.2.2 Variations of the Regression Model

Problem of Autocorrelation

The assumption of independence is more likely to be violated when the data being used in a regression problem is time series data, that is, data that have been collected in a
We can use the Durbin-Watson Statistic to test for positive or negative autocorrelation (first order). This statistic is given by

\[
\text{Durbin-Watson Statistic} = d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}
\]

where \(e_1, e_2, \ldots, e_n\) are the time-ordered residuals (Ref. Bowerman and O'Connell (1990)).

Consider testing the null hypothesis

\[H_0 : \text{The error terms are not auto-correlated}\]

versus the alternative hypothesis

\[H_1 : \text{The error terms are positively or negatively auto-correlated.}\]

For level of significance \(\alpha\), Durbin and Watson have shown that

(i) if \(d < d_{L/2}^\alpha\) or if \((4-d) > d_{U/2}^\alpha\), we reject \(H_0\)

(ii) if \(d > d_{U/2}^\alpha\) and if \((4-d) > d_{U/2}^\alpha\), we do not reject \(H_0\)

(iii) if \(d_{L/2}^\alpha \leq d \leq d_{U/2}^\alpha\) and \((4-d) \leq d_{U/2}^\alpha\), the test is inconclusive

Tables containing the points \(d_{L/2}^\alpha\) and \(d_{U/2}^\alpha\) are available for different values of \(\alpha\). In case, the test is inconclusive, second order Durbin-Watson statistics is used.

Problem of Multicollinearity

When the independent variables in a regression model are interrelated or are dependent on each other, multicollinearity is said to exist among the independent variables. The solution is to remove one or more of the highly correlated independent variables. Also, there are estimation procedures that are capable of producing point estimates that are ‘better’ than the least square point estimates in the sense that they are
closer to the true values of the parameters. In case of multicollinearity, one may use standardized regression or ridge regression techniques.

**Dummy variables**

In a linear regression model, we can model the effects of the different levels of a qualitative independent variable by using dummy variables.

1.2.3 **Generalized Linear Model**

The more general form of linear model is the Generalized Linear Model (GLM). This differs from the linear model in two aspects:

(i) the distribution of the response/dependent variable can be non-normal and does not have to be continuous, that is, it can belong to the exponential family with probability density function (pdf) given by

\[
 f_Y(y; \theta, \phi) = \exp \left\{ y \theta - b(\theta) - c(y, \phi) \right\}
\]

where \( \theta \) is called the natural parameter and \( \phi \) is called the scale parameter. The Normal, Binomial, Poisson and many other distributions belong to this family.

(ii) the response variable values are predicted from a linear combination of predictor/explanatory variables, which are connected to the response variable via a link function \( g \).

In this case, the model becomes:

\[
g(E[Y|X_1, \ldots, X_k]) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k,
\]

where \( \beta_0, \beta_1, \ldots, \beta_k \) can be estimated by Fisher scoring, an iterative procedure. Generally, ‘\( g \)’ is considered to be some monotonic differentiable function.

1.2.4 **Additive Model**

Additive model generalizes the linear model by modeling the expected value of \( Y \) as

\[
E[Y|X_1, \ldots, X_k] = S_0 + S_1(X_1) + \ldots + S_k(X_k) = S_0 + \sum_{i=1}^k S_i(X_i).
\]

This model is a non-parametric model where the linear coefficients \( \beta_0, \beta_1, \ldots, \beta_k \) in linear regression model are replaced by smoothers \( S_0, S_1, \ldots, S_k \) (where \( S_i \) is an unspecified non-parametric function). A **Smoother** is an arbitrary function that appears as a smooth curve through a scatter plot of points and is estimated in a nonparametric fashion. Sometimes, an additive model may contain a combination of parametric linear functions and non-parametric smooth functions.
1.3 Generalized Additive Model

If we combine additive models and generalized linear models, we have the notion of Generalized Additive Model (GAM) written as

\[ g(E[Y|X_1,...,X_k]) = S_0 + \sum_{i=1}^{k} S_i(X_i). \]

Generalized Additive Model consists of a random component, an additive component and a link function relating these two components. The response Y, the random component is assumed to possess a density in the exponential family. The quantity

\[ \eta = S_0 + \sum_{i=1}^{k} S_i(X_i) \]

where \( S_1,...,S_k \) are smooth functions and they define the additive component. Finally, the relationship between mean \( \mu \) of the response variable and \( \eta \) is defined through a link function \( g(\mu) = \eta \) where \( \mu = E[Y|X_1,...,X_k] \).

Generalized additive models represent a method of fitting a smooth relationship between two or more variables through a scatter plot of data points. They are useful where

(i) the relationship between the variables is expected to be of a complex form, not easily fitted by standard linear models

(ii) the data is expected to suggest the appropriate functional form

(iii) the estimated regression coefficients are small, and

(iv) there exist confounding factors that are modeled using nonparametric smoothers.

1.4 Logistic Model

Logistic model is derived from the logistic function written as

\[ f(\gamma) = \frac{1}{1+e^{-\gamma}}. \]

The model is designed to describe a probability, which is always some number between 0 & 1.

Let \( Y = \beta_0 + \beta_1X_1 + \ldots + \beta_kX_k \) and define
Introduction and Review of Literature

\[ P(X) = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{k} \beta_j X_j)}} \]

Therefore, logit \( P(X) = \log \frac{P(X)}{1-P(X)} = \beta_0 + \sum_{j=1}^{k} \beta_j X_j \).

This is the link known as log odds transformation, the logit, which converts this into linear form. The coefficients \( \beta_j \)'s are estimated through maximum likelihood technique.

1.5 The Proportional Hazards or Cox Regression Model

The regression method known as Cox regression can be used to account for the effect of continuous and discrete covariates (independent variables) when the dependent variable is possibly censored. It is based on the hazard function. For \( k \) covariates, we define the regression model as

\[ h(t_i) = h_0(t_i) e^{(\beta_1 z_{i1} + \beta_2 z_{i2} + \ldots + \beta_k z_{ik})}, \]

where \( h_0(t_i) \) as the baseline hazard function (Ref. Definition 1.6.14) and \( h(t_i) \) denotes the instantaneous failure rate at time \( t_i \).

The regression coefficients represent the change in the hazard that results from the risk factor \( z_{ik} \), that has been measured.

1.6 Definitions

Definition 1.6.1: Smoother

A smoother is a tool for summarizing the trend of response measurement \( Y \) as a function of one or more predictor measurements \( X_1, \ldots, X_p \). It produces an estimate of the trend that is less variable than \( Y \) itself and hence the name smoother. An important property of a smoother is its non-parametric nature implying that it does not assume a rigid form for the dependence of \( Y \) on \( X_1, \ldots, X_p \).

Definition 1.6.2: Bin Smoother

A bin smoother, also known as regressogram, looks like a categorical smoother by partitioning the predictor values into a number of disjoint and exhaustive regions and then averaging the response in each region. Formally, we choose cut points \( C_0, \ldots, C_K \) where \( C_0 = -\infty \) and \( C_K = \infty \) and define
Introduction and Review of Literature

\( R_k = \{i; c_k \leq x_i \leq c_{k+1}\}, \quad k = 0, 1, \ldots, K - 1, \)

the indices of the data points in each region. Then

\[ s = S(y|x) \text{ is given by } s(x_0) = \frac{\text{avg}(y_i)}{\text{avg}(y_i)} \text{ if } x_0 \in R_k. \]

**Definition 1.6.3: Symmetric Neighbourhood**

A symmetric neighbourhood consisting of \( k \)-points to the left as well as to the right of \( x_i \), is denoted by \( N'(x_i) \), where

\[ N'(x_i) = \{ \max(i - k, 1), \ldots, i-1, i, i+1, \ldots, \min(i + k, n)\}. \]

It is more convenient to think not in terms of \( k \) but instead, in terms of \( w = \frac{2k+1}{n} \), the proportion of points in each neighbourhood, called the *span*.

**Definition 1.6.4: Running Mean Smoothers**

The running mean for symmetric neighbourhood is defined as

\[ f(x_i) = \frac{\sum_{j \in N'(x_i)} y_j}{|N'(x_i)|} \]

where \(|N'(x_i)|\) is the size of the neighborhood.

**Definition 1.6.5: Running Median Smoothers**

The running median is calculated by finding the median of all the \( y \) values in the neighbourhood of \( x_i \). Running Mean was replaced by running median to make the smoother resistant to the outliers in the data.

**Definition 1.6.6: Running Line Smoothers**

The running line smoother for symmetric neighbourhood, is defined by

\[ f(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i \]

where \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) are least squares estimates for the data points in \( N'(x_i) \). For a running line smoother, the symmetrical neighbourhood consists of \( 2k+1 \) points. The parameter \( k \) controls the appearance of running line smooth. Large values of \( k \) tends to produce smooth curves while small values of \( k \) tends to produce more jagged curves. Again, we consider \( w = \frac{2k+1}{n} \) which is called span.
**Definition 1.6.7: Kernel Smoothers**

A kernel smoother uses an explicitly defined set of local weights, defined by the kernel, to produce the estimate at each target value. Usually, a kernel smoother uses weights that decrease in a smooth fashion as one moves away from the target point.

The weight given to the $j^{th}$ point in producing the estimate at $x_0$ is defined by

$$S_{0j} = \frac{c_0}{\lambda} d\left(\frac{x_0 - x_j}{\lambda}\right)$$

where $d(t)$ is an even function decreasing in $|t|$.

The parameter $\lambda$ is known as window-width or bandwidth, and the constant $c_0$ is usually chosen so that weights sum to unity. A natural choice for $d$ is the standard Gaussian density, which gives Gaussian kernel smoother. However, other kernels popular in literature are based on mean squared error and minimum variance.

**Definition 1.6.8: Splines**

A spline is a smooth function that is a piecewise polynomial and possesses high degree of smoothness at the places where the polynomial pieces connect. It is a piecewise polynomial real function $S : [a, b] \to \mathbb{R}$.

Suppose an interval $[a, b]$ is composed of $k$ subintervals $[t_{i-1}, t_i]$ with $a = t_0 < t_1 < \ldots < t_{k-1} < t_k = b$. The restriction of $S$ to an interval $i$ is a polynomial $P_i : [t_{i-1}, t_i] \to \mathbb{R}$, such that

$$S(t) = P_1(t), \quad t_0 \leq t \leq t_1,$$

$$S(t) = P_2(t), \quad t_1 \leq t \leq t_2,$$

$$\vdots$$

$$S(t) = P_k(t), \quad t_{k-1} \leq t \leq t_k.$$

The highest order of the polynomials $P_i(t)$ is said to be the order of spline $S$.

**Definition 1.6.9: Regression Splines**

Regression splines represent the fit as a piecewise polynomial. The regions that define the pieces are separated by a sequence of knots or breakpoints $\zeta_1, \ldots, \zeta_k$. In addition, it is necessary to force the piecewise polynomials to join smoothly at these knots. Many
different configurations for regression splines are possible, for example, bin smoothers or piecewise cubic polynomial constrained to be continuous and having continuous first and second derivatives at the knots.

**Definition 1.6.10: Cubic Smoothing Splines**

This smoother is not constructed explicitly like those described so far, but instead emerges as the solution to an optimization problem.

Consider the following optimization problem.

Among all the functions \( f(x) \) with two continuous derivatives, find one that minimizes the penalized residual sum of squares

\[
\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b (f''(t))^2 dt
\]

where \( \lambda \) is a fixed constant and \( a \leq x_1 \leq \ldots \leq x_n \leq b \). This criterion satisfies the requirements for a scatterplot smoother. The first term measures the closeness to the data while the second term penalizes the curvature in the function.

The parameter \( \lambda \) plays the same role as the span in the running line smoother. Large values of \( \lambda \) produce smoother curves while smaller values produce more wiggly curves. At the one extreme, as \( \lambda \to \infty \), the penalty term dominates, forcing \( f''(x) = 0 \) everywhere, and thus the solution is the least square line. At the other extreme, as \( \lambda \to 0 \), the penalty term becomes unimportant and the solution tends to an interpolating twice-differential function.

**Definition 1.6.11: Locally Weighted Running Line Smoothers**

Locally weighted running line smoother is also called the Loess in S, a statistical computing language. A locally weighted straight line smoother \( s(x_0) \) using \( k \) nearest neighbours is computed in a number of steps listed below:

(i) the \( k \) nearest neighbours of \( x_0 \) are identified and denoted by \( N(x_0) \).

(ii) the distance of the furthest near-neighbour from \( x_0 \), that is,

\[
\Delta(x_0) = \max_{N(x_0)} |x_0 - x_i|
\]

is computed.

(iii) weights \( w_i \) are assigned to each point in \( N(x_0) \), using the tri-cubic weight function:
Introduction and Review of Literature

\[
W \left[ \frac{x_0 - x_i}{\Delta(x_0)} \right]
\]

where

\[
W(u) = \begin{cases} 
(1-u^3)^3, & \text{for } 0 \leq u \leq 1 \\
0, & \text{otherwise.}
\end{cases}
\]

(iv) \(s(x_0)\) is the fitted value at \(x_0\) from the weighted least-squares fit of \(y\) to \(x\) confined to \(N(x_0)\) using the weights computed in (iii).

Nearest neighbourhoods are preferred to symmetric nearest neighbourhoods, because in a neighbourhood with fixed number of points, the average distance of the points to the target point is less in the nearest neighbourhood (unless the predictors are evenly spaced).

**Definition 1.6.12: Generalized Cross Validation**

Cross Validation (CV) works by leaving points \((x_i, y_i)\) and estimating the smoother at \(x_i\) based on remaining \((n-1)\) points. The cross-validation sum of squares is

\[
CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \hat{f}_\lambda(x_i) \right)^2
\]

where \(\hat{f}_\lambda(x_i)\) indicates the fit at \(x_i\) computed by leaving out the \(i^{th}\) data point. There is a simple way to define \(\hat{f}_\lambda(x_i)\) given only smoother matrix \(S_\lambda\). The Generalized Cross Validation (GCV) can be written as

\[
GCV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \hat{f}_\lambda(x_i)}{1 - tr(S_\lambda)/n} \right)^2
\]

where \(\hat{f}_\lambda(x_i) = \sum_{j \neq i} S_{ij}(\lambda) y_j\) and \(S = \{ S_{ij} \}\) is an \(n \times n\) matrix called smoother matrix.

**Definition 1.6.13: Survival Function**

Let \(f(t)\) be the probability density function of a lifetime random variable \(T \in [0, \infty)\) with cumulative distribution function as \(F(x)\). The survival function is defined as

\[
S(t) = P(T \geq t) = \int_t^{\infty} f(x)dx.
\]
Introduction and Review of Literature

Definition 1.6.14: Hazard Function

The hazard function \( h(t) \) is defined as

\[
h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}
\]

Definition 1.6.15: Type-I Censoring

Experiment is run over fixed time period in such a way that an individual’s life time will be known exactly only if it is less than some predetermined value. For example, in a life testing experiment, \( n \) items may be placed on test, but a decision is made to terminate the test after a time \( L \) has elapsed.

Definition 1.6.16: Type-II Censoring

In this censoring, a total of \( n \) items are placed on test, but instead of continuing until all \( n \) items fail, the test is terminated when the \( r \)th item fails where \( r \) is decided before the experiment.

Definition 1.6.17: Random Censoring

Each individual is assumed to have lifetime \( T \) and a censoring time \( C \), with \( C \) and \( T \) independent continuous random variables.

Definition 1.6.18: Interval Censoring

By interval censored data, we mean that a random variable of interest lies in an interval, instead of being observed exactly.

1.7 Review of Literature

Generalized Additive Models (GAMs) were originally developed by Hastie and Tibshirani (1990) to blend properties of generalized linear models and Additive models (Stone (1985)). GAMs assume that mean of the response variable depends on an additive predictor through a nonlinear link function. The distribution of the response variable belongs to the exponential family. For a single covariate, such a model could be specified as \( y = f(x) + \epsilon \), where \( f(x) \) is an unspecified smooth function and \( \epsilon \) is the error term. This function is generally estimated by scatter plot smoother viz. a running mean, running median, running line, kernel or a spline (Ref. Reinsch (1967) and Cleveland (1979)).
Non-parametric regression using splines is an attractive, flexible and widely applicable approach to curve estimation (Ref. Silverman (1985), Friedman and Stuetzle (1981)).

Hastie and Tibshirani (1990) proposed the fitting of GAMs using various methods. The general method of fitting these models is backfitting algorithm proposed by Hastie and Tibshirani (1986) and local scoring algorithm. The backfitting algorithm (Ref. Huimin Liu (2008)) is a general algorithm that can fit an additive model using regression type smoothers. It can be used with different smoothers such as smoothing splines and local regression smoothers.

PROC GAM fits the generalized additive models by using a modified term of adjusted dependent variable regression, with the additive predictor taking the role of linear predictor. The GAM procedure implements the B-spline (Ref. Marx and Eilers (1998)) and local regression methods for univariate smoothing components and the thin plate smoothing spline for bivariate smoothing components. Several studies were carried out for fitting a logistic GAM model and one may refer to Liu et al. (2007). Dong (2001) and WeijjieCai (2008) studied procedure of fitting logistic GAM on Kyphosis data and Prima Indian Diabetes data (Ref. Asuncion and Newman (2007)) to investigate the effects of Kyphosis (a severe forward flexion of the spline) and diabetes with a list of variables that represent physiological measurements and medical attributes respectively. Hastie and Tibshirani (1987) also proposed the additive non-parametric form of regression model.

Multivariate functions observed with noise can be approximated by thin-plate smoothing splines. It makes no assumptions of a parametric form of the model. The TPSPLINE procedure uses the penalized least squares method to fit a non-parametric regression model. The generalized cross Validation (GCV) (Ref. Craven and Wabha (1979)) or Akaike Information Criterion (AIC) function is generally used to select the amount of smoothing. Penalized least squares regression and generalizations have been studied extensively over the years. Initiated by Kimeldorf and Wabha (1970a, 1970b, 1971), it has also been carried forward by Wabha (1990), Green and Silverman (1994) and Gu (2002) for comprehensive studies.

GAM has also been a flexible and effective technique for conducting non-linear regression analysis in time-series studies. Time series designs are extremely useful to examine association between daily apparent temperature and daily mortality counts. For many years, the effect of temperature on mortality has been the subject of numerous studies, mostly examining the impact of extreme weather events (Ref. Basu and Malig
Introduction and Review of Literature

(2011), Ellis and Nelson (1978)). Many studies have shown a positive association between daily mortality and temperature during extreme hot and cold weather. Early studies found a relation between peak of daily mortality and a concomitantly peak of temperatures (Ref. Carson et al. (2006), Kim et al. (2006)). Mortality due to extreme temperatures has also been reported by Braga (2001).

In these studies, the elderly and those suffering from underlying medical conditions such as circulatory and respiratory diseases were found to be more vulnerable (Ref. Goldberg et al. (2011) and Hajat et al. (2002)). Most of the existing research has been conducted in Europe (Ref. Baker (2008), Ballester et al. (2011), McKay (1989)), Spain (Ref. Ballester et al. (1997)), Dutch (Ref. Huynen et al. (2001)), Germany (Ref. Hoffmann et al. (2008)), Canada (Ref. Goldberg et al. (2011)) and America (Ref. Anderson et al. (2009), Basu et al. (2009), Curriero et al. (2002), Ellis et al. (1978), Ostro (2009)). Very few studies in Asia, South Korea (Ref. Ha et al. (2011), Mckee (1989)), India (Ref. Kumar et al. (2010), Sharma et al. (2013)) which are rapidly developing and where effects of temperature on mortality are increasing, have been conducted. The effect of various temperature indicators on different mortality categories in a subtropical city of Brisbane has also been explored recently (Ref. Yu et al. (2011)). Reviews of epidemiologic studies from 2001 to 2008 have been conducted by Basu (2009). Modelling temperature effects on mortality using multiple segmented relationships with common break points was conducted by Muggeo (2008).

The analysis of survival data requires special techniques because the data are almost always incomplete and familiar parametric assumptions might be unjustifiable. Investigators follow subjects until they reach a pre-specified endpoint (for example, death or occurrence of some event). However, subjects sometimes withdraw from a study due to migration or lost to follow up or the study is completed before the endpoint is reached. In these cases, the survival times (also known as failure times) are censored and subjects survive to a certain time beyond which their status is unknown. The uncensored survival times are sometimes referred to as event times. Methods of survival analysis must account for both censored and uncensored data.

Many types of models have been used for survival data. Two of the popular types of models are the accelerated failure time model (Ref. Kalbfleisch and Prentice (1980)) and the Cox proportional hazards model (Ref. Cox (1972)). Each has its own assumptions about the underlying distribution of the survival times. Two closely related functions often used to describe the distribution of survival times are the survival
function and the hazard function. The accelerated failure time model assumes a parametric form for the effects of the explanatory variables and usually assumes a parametric form for the underlying survival function. Cox’s proportional hazards model also assumes a parametric form for the effects of the explanatory variables but it allows an unspecified form of the underlying survivor function. For Cox model, the partial likelihood is used as the likelihood which is justified by Sinha et al. (2003). PHREG generates a chain of posterior distribution samples by Gibbs sampler, using the adaptive rejection sampling algorithm to sample each parameter value from its full conditional distribution (Ref. Gilks and Wild (1992), Gilks et al. (1995)).

1.8 Outline of Thesis

We explore problems related to smoothing and estimation in Generalized Additive Models. Special attention has been given to well known distributions like Binomial, Poisson, Normal, Exponential and Gamma because of their obvious importance in real life situations. Various regression techniques have also been discussed. The results are expected to be practically useful in the areas of observational studies, clinical trials, air pollution and analysis of survival data. Programs have been written in Matlab, R and SAS for specific analysis of real life data. The results derived in this thesis are presented in the form of seven chapters including future directions and conclusions.

In Chapter 2, we study the various smoothing techniques and their related results. Different criteria which are useful for model building/validations have also been discussed. The running mean smoothers for Binomial, Poisson, Normal and Exponential distributions have been worked out. The running line smoothers and their properties are specified for Normal and Exponential distributions. Bias and Variance trade-off for running mean and linear smoothers is also explored. Simulation study has been conducted to compare efficiencies of various smoothers using average mean squared error.

In Chapter 3, we discuss two methods of fitting generalized additive logistic regression model, one based on Newton Raphson method and another based on iterative weighted least square method using the first and second order Taylor series expansion. The use of the GAM procedure with the specified set of weights, using local scoring
algorithm, has been applied to real life data sets. This procedure provides powerful tools for data analysis, based on nonparametric regression and smoothing techniques. The cubic spline smoother is applied to the independent variables.

In Chapter 4, we work on optimal choice of knots in TPSPLINE procedure. The procedure has been demonstrated by taking a real life data set. Another TRANSREG procedure fitting many types of linear models including ordinary, multiple, multivariate regression with variable transformations has been discussed. This procedure also fits regression functions with smooth, spline or penalized B-splines. It uses the method of alternating least-squares, that is, finding least-square estimates of the model parameters given the current scoring of the data, and then finding least-square estimates of the scoring parameters given the current set of model parameters. In this chapter, the fitting of the model through penalized B-splines using splines for AICC (Corrected AIC), CV, SBC (Schwarz’s Bayesian criterion) and GCV criteria have been discussed and compared using real life data sets.

In Chapter 5, we describe the temperature-mortality association for the period 2002-2004 in Ludhiana City of northern India by estimating the relative risks of mortality. GAM is used with Quasi-Poison function for mortality due to natural causes (excluding accidents) as the dependent variable and temperature as the independent variable with penalized spline smoothers for days of the week, relative humidity and wind speed. Current and recent day’s temperatures were the weather components most strongly predictive of mortality. Mortality risk generally decreases as temperature increases from the coldest days to a certain threshold temperature. However, after threshold temperature the risk of mortality increases. The model developed in this analysis is potentially useful for projecting the consequences of climate change scenarios and offering insights into susceptibility to the adverse effects of temperature.

In Chapter 6, we carry out the PHREG procedure using Bayesian approach. Many prior distributions for model parameters and hazard parameters have been compared using DIC (Deviance Information Criterion) and Pd (Effective number of parameters) criteria for the estimation of posterior densities. Examining of the trace plot, autocorrelation function plot and posterior density plot generated by Markov chain, help us in deciding whether the convergence has been reached or not.
Introduction and Review of Literature

In Chapter 7, the brief summary of the entire thesis with some concluding remarks along with future directions have been suggested.