Semiparametric ANCOVA Using Shape Restrictions

by

Yan Jiang

(Under the direction of Mary Meyer)

Abstract

For the semiparametric ANCOVA model, shape restrictions are applied when there is only qualitative information about the relationship between response variable and the covariate. We obtain the least-squares fits of the shape-restricted ANCOVA model and conduct a beta test statistic for the balanced shape-restricted ANCOVA model, from which the hypotheses tests about the categorical variable are performed. For the unbalanced shape-restricted ANCOVA, the hypotheses tests are performed by two Bayesian analysis methods, the Bayes credible interval method and Bayes factor. Both methods give reasonable test size and power, comparing with those of the related tests. The Bayes credible interval method is applied for simple hypotheses, and Bayes factor is a general method that can deal with any kind of hypotheses. Three real world examples are presented, both the least-square shape-restricted fit and Bayes shape-restricted fit are obtained, and the hypotheses tests about the shape-restricted ANCOVA models are performed.

Index words: ANCOVA, Shape-restricted Regression, Monotone, Convex, Increasing Concave, Test Size, Power, Simulation, Kernel Smoothing, Gibbs Sampler, Posterior Distributions, Bayes Credible Interval, Bayes Factor.

Semiparametric ANCOVA Using Shape Restrictions

by

Yan Jiang

B.S., Jilin University, China, 1995

M.S., Institute of Biophysics, Chinese Academy of Sciences, China, 1998

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

DOCTOR OF PHILOSOPHY

Athens, Georgia

2005

Semiparametric ANCOVA Using Shape Restrictions

by

Yan Jiang

Approved:

Major Professor: Mary Meyer

Committee: Ishwar Basawa
William P. McCormick
Somnath Datta
Daniel Hall

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2005

To my parents and my husband

Acknowledgments

I would like to express sincerest thanks to my advisor, Dr. Mary Meyer, for her wise guidance, constant encouragement and support throughout my research. Her kindness and patience has made the completion of this dissertation a successful and rewarding endeavor. She has been a model of excellence and scholar. I appreciate her dedication during my dissertation process and am sincerely grateful for everything she has done for me.

I am truly grateful to my committee members: Dr. Ishwar Basawa, Dr. William P. McCormick, Dr. Somnath Datta, Dr. Daniel Hall for their valuable suggestions and assistance to my dissertation and serving on my committees.

Special thanks go to Dr. Jaxk Reeves for bringing me to the Department of Statistics. It turns out that being a student in this department is a pleasure.

I would like to extend my appreciation to other faculties, staffs and graduate students in this department for their help and kindness. They make my stay at The University of Georgia both enjoyable and memorable.

I would like to thank my husband and my parents for their endless love and support. Without them, this dissertation would not have been possible.

Table of Contents

Chapter 1

Introduction

We consider a regression model with continuous response variable and two predictors: one categorical predictor and one continuous predictor. This model is the so called covariance model and the analysis related to this model is called the analysis of covariance (ANCOVA). When the main interest is in testing hypotheses concerning the categorical predictor, the continuous variable is included in the model to correctly specify the mean response and avoid biased estimation. Often the response is modeled as linear in the continuous predictor, by default. If there is no theory or evidence that the relationship is other than linear, this "simple" functional form is chosen.

In many real-world situations, there may be qualitative information about the relationship between the response and the continuous predictor. Perhaps the researcher knows only that the response is increasing with the predictor, or convex, or non-decreasing and concave. Then a shape-restricted family of functions is best performed and may provide a flexible fit to the data. Also, the shape-restricted fits can be compared with parametric models to see which one provides the best fit to the data, and can be used in hypotheses tests about the regression curves.

Taking shape restrictions into account can reduce the model mean square error or increase the power of the test, hence improve the efficiency of a statistical analysis, provided that the hypothesized shape restriction actually holds (Robertson, Wright and Dykstra, 1988, p.3). For data analysis problems, the "simplest" choice ought to be the one with fewest assumptions, not the most mathematically tractable, because minimizing assumptions reduces the possibility of lack of fit. Further, fewer assumptions result in more of the variation in response

"explained by" the continuous variable, so that the hypotheses tests concerning the categorical variable have more power.

The objective of this dissertation is to apply shape restrictions into ANCOVA and perform hypotheses tests about this model.

Chapter 2 devoted to a literature review. It describes some basic concepts of ANCOVA, semiparametric ANCOVA and the background on shape-restricted regression.

In Chapter 3, we apply shape restrictions to a semiparametric ANCOVA model, define a beta test statistic for balanced monotone and convex cases and compare its power with the traditional F-test power from corresponding parametric ANCOVA regression and an F^s-test from kernel smoothing ANCOVA regression. We also demonstrate the steps of conducting an unbalanced shape-restricted ANCOVA regression. Two examples using real data are provided.

In Chapter 4, we apply Bayesian approaches to the shape-restricted ANCOVA model. We gain all the information about the parameter estimates and some simple hypotheses tests based on the samples drawn from the posterior distribution. Through simulation studies, we compare the power from the Bayes credible interval method with that of the traditional F-test from the corresponding parametric ANCOVA regression and an F^s-test from kernel smoothing ANCOVA regression.

Bayes factor is applied in Chapter 5. A small simulation study is conducted to compare the behavior of the Bayes factor method with the traditional F-test from the corresponding parametric ANCOVA regression and an F^s-test from kernel smoothing ANCOVA regression. Three real world examples are utilized to show the feasibility and convenience of Bayes factors in hypotheses testings, when the model is extremely complicated.

Chapter 2

Literature review

2.1 Analysis of covariance

The covariance model contains both categorical and continuous predictor variables. The continuous predictor variables, usually called concomitant variables or covariates, are added in the model to correctly specify the mean response, avoid biased estimation, reduce the model error variability and enhance the power of the study.

The method of analysis of covariance (ANCOVA) is a commonly used procedure for comparing the mean responses from several groups when covariate effects are present. For example, in a clinical trial, the aim is to compare two drugs on treating Tinea pedis (athlete's foot). The investigator needs to get the information about the subject's response (signs/symptoms), which treatment he/she receives and also some other baseline information such as the subject's age. There are reasons for the investigator to believe that the difference between the two treatment groups is independent of the subject's age, but that the mean response changes linearly with age. Age, the continuous variable, is a covariate in this case.

When treatments interact with covariate, the regression lines won't be parallel and interaction terms must be included in the model. In this case, however, separate treatment regression functions must be used to compare the treatment effects.

In ANCOVA, the main goal is to compare different treatment effects. Assuming we have k treatments, then we can use $k-1$ indicator variables and an intercept (or constant term) to represent each treatment level. A simple ANCOVA model therefore can be written as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_{2i} + \cdots + \beta_k d_{ki} + \varepsilon_i$$

$i = 1, \cdots, n$, y is the response variable, x is a continuous variable and y is linear in x, d's are the indicator variables, where

$$d_{li} = \begin{cases} 1 & \text{if the } i^{th} \text{ observation is at the } l^{th} \text{ treatment level} \\ 0 & \text{otherwise.} \end{cases}$$

$l = 2, \cdots, k$. The ε_i's are independent, identically distributed normal random errors with mean 0 and variance σ^2. Assuming no interactions between x and d's, then the model implies k parallel lines, where β_0, $\beta_0 + \beta_2$, $\cdots$, $\beta_0 + \beta_k$ are the intercepts for treatment level 1, 2, $\cdots$, and k respectively. All treatment lines have a common slope β_1.

We define the balanced design as the following: at any fixed value of x_j, where x_j's are the distinct values of x, there are equal numbers $(= a)$ of y observed from each treatment group. Suppose $a = 1$ and $j = 1, \cdots, m$, k treatment levels, then the total sample size $n = mk$. We rewrite the above model in matrix format as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X}$ is the $n \times (k+1)$ design matrix,

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_m & 0 & 0 & \cdots & 0 \\ 1 & x_1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_m & 1 & 0 & \cdots & 0 \\ 1 & x_1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_m & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_m & 0 & 0 & \cdots & 1 \end{pmatrix}$$

$\boldsymbol{y}$ is the response vector, $\boldsymbol{\beta}$ is the unknown parameter vector, and $\boldsymbol{\varepsilon}$ is the random error vector. The least-squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

where $\boldsymbol{X}'$ is the transpose of $\boldsymbol{X}$. The solutions for β_i's are

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y}_1 - \frac{(\sum_{i=1}^{k}\sum_{j=1}^{m} x_j y_{ij} - n\bar{x}\bar{y})}{k\sum_{j=1}^{m}(x_j - \bar{x})^2}\bar{x} \\
\hat{\beta}_1 &= \frac{(\sum_{i=1}^{k}\sum_{j=1}^{m} x_j y_{ij} - n\bar{x}\bar{y})}{k\sum_{j=1}^{m}(x_j - \bar{x})^2} \\
\hat{\beta}_2 &= \bar{y}_2 - \bar{y}_1 \\
\hat{\beta}_3 &= \bar{y}_3 - \bar{y}_1 \\
&\vdots \\
\hat{\beta}_k &= \bar{y}_k - \bar{y}_1
\end{aligned}$$

y_{ij} is the j^{th} response at the i^{th} treatment level, $i = 1, \cdots, k$, $j = 1, \cdots, m$; $\bar{y}_i = \frac{\sum_{j=1}^{m} y_{ij}}{m}$, $\bar{x} = \frac{\sum_{j=1}^{m} x_j}{m}$, $\bar{y} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} y_{ij}}{n}$. For the linear model with unbalanced design, the solution for $\hat{\boldsymbol{\beta}}$ is also easy to obtain.

In ANCOVA, our main interest is to test different treatment effects, a special case of which is to test the hypotheses that all treatment effects are equal.

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

H_1: At least one $\beta_2, \cdots, \beta_k$ is not zero.

Let $\mathbf{1}$ be the first column of design matrix $\boldsymbol{X}$, $\boldsymbol{x}$ be the second column of $\boldsymbol{X}$, $\boldsymbol{d}_j$ be the $(j+1)^{th}$ column of $\boldsymbol{X}$, $j = 2, \cdots, k$. Define V to be the linear space spanned by vectors $\mathbf{1}, \boldsymbol{x}, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_k$, i.e., $V = \mathcal{L}(\mathbf{1}, \boldsymbol{x}, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_k)$. Define V_0 to be the linear space spanned by vectors $\mathbf{1}$ and $\boldsymbol{x}$, i.e., $V_0 = \mathcal{L}(\mathbf{1}, \boldsymbol{x})$. Then $dim(V) = k+1$ and $dim(V_0) = 2$, where dim refers to dimension. Define the test statistic

$$F^* = \frac{||\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{\mathbf{0}}||^2/(dim(V) - dim(V_0))}{||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2/(n - dim(V))} = \frac{||\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{\mathbf{0}}||^2/(k-1)}{||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2/(n-k-1)}$$

where $\hat{\boldsymbol{y}}$ is the orthogonal projection of $\boldsymbol{y}$ onto V and $\hat{\boldsymbol{y}}_{\boldsymbol{0}}$ is the orthogonal projection of $\boldsymbol{y}$ onto V_0. Under H_0, F^* has an F distribution, with degree of freedom $(k-1,\ n-k-1)$, i.e., $F^* \sim F_{(k-1,n-k-1)}$. F^* is actually a generalized likelihood ratio test statistic and can be written as

$$F^* = \frac{(SSE_0 - SSE_1)/(k-1)}{SSE_1/(n-k-1)}$$

where SSE_0 is the model error sum of squares under H_0 (the reduced model) and SSE_1 is the model error sum of squares under H_1 (the full model). We reject H_0 if F^* is large.

Another important test should be done as well: a test for non-parallel lines, i.e., to test if the interaction between the treatment factor and the covariate is significant. The role of this test is to check whether the simple ANCOVA model is appropriate.

2.2 Semiparametric analysis of covariance

Recently, more interest has been focused on semiparametric ANCOVA models, since they can provide greater flexibility to explain the data. A semiparametric ANCOVA model specifies the relationship between the response and the covariate term in a nonparametric form. In the previous example, suppose that in addition, the investigator believes that the response is linearly related to treatment effect, but the effect of age on the response is unknown. Then a semiparametric ANCOVA model will serve well in this case. Speckman (1988) studied the following semiparametric ANCOVA model

$$y_i = \boldsymbol{\xi}_i'\boldsymbol{\beta} + f(t_i) + \varepsilon_i, \quad 1 \le i \le n \tag{2.2.1}$$

where the $\boldsymbol{\xi}_i$ are fixed known p vectors, $\boldsymbol{\beta}$ is an unknown vector of parameters, t_i are known real numbers and f is a smooth but unknown function. The error terms ε_i's are assumed uncorrelated with mean zero and variance σ^2. In this model, the response y is linear in $\boldsymbol{\beta}$, but the relationship between y and t is hard to formulate. Due to this property, model (2.2.1) is also called "partial linear model" (Speckman 1988) or "partly linear model" (Heckman 1986). There are several approaches available to estimate $\boldsymbol{\beta}$ and the function f. Engle *et al.*

(1986), Green *et al.* (1985), Shiau *et al.* (1986) and Wahba (1984) studied the method of penalized least squares. Estimates are obtained by minimizing over $\boldsymbol{\beta}$ and f through

$$\sum_{i=1}^{n} [y_i - \boldsymbol{\xi}_i'\boldsymbol{\beta} - f(t_i)]^2 + \lambda J(f), \tag{2.2.2}$$

where $J(f)$ is an arbitary function of f, with inerpretation as a roughness penalty function for overfitting the data (Green 1985). For example, if $0 \leq t_i \leq 1$, $J(f)$ can be chosen as

$$J(f) = \int_0^1 \left[f^{(m)}(u)\right]^2 du;$$

when $m = 2$, it will fit a cubic smoothing spline. Solving this minimization problem produces simultaneous estimates of $\boldsymbol{\beta}$ and f. The parameter λ, also called the "tuning parameter", is a constant chosen by the statistician for a suitable fit. Denby (1986) and Engle *et al.* (1986) suggested a generalized cross-validation method to choose λ. Wahba (1984) called the estimate of f the "partial smoothing spline". Other approaches include arbitrary scatterplot smoothers method (Hastie and Tibshirani, 1986) and the projection method (Chen, 1988). Heckman (1986) showed that the estimates of $\boldsymbol{\beta}$ and f are equivalent to Bayes estimates under a diffuse prior on $\boldsymbol{\beta}$ and f, and under mild conditions on $\boldsymbol{\xi}$, $\boldsymbol{t}$, $\boldsymbol{\varepsilon}$ and the roughness penalty, the estimate of $\boldsymbol{\beta}$ is consistent and asymptotically normal.

Young and Bowman (1995) extended the Speckman's (1988) work by adding the possibility of different covariate effects in different groups and allowing the assumption of additivity to be tested within the more general model of different covariate effects for each group. They considered the ANCOVA model as

$$y_{ij} = \alpha_i + g_i(x_{ij}) + \varepsilon_{ij}$$

where $i = 1, \cdots, p$, $j = 1, \cdots, n_i$, $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$, $g_i(x)$ is assumed only to be a smooth function.

They constructed tests of the hypotheses of additivity and parallelism across groups as follows:

$$H_0 : y_{ij} = \alpha_i + g(x_{ij}) + \varepsilon_{ij}, \quad \alpha_1 = 0$$
$$H_1 : y_{ij} = g_i(x_{ij}) + \varepsilon_{ij}.$$

They proposed the test statistic as

$$TS = \frac{\sum_{i=1}^{p} \sum_{j=1}^{n_i} \{\hat{\alpha}_i + \hat{g}(x_{ij}) - \hat{g}_i(x_{ij})\}^2}{\hat{\sigma}^2},$$

where the α_i term is estimated via the least-squares method, $\hat{g}_i$ is the non-parametric curve estimator that can be obtained easily, for example, from a kernel method. For example, Gasser-Müller form of kernel estimator (Gasser and Müller, 1979) is

$$\hat{g}_i(x) = \sum_{j=1}^{n_i} y_{i[j]} \int_{t_{i,j-1}}^{t_{ij}} K_h(x - t)dt,$$

where $t_{i,0} = -\infty$, $t_{i,n} = \infty$, $t_{ij} = (x_{i[j]} + x_{i[j+1]})/2$ for $j = 1, \cdots, n-1$, and $n = \sum_{i=1}^{p} n_i$; $x_{i[j]}$ denotes the j^{th} largest value of x in the i^{th} group. Here, $y_{i[j]}$ denotes the value of y corresponding to $x_{i[j]}$. The estimator for σ^2 is constructed based on the assumption that the error variance is constant across groups,

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{p} (n_i - 1)\hat{\sigma}_i^2,$$

where the estimator of σ_i was proposed by Rice (1984)

$$\hat{\sigma}_i^2 = \frac{1}{2(n_i - 1)} \sum_{j=1}^{n_i - 1} (y_{i[j+1]} - y_{i[j]})^2.$$

They also proposed accurate moment-based approximations for the distribution of the test statistic.

Speckman *et al.* (2001) introduced a new test for ANCOVA when a one-sided analysis is appropriate for comparing two groups. This test is based on first smoothing the dependent variable on the covariates and then analyzing the residuals with a rank test.

2.2.1 Kernel smoothing

Speckman (1988) presented the following method of estimating $\boldsymbol{\beta}$ and f. From model (2.2.1), suppose

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{f}}$$

$$= \quad \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{K}_h(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

where $\boldsymbol{X} = (\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_n)'$ is the $n \times p$ design matrix, $\boldsymbol{f} = (f(t_1), \cdots, f(t_n))$, $\boldsymbol{K}_h$ is the kernel smoothing matrix with bandwidth parameter h.

To get the estimator $\hat{\beta}$, define

$$\tilde{\boldsymbol{X}} = (\boldsymbol{I} - \boldsymbol{K}_h)\boldsymbol{X}, \quad \tilde{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{K}_h)\boldsymbol{y},$$

$$\boldsymbol{P} = \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}',$$

$$\boldsymbol{A}_h = \boldsymbol{K}_h + \boldsymbol{P}(\boldsymbol{I} - \boldsymbol{K}_h),$$

$$\boldsymbol{U}' = (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'(\boldsymbol{I} - \boldsymbol{K}_h).$$

Then

$$\hat{\boldsymbol{\beta}} = \boldsymbol{U}'\boldsymbol{y},$$

$$SSTR = \hat{\boldsymbol{\beta}}'(\boldsymbol{U}'\boldsymbol{U})^{-1}\hat{\boldsymbol{\beta}},$$

$$RSS = \|\boldsymbol{y} - \boldsymbol{A}_h\boldsymbol{y}\|^2,$$

where $SSTR$ is the sum of squares of treatment, RSS is the residual sum of squares. Let $r = trace\,[(\boldsymbol{I} - \boldsymbol{A}_h)'(\boldsymbol{I} - \boldsymbol{A}_h)]$ be an estimate of the error degree of freedom, where $trace[\boldsymbol{A}] = \sum_{i=1}^{n} A_{ii}$ for an $(n \times n)$ matrix $\boldsymbol{A}$. Then the mean square of errors used to estimate σ^2 is

$$MSE = \frac{RSS}{r}.$$

Speckman proposed that $\hat{\boldsymbol{\beta}}$ has an asymptotic normal distribution, hence an approximate F test could be used to test the treatment effects. For the hypotheses

$$H_0\text{: } \beta_1 = \cdots = \beta_p = 0$$

$$H_1\text{: not } H_0$$

the test statistic is

$$F^s = \frac{SSTR/p}{MSE}. \tag{2.2.3}$$

Speckman showed that F^s has an approximate F distribution with degree of freedom (p, r).

2.2.2 Ordinary cross-validation method

The bandwidth parameter h in kernel smoothing can be chosen by an ordinary cross-validation method. Ordinary cross validation (OCV) is based on omitting one observation at a time and estimating the function at the left-out observation. It was suggested by Allen (1974) in the context of regression and by Wahba and Wold (1975) in the context of smoothing splines. And it has been studied in detail by Härdle and Marron (1985).

The OCV method can be summarized as follows:

(1) Let $\theta_h^{[k]}(x)$ be the kernel smoothing fit at a point x, where the fit is obtained by leaving out the k^{th} observation $(x_k,\, y_k)$, and using a particular value of h.

(2) Calculate the "ordinary cross-validation function" $V(h)$, where

$$V(h) = \frac{1}{n}\sum_{i=1}^{n}\left(y_k - \theta_h^{[k]}(x_k)\right)^2 \tag{2.2.4}$$

(3) Try a range of h values; OCV estimate of h is the minimizer of $V(h)$, since both oversmoothing and undersmoothing will tend to increase the value of $V(h)$.

2.3 Shape-restricted regression

Usually, we fit a straight line when the relationship between response variable and predictor variable is linear, and apply polynomial, logarithmic or exponential regression to fit the data otherwise. However, in many situations, when the underlying regression function has a particular shape or form that can be characterized by certain order or shape restrictions, order restricted classes of regression functions will be preferred. Taking shape restrictions into account can reduce the model mean square error or increase the power of the test, hence improve the efficiency of a statistical analysis, provided that the hypothesized shape restriction actually holds (Robertson, Wright and Dykstra, 1988, p.3).

Suppose the only information we have is that the underlying function is "nondecreasing" or "convex", then a shape-restricted family of functions may provide a flexible fit to the

data. Further, the shape-restricted models can be used to compare with parametric models to see which one performs a better fit to the data.

Shape-restricted models can also be used in hypotheses tests about the regression curves. When we fit a simple linear regression model, we might also want to check for the curvature of the regression function. Typically, we will add a quadratic term in the model, then do an F-test to see if the parabola explains significantly more variation than the linear model. A more general method is to get the best convex fit, then compare this with the straight line. The latter method might be more likely to capture a convex trend in the data that is other than quadratic.

In order restricted classes, monotone regressions and convex regressions are two important cases. Simple alternatives of these models include concave, nondecreasing convex, sigmoidal, etc.

2.3.1 Examples of shape-restricted regressions

The following are two examples of shape-restricted fit. In Figure 2.1 (a), the data are generated from a jump function

$$y = \begin{cases} 0 & \text{if} \quad 0 < x < 0.5 \\ 2 & \text{if} \quad 0.5 \leq x < 1. \end{cases}$$

We fit both monotone and linear regressions to the data, where the solid line is the monotone fit and the dotted line is a simple linear fit. In Figure 2.1 (b), the data are generated from a convex function $y = 2x + 1/x$. We fit both convex and quadratic regressions to the data, where the solid line is the convex fit and the dotted line is a quadratic fit. In both cases, it's easy to see that shape-restricted regressions fit the data better, which leads to a smaller error sum of squares.

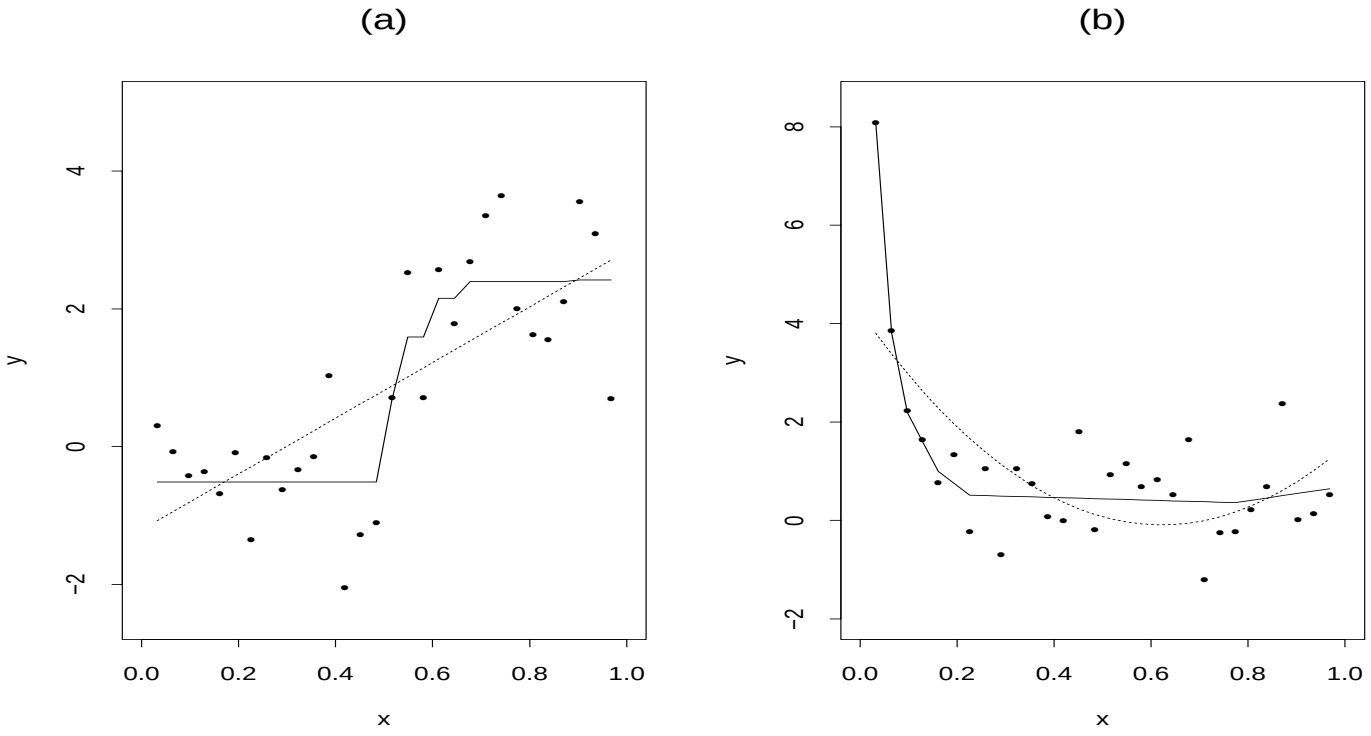

Figure 2.1: (a) Data are generated from a jump function, and fitted by both monotone and linear regressions, sample size is 30. The solid line is the monotone fit and the dotted line is a simple linear fit. (b) Data are generated from a convex function, and fitted by both convex and quadratic regressions, sample size is 30. The solid line is the convex fit and the dotted line is a quadratic fit.

2.3.2 Basic definitions

Shape-restricted inference constitutes a class of nonparametric statistics problems. Suppose we have the following model

$$y_i = f(x_i) + \sigma\epsilon_i, \quad i = 1, \cdots, n$$

where $f \in F$ shape-restricted class, σ is a known constant, ϵ_i's are independently distributed $N(0,1)$ random errors.

The shape-restricted least-squares estimator $\hat{f}$ is a function in F such that it minimizes

$$\sum_{i=1}^{n}\{y_i - f(x_i)\}^2 \text{ for all } f \in F$$

Constraint set and constraint matrix

Let $\theta_i = f(x_i)$, where the x_i's are ordered and distinct values, $i = 1, \cdots, n$. The monotone nondecreasing constraint can be written as

$$\theta_1 \leq \theta_2 \leq \cdots \leq \theta_n.$$

For convex regression and nondecreasing concave regression, we consider piecewise linear approximations to the regression function f, with knots at the x-values. The constraint for convexity can be written as

$$\frac{\theta_2 - \theta_1}{x_2 - x_1} \leq \frac{\theta_3 - \theta_2}{x_3 - x_2} \leq \cdots \leq \frac{\theta_n - \theta_{n-1}}{x_n - x_{n-1}},$$

the constraint for a non-decreasing concave assumption can be written as

$$\frac{\theta_2 - \theta_1}{x_2 - x_1} \geq \frac{\theta_3 - \theta_2}{x_3 - x_2} \geq \cdots \geq \frac{\theta_n - \theta_{n-1}}{x_n - x_{n-1}}, \quad \theta_{n-1} \leq \theta_n.$$

These inequalities define $(n-1)$ (for monotone and non-decreasing concave), and $(n-2)$ (for convex) half-spaces in $\mathbb{R}^n$, respectively. The intersection of the half-spaces can be written as a constraint set Ω, where

$$\Omega = \{\boldsymbol{\theta} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{\theta} \geq 0\}$$

$\boldsymbol{A}$ is called the constraint matrix. For monotone and non-decreasing concave regression, $\boldsymbol{A}$ is an $(n-1) \times n$ matrix; for convex regression, $\boldsymbol{A}$ is an $(n-2) \times n$ matrix. For $n = 6$, ordered and distinct x, monotone nondecreasing restriction has a constraint matrix as

$$\boldsymbol{A} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}.$$

For $n = 6$, ordered and equally spaced x, convex restriction has a constraint matrix as

$$\boldsymbol{A} = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \end{pmatrix}.$$

The constraint matrix for nondecreasing concave restriction is

$$\boldsymbol{A} = \begin{pmatrix} -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}.$$

The least-squares estimator $\hat{\boldsymbol{\theta}}$ is the orthogonal projection of the data vector $\boldsymbol{y}$ onto Ω that minimizes the Euclidean distance

$$d^2 = ||\boldsymbol{y} - \boldsymbol{\theta}||^2, \quad \boldsymbol{\theta} \in \Omega. \tag{2.3.1}$$

A set which can be expressed as the intersection of finitely many closed half spaces of $\mathbb{R}^n$ is called a polyhedral convex set (Rockafellar, 1970, p.11). This set Ω is a closed convex polyhedral cone (Robertson, Wright and Dykstra, 1988, p.15), where

1. Ω is closed in the topology induced by the metric given by (2.3.1).

2. Ω is convex: if $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Omega$ and $0 \leq \alpha \leq 1$, then $\alpha\boldsymbol{\theta}_1 + (1-\alpha)\boldsymbol{\theta}_2 \in \Omega$.

3. Ω is a cone: if $\boldsymbol{\theta} \in \Omega$ and $\alpha \geq 0$ then $\alpha\boldsymbol{\theta} \in \Omega$.

Least squares estimator $\hat{\boldsymbol{\theta}}$

Under the monotone constraint, the least-squares estimator for $\boldsymbol{\theta}$ is a step function with a closed form solution (Robertson, Wright and Dykstra, 1988, p.23)

$$\hat{\theta}_i = \min_{v \geq i} \max_{u \leq i} \frac{1}{v-u+1} \sum_{j=u}^{v} y_j, \quad i = 1, \cdots, n. \tag{2.3.2}$$

For convex and non-decreasing concave regression, the problem of finding the least-squares estimator $\hat{\boldsymbol{\theta}}$ is a quadratic programming problem. There is no known closed-form solution. But $\hat{\boldsymbol{\theta}}$ can be found using the mixed primal-dual bases algorithm (Fraser and Massam, 1989) or the hinge algorithm (Meyer, 1999t). Robertson, Wright and Dykstra (1988, p.17) proved the following important theorem about projection onto a convex cone.

Theorem 1 *The vector* $\hat{\boldsymbol{\theta}}$ *minimizes* $||\boldsymbol{y}-\boldsymbol{\theta}||^2$ *over* $\boldsymbol{\theta} \in \Omega$ *if and only if*

$$\langle \boldsymbol{y}-\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}} \rangle = 0 \tag{2.3.3}$$

and

$$\langle \boldsymbol{y}-\hat{\boldsymbol{\theta}}, \boldsymbol{\theta} \rangle \leq 0, \forall \boldsymbol{\theta} \in \Omega \tag{2.3.4}$$

where notation $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_i a_i b_i$ *refers to the vector inner product of* $\boldsymbol{a}$ *and* $\boldsymbol{b}$.

Constraint cone Ω^*

Define $\mathbf{1} = (1, \cdots, 1)'$ and $\boldsymbol{x} = (x_1, \cdots, x_n)'$, let V be the linear space spanned by $\mathbf{1}$, i.e., $V = \mathcal{L}(\mathbf{1})$, for discussion of monotone and non-decreasing concave regressions, and let V be the linear space spanned by $\mathbf{1}$ and $\boldsymbol{x}$, i.e., $V = \mathcal{L}(\mathbf{1}, \boldsymbol{x})$, for discussion of convex regression. Then $\boldsymbol{A}\mathbf{1} = \mathbf{0}$ for the first case, and in addition $\boldsymbol{A}\boldsymbol{x} = \mathbf{0}$ for the second case, where $\boldsymbol{A}$ is the corresponding constraint matrix. These imply that $V \subset \Omega$ and Ω actually contains a linear space.

Define the "constraint cone" Ω^* as the intersection of the constraint set and $V^\perp$,

$$\Omega^* = \Omega \bigcap V^\perp,$$

where $V^\perp$ is the set of all vectors in $\mathbb{R}^n$ that are orthogonal to V. By splitting Ω into two orthogonal spaces Ω^* and V, the projection of a vector $\boldsymbol{y}$ in $\mathbb{R}^n$ onto Ω therefore is the sum of the projections of $\boldsymbol{y}$ onto Ω^* and V, respectively, which can simplify the computation. Also, the edges of Ω^* are uniquely defined, provided each edge vector has unit length. However, the edges of Ω are not unique; there might be several sets of different edges that form the

same Ω. These will be shown later.

Polar cone Ω^0 of the constraint set

Each constraint set has a corresponding polar cone. The polar cone Ω^0 is defined as (Rockafellar, 1970, p.121)

$$\Omega^0 = \{\boldsymbol{\rho} : \langle \boldsymbol{\theta}, \boldsymbol{\rho} \rangle \leq 0, \quad \forall \boldsymbol{\theta} \in \Omega\}.$$

Geometrically, a "polar cone" may be defined as the set Ω^0 of points in $\mathbb{R}^n$ that form obtuse angles with all points in the constraint set Ω (Robertson, Wright and Dykstra, 1988, p.5). Alternatively, a vector $\boldsymbol{\rho}$ is in Ω^0 if the projection of $\boldsymbol{\rho}$ onto Ω is the origin. To see this, let $\boldsymbol{\rho} \in \Omega^0$ and let $\hat{\boldsymbol{\rho}}$ be the projection of $\boldsymbol{\rho}$ onto Ω, hence $\hat{\boldsymbol{\rho}} \in \Omega$. By Theorem 1,

$$\begin{aligned} \langle \boldsymbol{\rho} - \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\rho}} \rangle &= 0 \\ \Longrightarrow \langle \boldsymbol{\rho}, \hat{\boldsymbol{\rho}} \rangle &= ||\hat{\boldsymbol{\rho}}||^2 \end{aligned}$$

By the definition of the polar cone, $\langle \boldsymbol{\rho}, \hat{\boldsymbol{\rho}} \rangle \leq 0$. But $||\hat{\boldsymbol{\rho}}||^2 \geq 0$, hence $\hat{\boldsymbol{\rho}} = \mathbf{0}$ is the solution.

Edges of constraint cone and polar cone

The edges of the cone are the key to constructing the shape-restricted fits as well as the inference methods. The edges of a cone are a set of vectors in the cone such that any vector in the cone can be written as a non-negative linear combination of the edges, but an edge can not be written as the sum of two or more linearly independent vectors in the cone.

If every vector in the cone is a nonnegative linear combination of a finite collection of edges, the cone is said to be finitely generated and the edges of the finite collection are also called generators. It is well known that a convex cone is finitely generated if and only if it can be written as a finite intersection of closed half-spaces (Rockafeller,1970, p.171).

Let $\boldsymbol{\gamma}^i, i = 1 \cdots, m$ be the negative rows of the constraint matrix $\boldsymbol{A}$, i.e., $[\boldsymbol{\gamma}^1 \cdots \boldsymbol{\gamma}^m] = -\boldsymbol{A}'$, where $m = n-1$ for monotone and non-decreasing concave, and $m = n-2$ for convex.

Then the constraint set can be rewritten as

$$\Omega = \{\boldsymbol{\theta} : \langle \boldsymbol{\gamma}^i, \boldsymbol{\theta} \rangle \leq 0, i = 1, \cdots, m\}.$$

Clearly, $\boldsymbol{\gamma}^1 \cdots \boldsymbol{\gamma}^m \in \Omega^0$. These vectors are the generators of the polar cone. To see this, let W be the cone generated by $\boldsymbol{\gamma}^i$, i.e., each $\boldsymbol{\omega} \in W$ can be written as a nonnegative linear combination of the $\boldsymbol{\gamma}^i$,

$$W = \{\boldsymbol{\omega} : \boldsymbol{\omega} = \sum_{i=1}^{m} a_i \boldsymbol{\gamma}^i, a_i \geq 0\},$$

then for any $\boldsymbol{\theta} \in \Omega$, we have

$$\langle \boldsymbol{\theta}, \boldsymbol{\omega} \rangle = \sum_{i=1}^{m} a_i \langle \boldsymbol{\theta}, \boldsymbol{\gamma}^i \rangle \leq 0, \quad \forall \boldsymbol{\omega} \in W.$$

This shows that $\Omega \subseteq W^0$, where W^0 is the polar cone of W. For any $\boldsymbol{\zeta} \in W^0$, we have

$$\langle \boldsymbol{\zeta}, \boldsymbol{\gamma}^i \rangle \leq 0, i = 1, \cdots, m,$$

which shows that $W^0 \subseteq \Omega$. Therefore, $\Omega = W^0$. Since $W^{00} = W$ (Rockafellar, 1970, p.121 Theorem 14.1), we have $\Omega^0 = W^{00} = W$.

The edges of a constraint cone are unique up to constant multiplier. There is only one set of edges for a given constraint cone, if we normalize each edge; i.e., if we fix the length of each edge vector to be 1. The edges of a constraint set are not unique. There might be several different sets of edges that form the same closed convex cone. For example, if we set the linear space $V = \mathcal{L}(\mathbf{1}, \boldsymbol{x})$, then both the rows of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ define the same constraint set for a convex regression with $n = 6$ equally spaced x:

$$\boldsymbol{A}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 & 0 \\ 3 & 2 & 1 & 0 & 0 & 0 \\ 4 & 3 & 2 & 1 & 0 & 0 \end{pmatrix}$$

$$\boldsymbol{A}_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 1 & 2 & 3 & 4 \end{pmatrix}.$$

After constraining each edge vector to be perpendicular to the linear space V and have unit length, we get a unique set of edges for the corresponding constraint cone, which are the rows of $\boldsymbol{A}_3$,

$$\boldsymbol{A}_3 = \begin{pmatrix} 0.476 & -0.381 & -0.238 & -0.095 & 0.048 & 0.190 \\ 0.571 & -0.057 & -0.686 & -0.314 & 0.057 & 0.429 \\ 0.429 & 0.057 & -0.314 & -0.686 & -0.057 & 0.571 \\ 0.190 & 0.048 & -0.095 & -0.238 & -0.381 & 0.476 \end{pmatrix}.$$

Each row of $\boldsymbol{A}_3$ is the residual of the projection of the corresponding row of $\boldsymbol{A}_1$ (or $\boldsymbol{A}_2$) onto V.

Now we define $\boldsymbol{\delta}^j$, $j = 1, \cdots, m$ as the edges of the constraint cone Ω^*. Each of the edges $\boldsymbol{\delta}^j$ is perpendicular to V. Any vector $\boldsymbol{\theta} \in \Omega^*$ can be written as a non-negative linear combination of the edge vectors, i.e., $\boldsymbol{\theta} \in \Omega^*$ if and only if $\boldsymbol{\theta} = \sum_{j=1}^m b_j \boldsymbol{\delta}^j$ for $b_j \geq 0$. The constraint set Ω can now be written as

$$\Omega = \left\{ \boldsymbol{\theta} : \boldsymbol{\theta} = \sum_{j=1}^m b_j \boldsymbol{\delta}^j + \boldsymbol{v}; b_j \geq 0, \boldsymbol{v} \in V \right\}.$$

The $\boldsymbol{\delta}^j$s and $\boldsymbol{\gamma}^i$s are linked by the following relationship (Fraser and Massam, 1989)

$$\langle \boldsymbol{\delta}^j, \boldsymbol{\gamma}^i \rangle = \begin{cases} -1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

In order to get constraint cone edges $\boldsymbol{\delta}^j$'s, let

$$\boldsymbol{A}_* = \begin{bmatrix} \boldsymbol{A} \\ V' \end{bmatrix},$$

where V' is the transpose of V, then $\boldsymbol{A}_*$ is a $n \times n$ square matrix. Let $\boldsymbol{B} = \boldsymbol{A}_*^{-1}$, then the first m columns of $\boldsymbol{B}$ are the constraint cone edges $\boldsymbol{\delta}^j, j = 1, \cdots, m$. Define $(\boldsymbol{A}^0)' = [\boldsymbol{\delta}^1 \boldsymbol{\delta}^2 \cdots \boldsymbol{\delta}^m]$,

then $\boldsymbol{A}^0$ is the constraint matrix of Ω^0 and the constraint cone edges are the rows of $\boldsymbol{A}^0$. For n=6 with the monotone restriction, the matrix $\boldsymbol{A}^0$ is

$$\boldsymbol{A}_1^0 = \begin{pmatrix} -5 & 1 & 1 & 1 & 1 & 1 \\ -4 & -4 & 2 & 2 & 2 & 2 \\ -3 & -3 & -3 & 3 & 3 & 3 \\ -2 & -2 & -2 & -2 & 4 & 4 \\ -1 & -1 & -1 & -1 & -1 & 5 \end{pmatrix}$$

For convex constraint matrix as mentioned before, the edges of the constraint cone are the rows of

$$\boldsymbol{A}_2^0 = \begin{pmatrix} 10 & -8 & -5 & -2 & 1 & 4 \\ 20 & -2 & -24 & -11 & 2 & 15 \\ 15 & 2 & -11 & -24 & -2 & 20 \\ 4 & 1 & -2 & -5 & -8 & 10 \end{pmatrix}$$

Note, matrix $\boldsymbol{A}_2^0$ is the same as $\boldsymbol{A}_3$, except for multiplicative constant. Integers are used as elements of the constraint matrix only for the purpose of simplication.

Faces and Sectors

The faces of the constraint cone are constructed by subsets of the constraint cone edges. Any subset $J \subseteq \{1, \cdots, m\}$ determines a face of the constraint cone, i.e., a face consists of all nonnegative linear combinations of constraint cone edges $\delta^j, j \in J$. A face of the constraint cone can be written as

$$F_J = \left\{ \boldsymbol{\theta} \in \Omega^* : \boldsymbol{\theta} = \sum_{j \in J} b_j \boldsymbol{\delta}^j, b_j \geq 0, j \in J \right\}. \tag{2.3.5}$$

The projection of any $\boldsymbol{y} \in I\!R^n$ onto Ω^* falls onto one of the faces of Ω^*. Note that Ω^* itself is a face for $J = \{1, \cdots, m\}$.

The subspace $V^\perp$ can be partitioned into "sectors", which are constructed by either the subsets of the constraint cone edges and the subsets of the polar cone edges. Any subset $J \subseteq \{1, \cdots, m\}$ determines a sector of $V^\perp$.

Let Ω_J be the set of all $\boldsymbol{y} \in V^{\perp}$ such that

$$\Omega_J = \left\{ \boldsymbol{y} \in V^{\perp} : \boldsymbol{y} = \sum_{j \in J} b_j \boldsymbol{\delta}^j + \sum_{j \notin J} b_j \boldsymbol{\gamma}^j \right\}, \tag{2.3.6}$$

where $b_j \geq 0$ for $j \in J$, $b_j > 0$ for $j \notin J$. The polar cone is a sector defined by $J = \emptyset$ and the constraint cone is a sector with $J = \{1, 2, \cdots, m\}$. Each sector Ω_J is a convex polyhedral cone with m edges $\boldsymbol{\delta}^j, j \in J$ and $\boldsymbol{\gamma}^j, j \notin J$.

Let C_J be the set of all y in $I\!R^n$ such that

$$C_J = \left\{ \boldsymbol{y} \in R^n : \boldsymbol{y} = \sum_{j \in J} b_j \boldsymbol{\delta}^j + \sum_{j \notin J} b_j \boldsymbol{\gamma}^j + \boldsymbol{v} \right\}, \tag{2.3.7}$$

where $b_j \geq 0$ for $j \in J$; $b_j > 0$ for $j \notin J$ and $\boldsymbol{v} \in V$, then the C_J partition $I\!R^n$ and (2.3.7) are unique (Meyer 1999).

Meyer (1999) proved the following two propositions:

Proposition 1 *Given $\boldsymbol{y} \in R^n$ such that $\boldsymbol{y} = \sum_{j \in J} b_j \boldsymbol{\delta}^j + \sum_{j \notin J} b_j \boldsymbol{\gamma}^j + \boldsymbol{v}$, the projection of $\boldsymbol{y}$ onto the constraint set Ω is*

$$\hat{\boldsymbol{\theta}} = \sum_{j \in J} b_j \boldsymbol{\delta}^j + \boldsymbol{v} \tag{2.3.8}$$

and the residual vector $\hat{\boldsymbol{\rho}} = \boldsymbol{y} - \hat{\boldsymbol{\theta}} = \sum_{j \notin J} b_j \boldsymbol{\gamma}^j$ is the projection of $\boldsymbol{y}$ onto the polar cone Ω^0.

Proposition 2 *If $\boldsymbol{y} \in C_J$, then $\hat{\boldsymbol{\theta}}$ is the projection of $\boldsymbol{y}$ onto the linear space spanned by the vectors $\boldsymbol{\delta}^j, j \in J$, plus the projection of $\boldsymbol{y}$ onto V. Similarly, $\hat{\boldsymbol{\rho}}$ is the projection of $\boldsymbol{y}$ onto the linear space spanned by the vectors $\boldsymbol{\gamma}^j, j \notin J$.*

From these two propositions, the key idea for getting $\hat{\boldsymbol{\theta}}$ is to find the sector that contains $\boldsymbol{y}$. Once the sector is determined, $\hat{\boldsymbol{\theta}}$ can be easily obtained through ordinary least-squares regression, using edges of the constraint cone and $\boldsymbol{v} \in V$ as regressors. Also these propositions ensure that all $\boldsymbol{y}$ in a given sector will result in $\hat{\boldsymbol{\theta}}$ on the same face of the constraint cone, and $\hat{\boldsymbol{\rho}}$ on the same face of the polar cone.

2.3.3 Hypotheses testing, the test statistic and its distribution

Theoretically, once the sector is determined, the projection onto the cone can be found through ordinary least-squares regression, using the edges as regressors (Proposition 1). We might conclude that the model degrees of freedom under H_0 is the number of edges of the face on which the projection falls, plus the dimension of V, so the error degrees of freedom could be the sample size minus this number. However, this quantity of edges is a random variable, since a different data set will define a different face with a different set J.

Meyer (2003) tested the following hypotheses,

$$\begin{aligned} H_0 &: f(x) = a + bx \\ H_1 &: f(x) \in \mathcal{F}, \end{aligned}$$

where a and b are some constants, $\mathcal{F}$ is the class of convex functions. Let $\hat{\boldsymbol{y}}_0$ be the projection of $\boldsymbol{y}$ onto $V = \mathcal{L}(\mathbf{1}, \boldsymbol{x})$ and $\hat{\boldsymbol{\theta}}$ be the projection of $\boldsymbol{y}$ onto the convex constraint set according to data points $\boldsymbol{x}$.

Let $SSE_0 = \|\boldsymbol{y} - \hat{\boldsymbol{y}}_0\|^2$ and $SSE_1 = \|\boldsymbol{y} - \hat{\boldsymbol{\theta}}\|^2$. If the model variance σ^2 is known, the likelihood ratio test statistic for the above hypotheses is

$$\chi^2_{01} = \frac{1}{\sigma^2}(SSE_0 - SSE_1).$$

For unknown σ^2, the test statistic is

$$B_{01} = \frac{\chi^2_{01}}{\chi^2_{01} + SSE_1/\sigma^2} = \frac{SSE_0 - SSE_1}{SSE_0}.$$

Note that SSE_0/σ^2 has a $\chi^2(n-2)$ distribution, where n is the sample size.

Under H_0, the conditional distribution of SSE_1/σ^2, given $\boldsymbol{y} \in C_J$, is $\chi^2(d)$, where d is the number of elements in set J (Meyer 2003, Corollary 1).

Since $\hat{\boldsymbol{y}}_0 \in V$, we have $\langle \boldsymbol{y} - \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{y}}_0 \rangle = 0$ (by Proposition 2) and $\langle \boldsymbol{y} - \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}} \rangle = 0$, hence $\langle \boldsymbol{y} - \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{y}}_0 - \hat{\boldsymbol{\theta}} \rangle = 0$, which implies that $\boldsymbol{y} - \hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{y}}_0$ are independent. Furthermore,

$$\|\boldsymbol{y} - \hat{\boldsymbol{y}}_0\|^2 = \|\boldsymbol{y} - \hat{\boldsymbol{\theta}}\|^2 + \|\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{y}}_0\|^2,$$

hence $\chi^2_{01} = \|\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{y}}_0\|^2/\sigma^2$. These show that under H_0, the conditional distribution of χ^2_{01} given $D = d$ is $\chi^2(n-d-2)$, where D is the random variable that counts number of elements in set J. Similarly, under H_0, the conditional distribution of B_{01} given $D = d$ is $Beta\left(\frac{n-d-2}{2}, \frac{d}{2}\right)$.

The following theorem gives the null distribution of χ^2_{01} and B_{01} (Meyer 2003).

Theorem 2 *Under H_0,*

$$Pr(\chi^2_{01} \leq a) = \sum_{d=0}^{n-2} Pr\left\{\chi^2(n-d-2) \leq a\right\} Pr(D = d),$$

$$Pr(B_{01} \leq a) = \sum_{d=0}^{n-2} Pr\left\{Beta\left(\frac{n-d-2}{2}, \frac{d}{2}\right) \leq a\right\} Pr(D = d),$$

where $\chi^2(0) \equiv 0$, $Beta(0, \beta) \equiv 0$, and $Beta(\alpha, 0) \equiv 1$.

The values of $Pr(D = d)$, for $d = 1, \cdots, n-2$, is obtained from the "relative volume" of the sets C_J. When H_0 is true, the probability that $\boldsymbol{y}$ is in C_J, is equivalent to the probability that the value of a standard multivariate normal random vector in $\mathcal{R}^n$ falls into C_J. Meyer (2003) found these mixing probabilities by numerically generating N standard multivariate normal random vectors, and determining the value of D for each vector.

Robertson, Wright and Dykstra (1988, Chapter 2) tested the following hypotheses,

$$\begin{aligned} H_0 &: f(x) = c \\ H_1 &: f(x) \in \mathcal{F}, \end{aligned}$$

where c is a constant and $\mathcal{F}$ is the class of monotone nondecreasing functions. The test statistic and its distribution is the same as the above hypotheses, except that the $P(D = d)$ values are different. Theorem 2 applies for all the hypotheses in the format of $H_0 : \boldsymbol{\theta} \in V$ versus $H_1 : \boldsymbol{\theta} \in \Omega$.

2.3.4 The degree of freedom in shape-restricted regression

The error degrees of freedom for shape-restricted regression was explored by Meyer and Woodroofe (2000). For $\boldsymbol{\theta}$ in the interior of the constraint cone, it is shown that

$$n\sigma^2 - 2\sigma^2 E(D) \leq E(\|\boldsymbol{y} - \hat{\boldsymbol{\theta}}\|^2) \leq n\sigma^2 - \sigma^2 E(D) \tag{2.3.9}$$

where E denotes expectation and it depends on the values of the underlying $\boldsymbol{\theta}$ and σ^2. For the shape-restricted regression, the following equation holds

$$E(\|\boldsymbol{y} - \hat{\boldsymbol{\theta}}\|^2) = n\sigma^2 - c\sigma^2 E(D) + o(n^{1/3}). \tag{2.3.10}$$

It is shown through simulations that for monotone regression c is about 1.5; for convex regression, c is about 1.2.

2.3.5 Algorithms for shape-restricted regression

For shape-restricted regressions, the least-squares estimator is the projection onto a closed convex cone. The pool-adjacent-violators algorithm (Robertson, Wright and Dykstra, 1988, p.8-10) is a convenient algorithm to get the monotone regression estimate. Recall that there is a closed form solution (2.3.2) for monotone regression.

For other shape restriction problems, such as convex and nondecreasing concave regressions, this is a quadratic programming problem and there is no known closed-form solution.

Hildreth (1954) was the first one to address this issue and propose an algorithm that converges in an infinite number of steps. Wilhelmsen (1976) and Pshenichny & Danilin (1978) used different algorithms converging in a finite number of steps. Wu (1982) and Dykstra (1983) offered simpler solutions either convergent in an infinite number of steps or not necessarily convergent.

The mixed primal-dual bases algorithm (Fraser and Massam, 1989) provides an efficient way to find the least square regression estimate under inequality constraints. It converges to the exact solution in a finite number of iterations. But it is restricted to the case that the

number of constraints does not exceed the number of dimension n. Meyer (1999) extended the mixed primal-dual bases algorithm to the case of more constraints than dimensions. The hinge algorithm (Meyer, 1999t) is also an easy and efficient algorithm converging in a finite number of iterations. It appears to be faster than the mixed primal-dual bases algorithm for many applications.

The pool-adjacent-violators algorithm (PAVA)

The PAVA is a convenient algorithm to get monotone regression estimates. After the data is ordered by the x values, the PAVA starts with $\boldsymbol{y}$. If $\boldsymbol{y}$ is monotone then $\boldsymbol{\theta} = \boldsymbol{y}$. Otherwise, there must exist a smallest subscript i such that $y_{i-1} > y_i$. These two values are then replaced by their average $(y_{i-1} + y_i)/2$. If this new set of values is not monotone, then this process is repeated using the new values until a monotone set of values is obtained.

The mixed primal-dual bases algorithm

The mixed primal-dual bases algorithm is used to find the projection onto a closed convex cone. In this algorithm, the $\boldsymbol{\gamma}^j$'s are the "primal" vectors and $\boldsymbol{\delta}^j$'s are the"dual" vectors. The mixed primal-dual bases algorithm finds the correct set $\hat{J}$ by moving along a line segment connecting the point $\boldsymbol{z}^0 = \sum_{j=1}^{m} \boldsymbol{\delta}^j$ with $\boldsymbol{z}$, where $\boldsymbol{z}$ is the projection of the data $\boldsymbol{y}$ on the subspace spanned by $\boldsymbol{\delta}^j, j = 1, \cdots, m$. At the k^{th} iteration, the point $\boldsymbol{z}^k$ on the line segment is reached, such that the distance between $\boldsymbol{z}^k$ and $\boldsymbol{z}$ is strictly decreasing in k. This point is also on a face of Ω_{J_k}. The next iteration finds $\boldsymbol{z}^{k+1}$, farther along the segment, on a face of $\Omega_{J_{k+1}}$. At the beginning of the iteration, both $\boldsymbol{z}$ and $\boldsymbol{z}^k$ are expressed in the basis defined by J_k, such as

$$\boldsymbol{z} = \sum_{j \in J_k} b_j \boldsymbol{\delta}^j + \sum_{j \notin J_k} b_j \boldsymbol{\gamma}^j$$

and

$$\boldsymbol{z}^k = \sum_{j \in J_k} a_j \boldsymbol{\delta}^j + \sum_{j \notin J_k} a_j \boldsymbol{\gamma}^j$$

where $a_j \geq 0$ for $j \in J_k$ and $a_j > 0$ for $j \notin J_k$. If $b_j \geq 0$ for $j \in J_k$ and $b_j > 0$ for $j \notin J_k$, the algorithm stops. Otherwise, find

$$\boldsymbol{z}^{k+1} = \boldsymbol{z}^k + \alpha_{k+1}(\boldsymbol{z} - \boldsymbol{z}^k)$$

where $\alpha_{k+1} \in (0, 1)$ is as large as possible while the coefficients of $\boldsymbol{z}^{k+1}$ are all positive or nonnegative as they are in J_k or not, respectively. The point $\boldsymbol{z}^{k+1}$ is on the face of Ω_{J_k}, which divides Ω_{J_k} and $\Omega_{J_{k+1}}$. The algorithm terminates at the face of the sector containing $\boldsymbol{z}$. Clearly, it takes a finite number of iterations since there are a finite number of sectors.

The hinge algorithm

The hinge algorithm is another convenient way to get the least-squares fit for a shape restriction regression. This algorithm uses a set of vectors $\boldsymbol{\delta}^1, \cdots, \boldsymbol{\delta}^n$ to characterize the constraint space. One special choice of the vectors $\boldsymbol{\delta}^j$ for a convex regression could be as follow:

$$\delta_i^j = x_{j+1} - x_i, \quad \delta_i^{n-1} = x_i, \quad \delta_i^n = 1,$$

where $i = 1, \cdots, n$, $j = 1, \cdots, n-2$. The algorithm finds $\hat{\boldsymbol{\theta}}$ by finding $\hat{J}$ through a series of guesses J_k. At a typical iteration, the current estimate $\boldsymbol{\theta}^k$ can be obtained by the least-squares regression of $\boldsymbol{y}$ on the $\boldsymbol{\delta}^j$, for $j \in J_k$. We call $\boldsymbol{\delta}^j$ the "hinges" since for the convex regression problem, the points (x_j, θ_j), $j \in J$, are the bending points at which the line segments change slope, and there is only one way that the bends are allowed to go.

The initial guess J_0 is set to be $\{n-1, n\}$. Since there are no constraints on the coefficients of the $\boldsymbol{\delta}^{n-1}, \boldsymbol{\delta}^n$, $J_0 \subseteq J_k$ for all k.

The algorithm can be summarized in four steps:

(1) Using $\boldsymbol{\delta}^j, j \in J_0$ as regressors, obtain the least-squares estimate $\boldsymbol{\theta}^0$.

(2) At the k^{th} iteration, compute $\langle \boldsymbol{y} - \boldsymbol{\theta}^k, \boldsymbol{\delta}^j \rangle$ for each $j \notin J_k$. If these are all non-positive, then stop. If not, then add the vector $\boldsymbol{\delta}^j$ to the model for which this inner product is largest.

(3) Get the least-squares fit with the new set of $\boldsymbol{\delta}$-vectors.

(4) Check to see if the regression function satisfies the constraints on the coefficients, which are $b_j \geq 0$, for $j \in J$ and $j \notin J_0$

   (a) If yes, go to step (2).

   (b) If no, choose the hinge with the largest negative coefficient and remove it from the current set J. Go to step (3).

Intuitively, at each stage, the new hinge is added where it is "most needed", and other hinges are removed if the new fit does not satisfy the constraints. It is clear that if the algorithm ends, it gives the correct solution. The algorithm does end, see Meyer (1999) for proof.

This algorithm is fast and does not depend on an initial guess. This is important for bootstrapping ideas and iterative projections algorithm, since the projection is repeated in a loop.

2.3.6 Other numerical methods used

In order to fit unbalanced semiparametric ANCOVA model, we need to combine cone projections with other numerical methods to get the final estimates of the coefficients $\boldsymbol{\beta}$. When there are two treatment levels, i.e., $k = 2$, golden section search method is used; for more than two treatment groups, i.e., $k > 2$, downhill simplex method is applied.

Golden section search method

The golden section search algorithm provides a method of function minimization.[1] A minimum is known to be bracketed only when there is a triplet of points, $a < b < c$, such that $f(b)$ is less than both $f(a)$ and $f(c)$. We choose a new point x, either between a and b or between b and c; suppose we make the latter choice, then we evaluate $f(x)$. If $f(b) < f(x)$,

[1] Numerical Recipes in Fortran 90, Chapter 10

then the new bracketing triplet of points is (a, b, x); otherwise, the new bracketing triplet is (b, x, c).

It remains to decide on a strategy for choosing the new point x, given (a, b, c). Suppose that b is a fraction w of the distance between a and c, i.e.,

$$\frac{b-a}{c-a} = w \quad and \quad \frac{c-b}{c-a} = 1-w.$$

Also suppose that x is an additional fraction z beyond b, such as

$$\frac{x-b}{c-a} = z.$$

Then the next bracketing segment will either be of length $w + z$ relative to the current one, or else of length $1 - w$. If we want to minimize the worse case possibility, then we will choose z to make these equal, namely

$$w + z = 1 - w. \tag{2.3.11}$$

The scale similarity implies that x should be the same fraction of the distance from b to c as was b from a to c, in other words,

$$w = \frac{z}{1-w}. \tag{2.3.12}$$

From equations (2.3.11) and (2.3.12), we get $w = \frac{3-\sqrt{5}}{2} \approx 0.38197$. In other words, the optimal bracketing interval (a, b, c) has its middle point b a fractional distance 0.38197 from one end, and 0.61803 from the other end. These fractions are the so-called *golden section*. This method of function minimization is thus called the *golden section search*. We summarize the golden section search method as follows:

At each stage, given a bracketing triplet of points, the next point to be tried is a fraction 0.38197 into the larger of the two intervals (measuring from the central point of the triplet). If we start out with a bracketing triplet whose segments are not in the golden ratios, the procedure of choosing successive points at the golden mean point of the larger segment will quickly converge to the proper, self-replicating ratios.

The golden section search guarantees that each new function evaluation will (after self-replicating ratios have been achieved) bracket the minimum to an interval just 0.61803 times the size of the preceding interval.

Downhill simplex method in multidimensions

The downhill simplex method is due to Nelder and Mead.[2] It is used to find the minimum of a function of more than one independent variable. The method requires only function evaluations, not derivatives. This method has a geometrical appeal which makes it delightful to describe or work through:

A "simplex" is the geometrical figure consisting, in N dimensions, of $N + 1$ points and all their interconnecting line segments, polygonal faces, etc. In two dimensions, a simplex is a triangle. In three dimensions it is a tetrahedron, not necessarily the regular tetrahedron. In general we are only interested in simplexes that are nondegenerate, i.e., that enclose a finite inner N-dimensional volume. If any point of a nondegenerate simplex is taken as the origin, then the N other points define vector directions that span the N-dimensional vector space. The downhill simplex method must be started not just with a single point, but with $N+1$ points, defining an initial simplex. If we think of one of these points as being our initial starting point $\boldsymbol{P}_0$, then we can take the other N points to be

$$\boldsymbol{P}_i = \boldsymbol{P}_0 + \lambda \boldsymbol{e}_i,$$

where $\boldsymbol{e}_i$'s are N unit vectors, and λ is a constant which is our guess of the problem's characteristic length scale.

The downhill simplex method now takes a series of steps, most steps just moving the point of the simplex where the function is largest ("highest point") through the opposite face of the simplex to a lower point. These steps are called reflections, and they are constructed to conserve the volume of the simplex. When it can do so, the method expands the simplex in one or another direction to take larger steps. When it reaches a "valley floor", the method

[2]Numerical Recipes in Fortran 90, Chapter 10

contracts itself in the transverse direction and tries to ooze down the valley. If there is a situation where the simplex is trying to "pass through the eye of a needle", it contracts itself in all directions, pulling itself in around its lowest (best) point.

Chapter 3

Semiparametric ancova using shape restrictions

In an ANCOVA model, suppose there is only qualitative information about the relationship between the response and the covariate, such that the response is nondecreasing with the covariate, or convex, or increasing and concave. Then a semiparametric ANCOVA model will be a reasonable choice to fit the data.

Previous studies (Speckman 1988; Heckman 1986; Wahba 1984; *etc.*) apply some smoothing techniques such as kernel or spline smoothing to characterize the property of the underlying function. However, smoothing techniques are generally used when the relationship between the response and the covariate is hard to formulate. In our case, when the underlying function has a special shape, using only a smoothing technique might waste some useful information, and result in a loss of power for the hypotheses tests. Also, by applying a smoothing technique, a suitable smoothing parameter must be chosen, which might involve intensive numerical computations. And the test results might be sensitive to the choice of the smoothing parameter.

In this chapter, we apply shape restrictions to the semiparametric ANCOVA model. They can provide flexible fits to the data, which are comparable to smoothing techniques. If the assumption of the shape restriction holds, the shape-restricted ANCOVA might have higher power than that of the parametric models when performing hypothesis tests.

We study two kinds of shape restrictions, monotone and convex, and develop the beta test statistics and their distributions. For monotone ANCOVA, we compare the power of the beta test to both the power of an approximate F-test, denoted by F^s-test, from kernel smoothing (see section 2.2.1) and the power of the traditional F-test from a parametric

ANCOVA model with a linear underlying function; for convex ANCOVA, the power of the beta test is compared to both the power of the F^s-test and the power of the F-test from a parametric ANCOVA model with a quadratic underlying function.

For the sake of simplicity, we consider a balanced design. The balanced design is again defined as: at any fixed value of x, there are equal numbers ($= a$) of y observed from each treatment group.

A simple shape-restricted ANCOVA model can be written as

$$y_i = f(x_i) + \beta_1 d_{1i} + \cdots + \beta_{k-1} d_{(k-1)i} + \sigma\varepsilon_i$$

where $i = 1, \cdots, n$, k is the number of treatment levels, ε_i are assumed to be independent, identically distributed standard normal errors, and f is any function belongs to the shape-restricted family. The σ and β_j, $j = 1, \cdots, k-1$ are some unknown constants. The d_{ji}'s, $i = 1, \cdots, n$, $j = 1, \cdots, k-1$, are indicator variables, where

$$d_{ji} = \begin{cases} 1 & \text{if } i^{th} \text{ obs is in group } j \\ 0 & \text{otherwise.} \end{cases}$$

We assume no interaction term in the model, then the fit is k parallel curves.

Our main interest is to test

$$H_0\text{: } \beta_1 = \cdots = \beta_{k-1} = 0$$

$$H_1\text{: at least one } \beta_1, \cdots, \beta_{k-1} \text{ is not zero.}$$

3.1 The test statistic and its distribution

The test statistic is formed by comparing SSE under the null hypothesis and the alternative hypothesis. Both involve projection onto the same constraint cone. The following theorem shows that for a balanced design, the difference in error degrees of freedom under the hypotheses is simply $k-1$, where k is the number of treatment levels.

Theorem 3 *For a balanced shape-restricted semiparametric ANCOVA model with k treatment levels, define $\chi^2_{01} = \frac{SSE_0 - SSE_1}{\sigma^2}$, where σ is a known constant, SSE_0 and SSE_1 are the*

residual sum of squares under the null hypothesis and the alternative hypothesis, respectively. Then χ^2_{01} has a chi-square distribution with degrees of freedom $k-1$.

Proof: Condition on a J^*, and let s be the number of elements in J^*. We can write a "design matrix" as $(\mathbf{1}, \boldsymbol{x}, \boldsymbol{d}_1, \cdots, \boldsymbol{d}_{k-1}, \boldsymbol{\delta}^j, j \in J^*)$, where $\boldsymbol{\delta}^j, j \in J^*$ are some edges of the constraint cone, which are orthogonal to $\mathcal{L}(\mathbf{1}, \boldsymbol{x}, \boldsymbol{d}_1, \cdots, \boldsymbol{d}_{k-1})$.

For example, the "design matrix" for a convex constraint with $k = 2$, group size $m = 6$, equally spaced x and $J^* = \{1, 3\}$ can be written as

$$\begin{pmatrix} 1 & x_1 & 0 & 10 & 15 \\ 1 & x_2 & 0 & -8 & 2 \\ 1 & x_3 & 0 & -5 & -11 \\ 1 & x_4 & 0 & -2 & -24 \\ 1 & x_5 & 0 & 1 & -2 \\ 1 & x_6 & 0 & 4 & 20 \\ 1 & x_1 & 1 & 10 & 15 \\ 1 & x_2 & 1 & -8 & 2 \\ 1 & x_3 & 1 & -5 & -11 \\ 1 & x_4 & 1 & -2 & -24 \\ 1 & x_5 & 1 & 1 & -2 \\ 1 & x_6 & 1 & 4 & 20 \end{pmatrix}$$

Write $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n$ as an orthonormal basis for $\mathbb{R}^n$, where $\boldsymbol{u}_1$ spans $\mathcal{L}(\mathbf{1})$, $\boldsymbol{u}_2$ spans $\mathcal{L}(\boldsymbol{x})$, $\boldsymbol{u}_3$ spans $\mathcal{L}(\boldsymbol{d}_1)$, $\cdots$, $\boldsymbol{u}_{k+1}$ spans $\mathcal{L}(\boldsymbol{d}_{k-1})$, and $\boldsymbol{u}_{k+2}, \cdots, \boldsymbol{u}_{k+2+s}$ span $\mathcal{L}(\boldsymbol{\delta}^j, j \in J^*)$, $\mathcal{L}(\boldsymbol{a}, \boldsymbol{b})$ refers to the linear space spanned by vector $\boldsymbol{a}$ and $\boldsymbol{b}$. Then the projection of $\boldsymbol{y}$ onto the design matrix can be written as

$$\hat{\boldsymbol{y}} = \sum_{j=1}^{k+1} a_j \boldsymbol{u}_j + \sum_{j=k+2}^{k+2+s} a_j \boldsymbol{u}_j$$

where $\sum_{j=1}^{k+1} a_j \boldsymbol{u}_j$, $\sum_{j=k+2}^{k+2+s} a_j \boldsymbol{u}_j$ and $\sum_{k+3+s}^{n} a_j \boldsymbol{u}_j$ are the projections of $\boldsymbol{y}$ onto three orthogonal subspaces of $\mathbb{R}^n$. Since

$$a_j \boldsymbol{u}_j = \Pi(\boldsymbol{y}|\boldsymbol{u}_j) = \frac{\langle \boldsymbol{y}, \boldsymbol{u}_j \rangle}{\|\boldsymbol{u}_j\|^2} \boldsymbol{u}_j,$$

so that

$$a_j = \langle \boldsymbol{y}, \boldsymbol{u}_j \rangle,$$

hence,

$$a_j \sim N\left(\langle \boldsymbol{\theta}, \boldsymbol{u}_j \rangle, \sigma^2\right),$$

where $\boldsymbol{\theta} = E[\boldsymbol{y}]$, and E refers to expected value. Therefore,

$$\frac{a_j^2}{\sigma^2} \sim \chi^2_{\tau_j}(1),$$

where τ_j is the noncentality parameter with $\tau_j = \frac{\|\langle \boldsymbol{\theta}, \boldsymbol{u}_j \rangle\|^2}{\sigma^2}$

Under H_0, $\tau_j = 0$ for $j = 3, \cdots, k+1$, hence

$$\frac{a_j^2}{\sigma^2} \sim \chi^2_0(1) \quad for \quad j = 3, \cdots, k+1.$$

Let $\hat{\boldsymbol{y}}_0$ and $\hat{\boldsymbol{y}}_1$ be the fitted values of $\boldsymbol{y}$ under H_0 and H_1, respectively. Then

$$\hat{\boldsymbol{y}}_1 - \hat{\boldsymbol{y}}_0 = a_3 \boldsymbol{u}_3 + \cdots + a_{k+1} \boldsymbol{u}_{k+1}.$$

Further,

$$\|\boldsymbol{y} - \hat{\boldsymbol{y}}_0\|^2 = \|\boldsymbol{y} - \hat{\boldsymbol{y}}_1\|^2 + \|\hat{\boldsymbol{y}}_1 - \hat{\boldsymbol{y}}_0\|^2,$$

so,

$$\begin{aligned} \|\hat{\boldsymbol{y}}_1 - \hat{\boldsymbol{y}}_0\|^2 &= \|\boldsymbol{y} - \hat{\boldsymbol{y}}_0\|^2 - \|\boldsymbol{y} - \hat{\boldsymbol{y}}_1\|^2 \\ &= SSE_0 - SSE_1, \end{aligned}$$

where SSE_0 and SSE_1 are the residual sum of squares under the null hypothesis and the alternative hypothesis, respectively. Hence, we have

$$\frac{SSE_0 - SSE_1}{\sigma^2} = \frac{a_3^2 + \cdots + a_{k+1}^2}{\sigma^2}.$$

Since

$$\frac{a_3^2 + \cdots + a_{k+1}^2}{\sigma^2} \sim \chi^2(k-1),$$

therefore,

$$\frac{SSE_0 - SSE_1}{\sigma^2} \sim \chi^2(k-1).$$

This is true for any realization of J^*, i.e., $\left(\frac{a_3^2+\cdots+a_{k+1}^2}{\sigma^2}|J^*\right) \sim \chi^2(k-1)$ for any J^*. Hence,

$$\frac{SSE_0 - SSE_1}{\sigma^2} \sim \chi^2(k-1). \qquad \diamond$$

For the case of σ unknown, we define the test statistic as

$$B_{01} = \frac{\chi_{01}^2}{\chi_{01}^2 + SSE_1/\sigma^2} = \frac{SSE_0 - SSE_1}{SSE_0}.$$

If SSE_1/σ^2 has a $\chi^2(r)$ density, then the statistic B_{01} would be distributed as $Beta(\frac{k-1}{2}, \frac{r}{2})$, because χ_{01}^2 and SSE_1 are independent. However, from the simulation studies, the statistic SSE_1/σ^2 appears to behave similarly as a density which is a mixture of chi-squares, but the density and mixing distribution can not be determined exactly since the true $\boldsymbol{\theta}$ is not in the linear space V, but rather in the interior of the constraint cone.

The theory of degrees of freedom in shape-restricted regression (Meyer and Woodroofe, 2000) suggests that $n - cD$ might be used for r in the beta test statistic above, where n is the total sample size, c is a constant, where $c \approx 1.5$ for monotone restrictions, and $c \approx 1.2$ for convex restrictions, D is the dimension of the face of the constraint cone on which the projection falls.

Figure 3.1 contains the probability plots of B_{01} under monotone restriction. A total of 100000 data sets were generated to construct these plots. For each data set, the test statistic B_{01} was computed and its p-value was obtained from $Beta(\frac{1}{2}, \frac{n-1.5D}{2})$ distribution. Plots (a) and (b) show the case of an underlying jump function

$$f(x_i) = \begin{cases} 0 & \text{if} \quad 0 < x_i < 0.5 \\ 2 & \text{if} \quad 0.5 \leq x_i < 1, \end{cases}$$

with sample size 20 and 80, respectively. Plots (c) and (d) show the case of an underlying log function $f(x_i) = 2log(x_i)$, with sample size 20 and 80, respectively. The x-axis is the

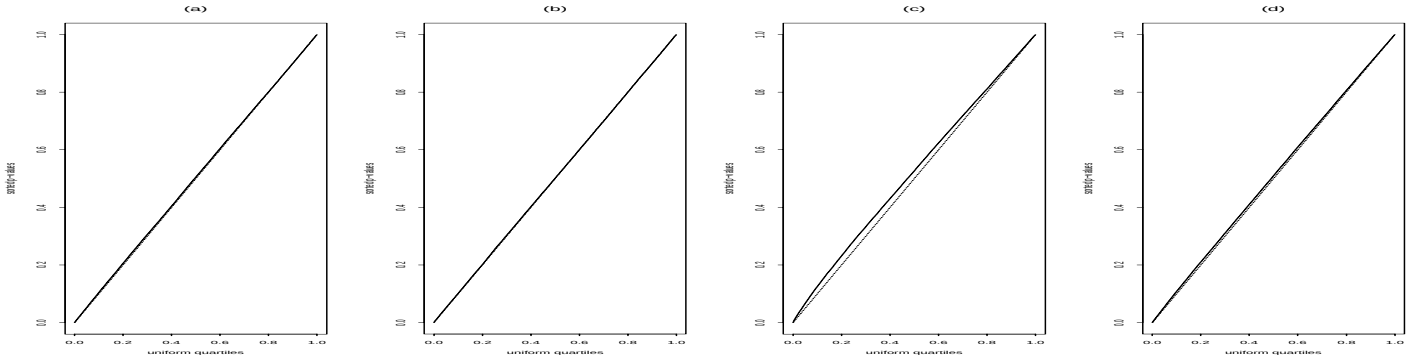

Figure 3.1: Probability plot of test statistic B_{01} under monotone restrictions. (a) sample size 20, jump underlying function; (b) sample size 80, jump underlying function; (c) sample size 20, log underlying function; (d) sample size 80, log underlying function. The x-axis is the uniform quantiles and y-axis is the sorted p-values. The solid line is the probability plot and the dotted line is $y = x$.

uniform quantiles and y-axis is the sorted p-values. The solid line is the probability plot and the dotted line is $y = x$. The probability plot and the line $y = x$ highly overlap, which suggests that the chosen distribution of B_{01} appears to be close to correct.

Figure 3.2 shows the probability plot for B_{01} under convex restriction. Total of 100000 data sets were generated. For each data set, the test statistic B_{01} was computed and its p-value was obtained from $Beta(\frac{1}{2}, \frac{n-1.2D}{2})$ distribution. Plots (a) and (b) show the case of an convex underlying function $f(x_i) = \frac{1}{4}(2x_i + \frac{1}{x_i})$, with sample size of 20 and 60, respectively. Plots (c) and (d) show the case of an ramp underlying function

$$f(x_i) = \begin{cases} 0 & \text{if} \quad 0 < x_i < 0.5 \\ x_i - 0.5 & \text{if} \quad 0.5 \leq x_i < 1, \end{cases}$$

with sample size of 20 and 60, respectively. The x-axis is the uniform quantiles and y-axis is the sorted p-values. The solid line is the probability plot and the dotted line is $y = x$. Again the chosen distribution of B_{01} seems correct.

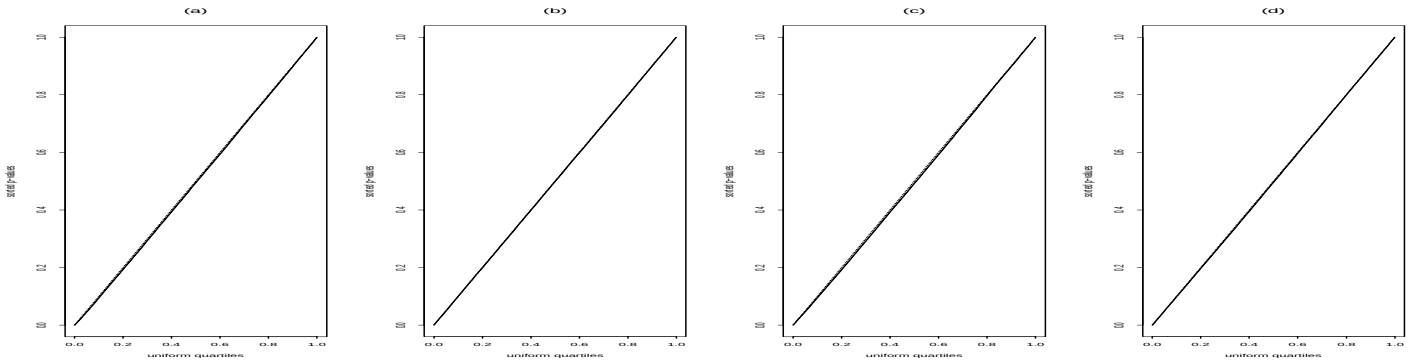

Figure 3.2: Probability plot of test statistic B_{01} under convex restrictions. (a) sample size 20, convex underlying function; (b) sample size 60, convex underlying function; (c) sample size 20, ramp underlying function; (d) sample size 60, ramp underlying function. The x-axis is the uniform quantiles and y-axis is the sorted p-values. The solid line is the probability plot and the dotted line is $y = x$.

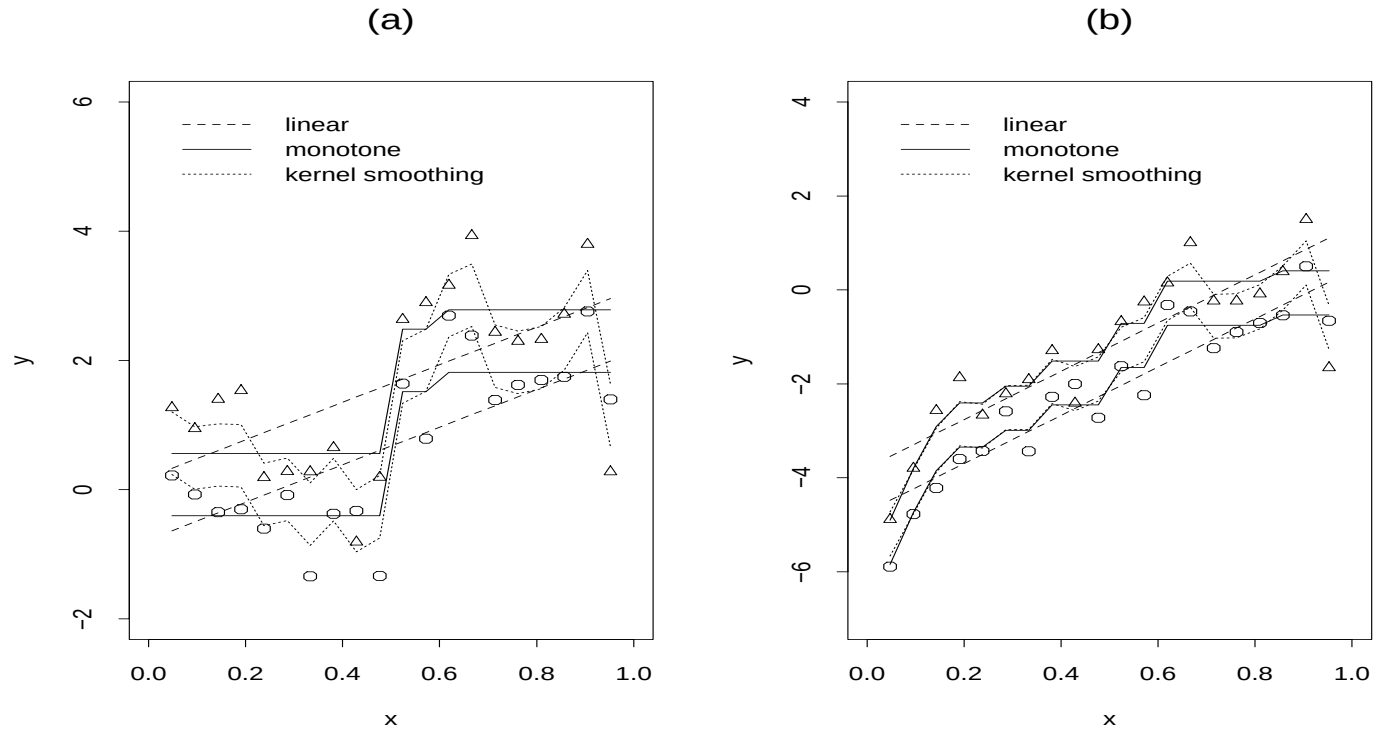

Figure 3.3: Fit monotone, kernel smoothing and linear regressions when the underlying function is jump (a) and log (b). Sample size is 40, model variance is chosen to keep F-test power=0.5.

3.2 Semiparametric ancova using monotone restriction

We choose three different underlying functions to study the test size and power of the beta test under monotone restrictions. They are the linear function $f(x_i) = x_i$, the log function $f(x_i) = 2log(x_i)$, and the jump function

$$f(x_i) = \begin{cases} 0 & \text{if} \quad 0 < x_i < 0.5 \\ 2 & \text{if} \quad 0.5 \leq x_i < 1 \end{cases}$$

In order to make comparisons, the test size and power from simple linear regression and kernel smoothing regression are also computed, where the smoothing parameter h in kernel smoothing is chosen by the method of ordinary cross-validation (see section 2.2.2). The kernel smoothing matrix we used through our study is

$$\boldsymbol{K}_{h(ij)} = \frac{s(\frac{x_j - x_i}{h})}{\sum_{i=1}^{n} s(\frac{x_j - x_i}{h})}, \tag{3.2.1}$$

where h is a chosen bandwidth, $\boldsymbol{K}_{h(ij)}$ is the element at the i^{th} row and j^{th} column of the kernel smoothing matrix $\boldsymbol{K}_h$, $s(x)$ is the standard normal density, $s(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

Figure 3.3 shows the fits of monotone regression, simple linear regression and kernel smoothing regression on two given data sets. The data are generated from the jump function (a) and the log function (b), and the sample size is 40. We choose the model variance in order to keep the F-test at power= 0.50. The solid curve is the monotone fit, the dotted line is the simple linear fit and the dashed line is the kernel smoothing fit.

3.2.1 Test size analysis

Test size is defined as the probability of rejecting H_0 when H_0 is true. Table 3.1 lists the test size of the beta test (under monotone constraints), the F^s-test and the F-test, for different underlying functions and different sample sizes. One million data sets were generated for each underlying function and sample size combination.

For the linear underlying function, F-test has the test size 0.050 and the test sizes for both the beta test and the F^s-test are very close to 0.05, no matter what the sample size n

is. For the jump underlying function, F-test has deflated test size 0.028 for all n; however, the test sizes for the beta test and the F^*-test are much better and they get closed to 0.05 as n increases. For the log underlying function there is a strange trend in test size for F-test; it differs from 0.05 by substantial amount, which increases with n. The test sizes for the beta test and F^s-test increase toward 0.05 as n increases.

Table 3.1: Test sizes for the beta test (monotone restriction), the F^s-test and the F-test when underlying functions are log, jump, and linear respectively.

log regression function

Tests	$n = 20$	$n = 40$	$n = 80$
Beta test	0.036	0.041	0.044
F^s-test	0.043	0.047	0.046
F-test	0.033	0.023	0.016

jump regression function

Tests	$n = 20$	$n = 40$	$n = 80$
Beta test	0.040	0.043	0.046
F^s-test	0.043	0.049	0.044
F-test	0.028	0.028	0.028

linear regression function

Tests	$n = 20$	$n = 40$	$n = 80$
Beta test	0.047	0.048	0.048
F^s-test	0.051	0.050	0.050
F-test	0.050	0.050	0.050

3.2.2 Power comparison

Power is defined as the probability of rejecting H_0 when H_0 is false. The powers of the tests are computed numerically for several choices of sample size n, model variance and true underlying regression function. The power of the beta test is compared to that of the F^s-test and the F-test (Table 3.2).

The sample sizes n are 20, 40, and 80 respectively. The model variances are chosen so that the power of the corresponding F-test is 0.250, 0.500 and 0.750, respectively. One million

datasets were generated for each sample size, model variance and underlying regression function combinations. We assume the true parameter value $\beta = 0.6$.

When the true underlying function is linear, the F-test has the highest power, but as n increases, all three tests have similar power. However, when the true underlying function is jump, the beta test and the F^s-test have more power, especially for small sample sizes. This result is reasonable since a monotone regression line will best explain the underlying jump function. Similar results were obtained from the log underlying function.

Table 3.2: Power of the beta test (monotone restriction), compared with the F^s-test and F-test.

log regression function

F-test	Beta test			F^s-test		
	$n = 20$	$n = 40$	$n = 80$	$n = 20$	$n = 40$	$n = 80$
0.250	0.263	0.295	0.285	0.311	0.312	0.278
0.500	0.623	0.666	0.618	0.757	0.697	0.605
0.750	0.931	0.943	0.898	0.989	0.959	1.000

jump regression function

F-test	Beta test			F^s-test		
	$n = 20$	$n = 40$	$n = 80$	$n = 20$	$n = 40$	$n = 80$
0.250	0.367	0.278	0.260	0.374	0.288	0.263
0.500	0.921	0.622	0.547	0.916	0.623	0.540
0.750	1.000	0.911	0.822	1.000	0.907	0.812

linear regression function

F-test	Beta test			F^s-test		
	$n = 20$	$n = 40$	$n = 80$	$n = 20$	$n = 40$	$n = 80$
0.250	0.235	0.242	0.246	0.246	0.241	0.248
0.500	0.471	0.489	0.494	0.482	0.483	0.496
0.750	0.714	0.738	0.745	0.726	0.731	0.747

3.3 Semiparametric ancova using convex restriction

We chose three underlying functions to study the test size and power of the beta test under convex restrictions. They are the convex function $f(x_i) = 2x_i + 1/x_i$, the quadratic function

$f(x_i) = x_i^2$, and the ramp function

$$f(x_i) = \begin{cases} 0 & \text{if} \quad 0 < x_i < 0.5 \\ x_i - 0.5 & \text{if} \quad 0.5 \leq x_i < 1. \end{cases}$$

Again, the test size and power from a quadratic regression and a kernel smoothing regression were computed as comparisons.

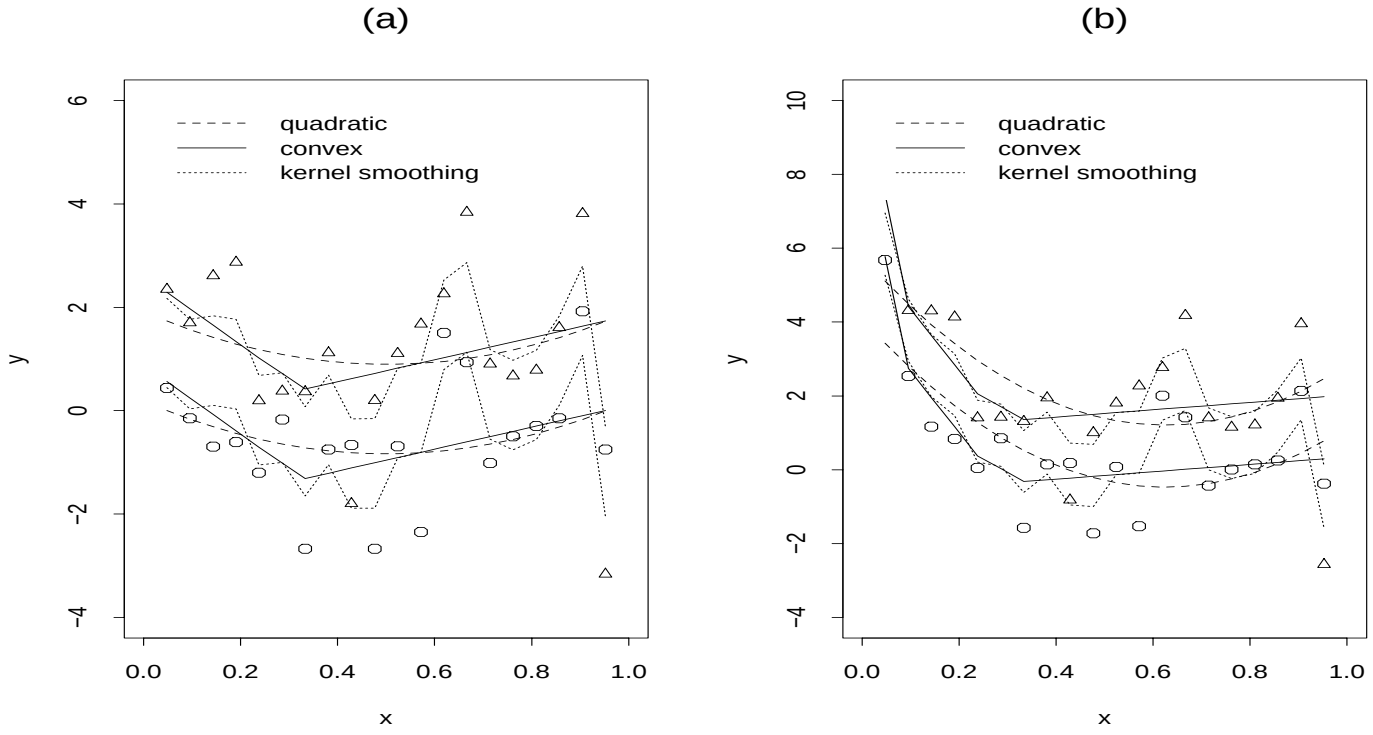

Figure 3.4: Fitted convex, kernel smoothing and quadratic regressions when the underlying function is ramp (a) and convex (b). Sample size is 40. Model variance was chosen to keep F-test power=0.5.

Figure 3.4 shows the fits of convex regression, kernel smoothing regression and quadratic regression on two given data sets. The data are generated from the ramp function (a) and the convex function (b), and the sample size is 40. We chose the model variance to keep the F test at power= 0.50. The solid curve is the convex fit, the dotted line is the quadratic fit and the dashed line is the kernel smoothing fit.

Table 3.3: Test sizes for the beta test (convex restriction), F^s-test, F-test when underlying functions are convex, ramp, and quadratic, respectively.

convex regression function

Tests	$n = 20$	$n = 40$	$n = 60$
Beta test	0.053	0.051	0.050
F^s-test	0.044	0.047	0.046
F-test	0.044	0.025	0.010

ramp regression function

Tests	$n = 20$	$n = 40$	$n = 60$
Beta test	0.053	0.050	0.050
F^s test	0.050	0.051	0.051
F-test	0.050	0.050	0.050

quadratic regression function

Tests	$n = 20$	$n = 40$	$n = 60$
Beta test	0.053	0.051	0.050
F^s-test	0.048	0.050	0.051
F-test	0.050	0.050	0.050

3.3.1 Test size analysis

Table 3.3 lists the test size of the beta test (under convex constraint), F^s-test and F-test, for different underlying functions and sample sizes. One million data sets were generated for each underlying function and sample size combination. For ramp underlying function and quadratic underlying function, all three tests have test size very close to 0.050 for different sample size n. For convex underlying function, the test size of F-test is bad deflated and decreases as n increases. The Beta test has the test size closest to 0.05.

Table 3.4: Power of the beta test (convex restriction), compared with the F^s-test and the F-test.

convex regression function

F-test	Beta test			F^s-test		
	$n = 20$	$n = 40$	$n = 60$	$n = 20$	$n = 40$	$n = 60$
0.250	0.267	0.273	0.282	0.241	0.258	0.263
0.500	0.533	0.562	0.591	0.488	0.538	0.551
0.750	0.788	0.831	0.865	0.744	0.809	0.848

ramp regression function

F-test	Beta test			F^s-test		
	$n = 20$	$n = 40$	$n = 60$	$n = 20$	$n = 40$	$n = 60$
0.250	0.258	0.252	0.251	0.251	0.241	0.248
0.500	0.509	0.505	0.502	0.499	0.484	0.496
0.750	0.757	0.753	0.752	0.750	0.731	0.746

quadratic regression function

F-test	Beta test			F^s-test		
	$n = 20$	$n = 40$	$n = 60$	$n = 20$	$n = 40$	$n = 60$
0.250	0.257	0.253	0.252	0.250	0.241	0.248
0.500	0.510	0.504	0.502	0.496	0.484	0.496
0.750	0.758	0.753	0.752	0.740	0.731	0.747

3.3.2 Power comparison

The power of the beta test is computed numerically for several choices of sample size n, model variance and true underlying regression function. And the power of the beta test is also compared to that of the F^s-test and the F-test (Table 3.4).

The sample sizes n are 20, 40, and 60 respectively. The model variances are chosen so that the power of the corresponding F-test is 0.250, 0.500 and 0.750 respectively. One million datasets were generated for each sample size, model variance and underlying regression function combination. We assume the true β is 1.0.

For ramp, quadratic and convex underlying functions, the beta test always has higher power than the corresponding F tests. A possible reason is that convex regression might explain more model variability than a parabola. The beta test has similar power when the underlying functions are ramp and quadratic. Usually, we expect higher power from the F-test when the data are generated from a quadratic underlying function. However, our simulation results show that the beta test has better power than the F-test in this case, although the difference in power is trivial. For convex underlying function, the superiority of the beta test enhances as n increases and the F^s-test has slightly lower power comparing to that of the beta test.

From the above test size and power analysis, we conclude that the beta test statistic we have conducted for the balanced shape-restricted ANCOVA regression has good test size and better power compared with the F test statistic and F^s test statistic.

3.4 Unbalanced shape-restricted ancova

We have already studied the test statistic and power comparison for the balanced shape-restricted ANCOVA model. However, the case of balanced design is so ideal that in the real world we usually have the unbalanced design, i.e., at a given x, there are different number of observations from each treatment level. Hence we need to figure out a way to apply the shape-restricted ANCOVA in an unbalanced design.

3.4.1 Procedure of fit

Suppose we have k treatment levels and treatment level one is set to be the baseline, then the unbalanced shape-restricted ANCOVA proceeds as follows:

(1) Define a new variable z, where z=y for treatment level one and $z = y - r_1$ for treatment level two, $\cdots$, $z = y - r_{k-1}$ for treatment level k, $r_1, \cdots, r_{k-1}$ are some constants obtained by golden section search method (for $k = 2$) and downhill simplex method (for $k \geq 3$) in such a way that the sum of squares of errors is minimized.

(2) At any fixed x, average all the z's corresponding to this x, get $\bar{z}$.

(3) Now, our model is simply

$$\bar{\boldsymbol{z}} = \boldsymbol{\theta} + \boldsymbol{\varepsilon},$$

where $\theta_i = f(x_i)$, $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ is the weight matrix with $\boldsymbol{\Sigma}_{ii} = 1/w_i$ and $\boldsymbol{\Sigma}_{ij} = 0$, for $j \neq i$, w_i is the weight at x_i, i.e., the number of observations corresponding to x_i, $i = 1, \cdots, n$.

(4) Multiply $\boldsymbol{\Sigma}^{-1/2}$ on both side of the model to get

$$\boldsymbol{\Sigma}^{-1/2}\bar{\boldsymbol{z}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta} + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}.$$

Define $\boldsymbol{z}^* = \boldsymbol{\Sigma}^{-1/2}\bar{\boldsymbol{z}}$, $\boldsymbol{\theta}^* = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta}$, $\boldsymbol{\varepsilon}^* = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}$, then $\boldsymbol{\varepsilon}^* \sim N(0, \sigma^2 \boldsymbol{I})$, where $\boldsymbol{I}$ is the identity matrix with $\boldsymbol{I}_{ii} = 1$ and $\boldsymbol{I}_{ij} = 0$, for $j \neq i$.

(5) Define $\boldsymbol{A}^* = \boldsymbol{A}\boldsymbol{\Sigma}^{1/2}$, where $\boldsymbol{A}$ is the constraint matrix. The shape restriction $\boldsymbol{A}\boldsymbol{\theta} \geq 0$ can be written as $\boldsymbol{A}^*\boldsymbol{\theta}^* \geq 0$. Define $V^* = \boldsymbol{\Sigma}^{-1/2}V$, where $V = \mathcal{L}(\mathbf{1})$ for monotone restrictions and $V = \mathcal{L}(\mathbf{1}, \boldsymbol{x})$ for convex restrictions. From $\boldsymbol{A}^*$ and V^*, we can get the edge vectors $\boldsymbol{\delta}^*$ (according to the method in section 2.3.2),

(6) Apply the hinge algorithm on $\boldsymbol{y}^*$, $\boldsymbol{\delta}^*$, and V^*, using $\boldsymbol{\delta}^*$ and V^* as hinges to get $\hat{\boldsymbol{\theta}^*}$.

(7) Obtain the least squares estimator for $\boldsymbol{\theta}$ via $\hat{\boldsymbol{\theta}} = \boldsymbol{\Sigma}^{1/2}\hat{\boldsymbol{\theta}^*}$.

3.4.2 Feet data example

Dr. Meyer found that in general boy's shoes are wide and comfortable, but girl's shoes are narrow and uncomfortable, especially for dress shoes, given that the shoes are of the same length. However, she did not believe that girl's feet are narrower than those of boy's, when they have the same feet length. Hence, she collected data on feet dimensions from her daughter's class to check if boys have wider feet than girls. All students in the class were asked to measure the width and length of their feet. Table 3.5 is a simple display of the

feet data (see Appendix one for the whole data set), where ID is the student's identification number, *length* and *width* are in the unit of centimeters, the symbol 'G' is for girls and 'B' is for boys.

Table 3.5: Feet data

ID	Length	Width	Gender
1	21.6	7.9	G
2	22.5	8.6	G
3	22.9	8.8	B
4	22.9	8.5	G
5	23	8.8	G
6	23.6	9	B
7	23.6	9.3	G
8	23.7	7.9	G
9	23.9	9.3	B
10	24	9.2	B
⋮	⋮	⋮	⋮
39	27.5	9.8	B

Since longer feet might tend to be wider than shorter feet. Also, as feet get longer, the increasing rate of their width might slow down, but their positive association might still remain. Based on these, a reasonable choice of the relationship between the width and the length is increasing concave. We are interested to know if there are significant feet width differences between boys and girls.

Figure 3.5(a) shows the increasing concave fit for the feet data. For those boys and girls that have the same feet length, the average width of boys' feet is 0.2268 centimeters wider than those of girls'. As a comparison, we also apply monotone restriction (Figure 3.5(b)) and the gap between boys and girls is 0.2428, where boys' feet are again wider. From the fits, we can see that monotone regression might not be a realistic choice, since it has so many jumps and the fit is far from smooth. The gap between boys and girls is 0.2325 from an ordinary ANCOVA regression where feet width is linear in feet length, with a p-value from F-test 0.0806, which suggests that the gap is not statistically significant.

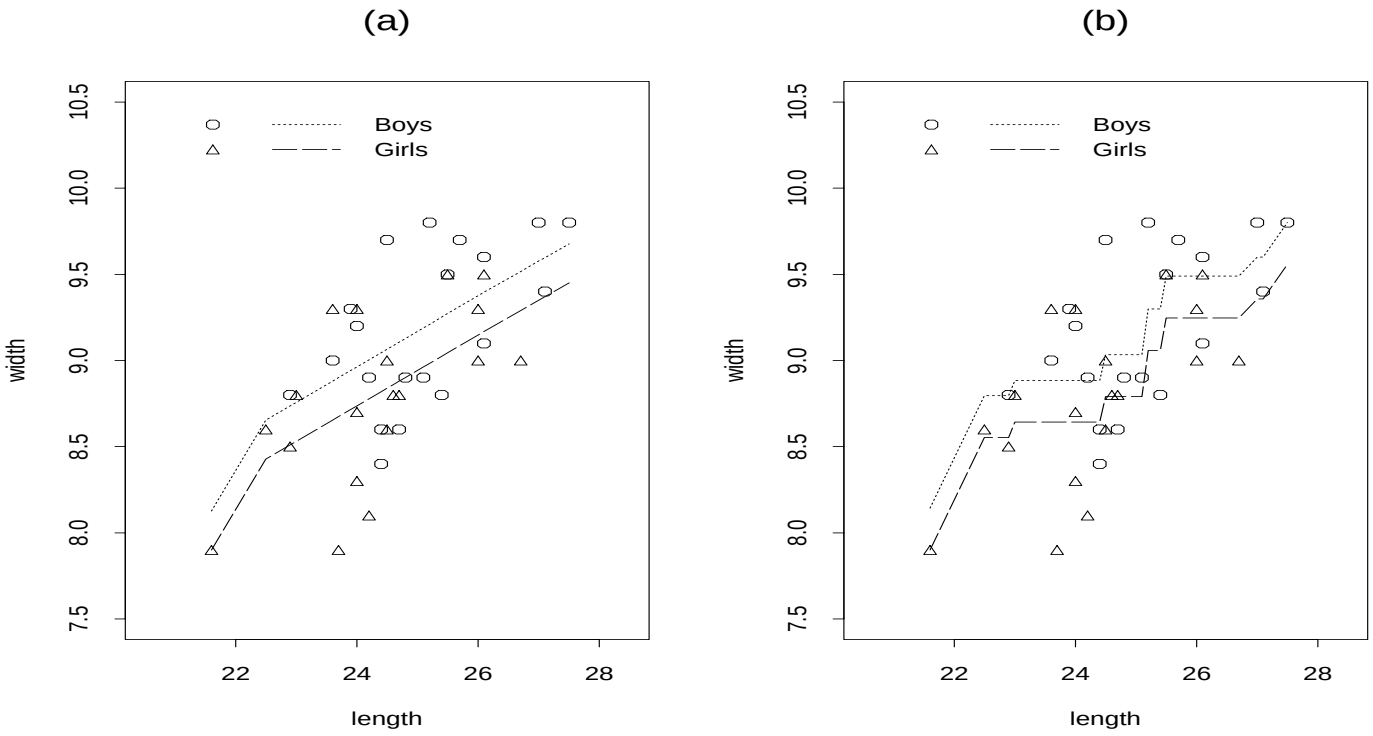

Figure 3.5: (a) Feet data: parallel increasing concave fit. (b) Feet data: parallel monotone fit. The dotted line refers to boys and the dashed line refers to girls.

3.4.3 Senic data example

The senic data is from *"Special issue: the senic project", American Journal of Epidemiology 111(1980), p.465-653.* A total of 113 hospitals from four geographic regions were examined for their infection risks, where the infection risk is defined as the average estimated probability of patients acquiring infection in a hospital.

Table 3.6 lists part of the senic data, where variable ID is the hospital identification number; variable $InfctRsk$ is the infection risk; variable $Region$ refers to geographic regions, where 1=NE, 2=NC, 3=S, 4=W; variable $Census$ counts the average number of patients in hospital per day during the study period.

Table 3.6: Senic data

ID	InfctRsk	Region	Census
1	3.1	3	20
2	2.9	1	37
3	3.7	4	37
4	4.2	2	38
5	3.1	1	39
6	1.3	2	40
7	2.7	4	40
8	5.4	4	42
9	2.1	2	44
10	4.2	4	47
11	2.6	4	47
12	1.3	3	49
13	4.6	3	50
14	1.6	2	51
15	4.5	4	51
16	2	3	52
17	5.6	4	53
18	5.3	4	55
⋮	⋮	⋮	⋮
113	5.9	1	791

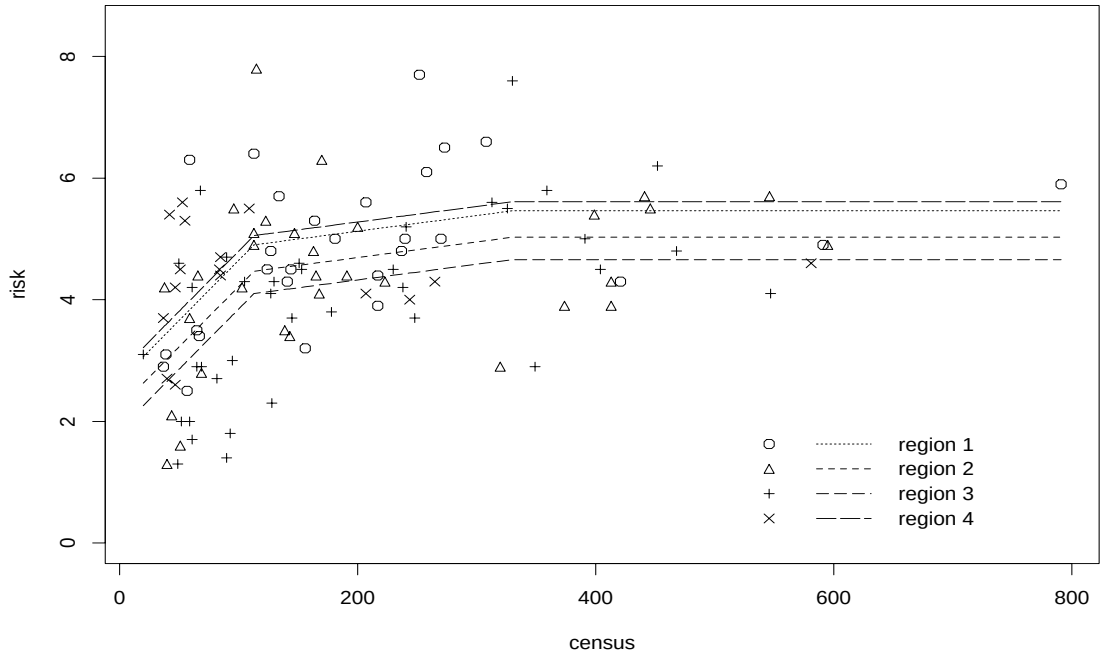

Figure 3.6: Senic data: parallel increasing concave fit.

Intuitively, the infection risk in a hospital might increase when more patients enrolled in the hospital, but the rate of increase might decrease. Hence, an increasing concave curve might be a good choice to reflect the relationship between the infection risk and the census. Our primary interest is to check if the infection risks are significantly different in the four regions. Figure 3.6 shows the increasing concave fit for the senic data. If we choose region 4 as baseline, and define r_i, $i = 1, 2, 3$, as the gap between region i and the baseline, then the least squares estimates of r_i's are $r_1 = -0.1515$, $r_2 = -0.5852$, $r_3 = -0.9531$. Region 4 has the higest infection risk and region 3 has the lowest infection risk. We also transform the covariate variable census into logarithmic scale, then fit an ordinary ANCOVA regression, the results are shown in the following table.

Table 3.7: Senic data: parameter estimates from ordinary ANCOVA regression

Parameter	Estimate	SD	P-Value
r_1	-0.0132	0.3730	0.9719
r_2	-0.4511	0.3637	0.2175
r_3	-0.8365	0.3528	0.0195

Other methods, such as asymptotic regression via nonlinear least squares can also give reasonable fit to this senic data set.

For the unbalanced shape-restricted ANCOVA, we can obtain the shape-restricted fit, but we have difficulty to define a suitable test statistic, hence the statistical inference for the unbalanced shape-restricted ANCOVA is not available so far. However, we will revisit these two examples in Chapter 5, where the inference is conducted by Bayesian methods.

Chapter 4

A Bayesian approach to shape-restricted semiparametric ancova

For the shape-restricted model, the model degrees of freedom are not clear, which makes frequentist hypothesis testing difficult. Also the least-squares estimator from the shape-restricted model is not smooth. We apply Bayesian approach with vague priors to the shape-restricted semiparametric ANCOVA model to act as a smoothing technique and to perform hypothesis testing on parameters related to the categorical variables.

We study the same model as before, namely

$$y_i = f(x_i) + \beta_1 d_{1i} + \cdots + \beta_{k-1} d_{(k-1)i} + \sigma\varepsilon_i, \tag{4.0.1}$$

where $i = 1, \cdots, n_T$, k is the number of treatment levels, f is any shape-restricted function, $\beta_1, \cdots, \beta_{k-1}$ and σ are unknown constants, and the ε_i's are independently distributed standard normal random errors. The d_{ji}'s, $i = 1, \cdots, n_T$, $j = 1, \cdots, k-1$, are indicator variables, where

$$d_{ji} = \begin{cases} 1 & \text{if } i^{th} \text{ obs is in group } j+1 \\ 0 & \text{otherwise.} \end{cases}$$

We assume no interaction terms in the model.

4.1 Backgroud on Bayesian analysis

Recently, the use of Bayesian methods has been greatly extended. One reason is that improvements in computation have made these methods more widely feasible. The Gibbs sampler is an intensive Bayesian computational method, which has been broadly applied. In addition to being natural and appealing when prior information on parameters is available, Bayesian

methods are sometimes the most natural useful ones in practice, and they are able to get useful solutions in some applications where frequentist approaches can not.

Bayesian and frequentist statistical methods make statistical inference based on different frameworks. The latter is based on the idea of an experiment that can be repeated many times, while the former regards parameters as random and combines the prior information on the parameters with the data from a single performance of an experiment. Prior information may be formulated as "vague" or "noninformative" if inference is to be based solely on the data.

Let $\boldsymbol{\phi}$ be the vector containing unknown population parameters, Bayesian analysis is performed by combining the prior information $\pi(\boldsymbol{\phi})$ and the data $\mathcal{D}$ into the posterior distribution of $\boldsymbol{\phi}$, given $\mathcal{D}$, from which all the decisions and inferences are made. We denote $p(\boldsymbol{\phi}|\mathcal{D})$ as the posterior distribution of $\boldsymbol{\phi}$ given $\mathcal{D}$.

We do need prior information in Bayesian analysis. On the one side, the advantage of a prior is to include some valuable information about the free parameters in the model. However, on the other side, it is hard to choose a prior density when we know nothing about the parameters. Asymptotically, the data will overwhelm the choice of prior, so if we have infinite data sets, priors would be irrelevant and Bayesian and frequentist results would converge (Kass and Raftery, 1994).

4.1.1 Bayes' theorem

Theorem 4 *Conditional on the observed data $\mathcal{D}$, the distribution of $\boldsymbol{\phi}$ is*

$$p(\boldsymbol{\phi}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\phi})\pi(\boldsymbol{\phi})}{p(\mathcal{D})}$$

where

- *$\boldsymbol{\phi}$ is the vector containing unobservable population parameters of interest;*
- *$\pi(\boldsymbol{\phi})$ is the prior density;*
- *$p(\mathcal{D})$ is the density of the data and can be considered as a constant;*

- $p(\mathcal{D}|\boldsymbol{\phi})$ *is the density of the data given the parameter valves and is also known as the likelihood function of the parameter* $\boldsymbol{\phi}$*, denoted by* $L(\boldsymbol{\phi}|\mathcal{D}) = p(\mathcal{D}|\boldsymbol{\phi})$*;*

- $p(\boldsymbol{\phi}|\mathcal{D})$ *is the posterior density.*

Bayes' theorem can also be expressed as

$$p(\boldsymbol{\phi}|\mathcal{D}) \propto L(\boldsymbol{\phi}|\mathcal{D})\pi(\boldsymbol{\phi})$$

4.2 Prior distributions

In order to proceed in a Bayesian framework, we must put priors over all the unknown parameters. The unknown parameters in our case are the coefficients of the edge vectors, b_j, $j = 1, \cdots, m$, where $m = n - 1$ for monotone restriction, $m = n - 2$ for convex restriction, n is the number of non-duplicated covariate values; the coefficients of the linear vectors, c_j, $j = 1, \cdots, r$, where $r = 1$ for monotone and $r = 2$ for convex constraint; the coefficients β_j, $j = 1, \cdots, k - 1$ and the model variance σ^2.

Gamma prior for edge coefficients

A gamma prior is a reasonable choice for the coefficients of the edge vectors (Meyer and Laud, 2005), because the coefficients need to be non-negative and we believe that their priors should be unimodal. In addition, a gamma prior can have small mean and large variance (vague). Suppose $b_j, j = 1, \cdots, m$ have independent, identical Gamma (α, γ) distributions, with the density

$$f(b) = \frac{b^{\alpha-1}\gamma^{\alpha}}{\Gamma(\alpha)} e^{-\gamma b},$$

where

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt, \quad x > 0.$$

The γ parameter here is considered as a "smoothing" parameter, a larger choice of γ will produce a smoother fit, a smaller γ will produce a fit closer to the maximum likelihood estimate.

For monotone regression with equally spaced x values, let $m = n - 1$, the edge vectors $\boldsymbol{\delta}^j$, $j = 1, \cdots, m$ can be chosen as

$$\delta_i^j = \begin{cases} \frac{j-n}{n} & \text{if } 1 \le j \le i \\ \frac{j}{n} & \text{if } i+1 \le j \le n. \end{cases}$$

Then the estimate of θ_i is

$$\hat{\theta}_i = \sum_{j=1}^{m} \hat{b}_j \delta_i^j + w, i = 1, \cdots, n$$

where w is a constant. Hence

$$\hat{\theta}_1 = \sum_{j=1}^{m} \hat{b}_j \delta_1^j + w, \quad \hat{\theta}_n = \sum_{j=1}^{m} \hat{b}_j \delta_n^j + w.$$

Since $\delta_n^j - \delta_1^j = 1$ for each $j = 1, \cdots, m$, we have

$$\sum_{j=1}^{m} \hat{b}_j = \hat{\theta}_n - \hat{\theta}_1.$$

The prior mean of b_j is α/γ, the parameter α hence can be chosen by the relationship $m\alpha/\gamma = G$, where G is a "guess" of the range of the fitted values.

For convex regression with equally spaced x values, let $m = n - 2$, we choose the hinge vectors $\boldsymbol{\sigma}^j, j = 1, \cdots, m$, to be the rows of the following matrix

$$\begin{pmatrix} 0 & 0 & \frac{1}{n-2} & \cdots & \cdots & \frac{n-3}{n-2} & 1 \\ 0 & 0 & 0 & \frac{1}{n-3} & \cdots & \frac{n-4}{n-3} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

Then we project each $\boldsymbol{\sigma}^j$ onto $\mathcal{L}(\mathbf{1}, \boldsymbol{x})$ and get the projections $\hat{\boldsymbol{\sigma}}^j$ as

$$\hat{\boldsymbol{\sigma}}^j = a_0 \mathbf{1} + a_1 \boldsymbol{x}, \quad j = 1, \cdots, m,$$

where a_0 and a_1 are constants. Define $\boldsymbol{\delta}^j = \boldsymbol{\sigma}^j - \hat{\boldsymbol{\sigma}}^j$, then the estimate of θ_i is

$$\begin{aligned}\hat{\theta}_i &= \tilde{w}_0 + \tilde{w}_1 x_i + \sum_{j=1}^{m} \hat{b}_j \delta_i^j \\ &= w_0 + w_1 x_i + \sum_{j=1}^{m} \hat{b}_j \sigma_i^j\end{aligned}$$

where $i = 1, \cdots, n$, $\tilde{w}_0$, $\tilde{w}_1$, w_1 and w_2 are some constants. Hence,

$$\hat{\theta}_1 = w_0 + w_1 x_1, \quad \hat{\theta}_2 = w_0 + w_1 x_2, \quad \hat{\theta}_n = w_0 + w_1 x_n + \sum_{j=1}^{m} \hat{b}_j \tag{4.2.1}$$

From (4.2.1), we can get

$$\begin{aligned}\sum_{j=1}^{m} \hat{b}_j &= \hat{\theta}_n - w_0 - w_1 x_n \\ &= (\hat{\theta}_n - \hat{\theta}_1) - \frac{\hat{\theta}_2 - \hat{\theta}_1}{x_2 - x_1}(x_n - x_1).\end{aligned}$$

Geometrically, if we draw a line by connecting points $(x_1, \hat{\theta}_1)$ and $(x_2, \hat{\theta}_2)$ and let it join with the line $x = x_n$, and denote their joint point as (x_n, z), then the distance between the points (x_n, z) and $(x_n, \hat{\theta}_n)$ is equal to $\sum_{j=1}^{m} \hat{b}_j$, which can be interpreted as "curvature" of the function. If we make a reasonable guess "G" about the distance, then α is again evaluated from the equation $m\alpha/\gamma = G$.

Normal prior for the coefficients c's and β's

We choose c_j's, $j = 1, \cdots, r$, where $r = 1$ for monotone restriction, $r = 2$ for convex restriction, to have independent normal distribution with mean $\mu_j = 0$, and variance $M_j = 10000$. The prior distribution of β_j's, $j = 1, \cdots, k-1$ is chosen to be independent normal with mean $w_j = 0$ and variance $W_j = 10000$. The reason for choosing normal prior is that it is convenient, the prior mean can take any value along the real line and our prior beliefs are symmetric and unimodal about some point along the real line. Large variance means that prior does not affect the final fit, only the data itself determine the estimates of c's and β's.

Inverse-gamma prior for model variance σ^2

The commonly chosen priors for the model variance σ^2 are inverse-gamma prior (Gelfand & Smith 1990, Gelfand *et al* 1990, Carlin & Polson 1991, *etc*) and Jeffreys' prior. Both of them can produce positive σ values with large variance. However, Jeffreys' prior can not be integrated to 1, and we found that the use of Jeffreys' prior made our final results unstable. The inverse-gamma prior $IG(g_1, g_2)$ can be integrated to 1, with the density

$$f(\sigma^2) = \frac{(\sigma^2)^{-(g_1+1)} g_2^{g_1}}{\Gamma(g_1)} e^{-g_2/\sigma^2}, \quad \sigma^2 > 0.$$

The mean and variance of σ^2 are

$$E(\sigma^2) = \frac{g_2}{g_1 - 1}, \quad V(\sigma^2) = \frac{g_2^2}{(g_1-1)^2(g_1-2)}.$$

We choose $g_1 = 2.0001$ and $g_2 = 1.0001$, so that

$$E(\sigma^2) = 1, \quad V(\sigma^2) = 10000.$$

For convenience, we use $\tau = 1/\sigma^2$ instead σ^2 in the likelihood functions, hence τ has a Gamma prior with parameters (g_1, g_2). The density of τ is

$$f(\tau) = \frac{\tau^{g_1-1} g_2^{g_1}}{\Gamma(g_1)} e^{-g_2\tau}, \quad \tau > 0.$$

4.3 Likelihood function, posterior distribution, and full conditional distributions

In order to simplify the problem, we focus on the balanced design with $k = 2$ treatment levels and we assume $\beta = \beta_1$ in model (4.0.1).

Let $\theta_i = f(x_i) = \sum_{j=1}^{m} b_j \delta_i^j + \sum_{j=1}^{r} c_j v_i^j$, where $m = n - 1$ and $r = 1$ for monotone restriction; $m = n - 2$ and $r = 2$ for convex restriction, $\tau = \sigma^{-2}$, total sample size is $2n$. the likelihood function L for $\boldsymbol{\phi}$, is proportional to

$$L \propto \tau^n \exp\left\{ -\frac{\tau}{2} \sum_{i=1}^{n} \left[(y_i - \theta_i)^2 + (y_{n+i} - \theta_i - \beta)^2 \right] \right\}$$

$$= \tau^n \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}\left[y_{n+i}^2 + y_i^2 + \beta^2 - 2y_{n+i}\beta + 2\left(\sum_{j=1}^{r} c_j v_i^j + \sum_{j=1}^{m} b_j \delta_i^j\right)^2\right]\right\}$$
$$\times \exp\left\{-\tau\sum_{i=1}^{n}\left[\left(\sum_{j=1}^{r} c_j v_i^j + \sum_{j=1}^{m} b_j \delta_i^j\right)(\beta - y_i - y_{n+i})\right]\right\}$$

We apply Gamma(α_j, γ) prior on the edge coefficients b_j, $N(\mu_j, M_j)$ prior on the linear coefficients c_j, $N(w, D)$ prior on β, and Gamma(g_1, g_2) prior on τ, all independent, then the joint prior density is proportional to

$$\prod_{j=1}^{m} b_j^{\alpha_j - 1}\exp(-b_j\gamma)\exp\left\{\sum_{j=1}^{r} -\frac{(c_j - \mu_j)^2}{2M_j}\right\}\exp\left\{-\frac{(\beta - w)^2}{2W}\right\}\tau^{g_1 - 1}\exp\{-g_2\tau\},$$

Hence, the posterior density is proportional to

$$\tau^n \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}\left[y_{n+i}^2 + y_i^2 + \beta^2 - 2y_{n+i}\beta + 2\left(\sum_{j=1}^{r} c_j v_i^j + \sum_{j=1}^{m} b_j \delta_i^j\right)^2\right]\right\}$$
$$\times \exp\left\{-\tau\sum_{i=1}^{n}\left[\left(\sum_{j=1}^{r} c_j v_i^j + \sum_{j=1}^{m} b_j \delta_i^j\right)(\beta - y_i - y_{n+i})\right]\right\}$$
$$\times \prod_{j=1}^{m} b_j^{\alpha_j - 1}\exp(-b_j\gamma)\exp\left\{\sum_{j=1}^{r} -\frac{(c_j - \mu_j)^2}{2M_j}\right\}\exp\left\{-\frac{(\beta - w)^2}{2W}\right\}\tau^{g_1 - 1}\exp\{-g_2\tau\}$$

The posterior distribution gives full information about the parameters. Various summary characteristics, such as posterior mean, posterior mode, posterior median, are used to interpret the nature of the parameters.

Let $[a|b]$ represent the conditional distribution of a given b. The conditional distributions of the model parameters $\tau, b_j, j = 1, \cdots, m$, $c_j, j = 1, \cdots, r$, and β are as follows:

1. $[b_s \mid c_j, j = 1, \cdots, r;\, b_j, j \neq s;\, \tau;\, \beta]$ is from a distribution with the density proportional to

$$b_s^{\alpha_s - 1}\exp\left\{-\tau\sum_{i=1}^{n}(\delta_i^s)^2\left[b_s - \frac{\sum_{i=1}^{n}\delta_i^s l_i - \gamma/\tau}{2\sum_{i=1}^{n}(\delta_i^s)^2}\right]^2\right\},$$

where $l_i = y_{n+i} + y_i - \beta - 2\left(\sum_{j\neq s} b_j\delta_i^j + \sum_{j=1}^{r} c_j v_i^j\right)$.

2. $[c_s \mid b_j, j = 1, \cdots, m;\, c_j, j \neq s;\, \tau;\, \beta]$ has a normal distribution with

$$\text{mean} = \frac{\mu_s/M_s + \tau\sum_{i=1}^{n} l_i v_i^s}{1/M_s + 2\tau\sum_{i=1}^{n}(v_i^s)^2}, \quad \text{variance} = \frac{1}{2\tau\sum_{i=1}^{n}(v_i^s)^2 + 1/M_s},$$

where $l_i = y_{n+i} + y_i - \beta - 2\left(\sum_{j=1}^{m} b_j\delta_i^j + \sum_{j\neq s} c_j v_i^j\right)$.

3. $[\tau \mid b_j, j = 1, \cdots, m;\ c_j, j = 1, \cdots, r;\ \beta]$ has a gamma distribution with parameters $(n + g_1, \frac{sse}{2} + g_2)$, its density is proportional to

$$\tau^{n+g_1-1} \exp\left\{-\tau \left(sse/2 + g_2\right)\right\},$$

where $sse = \sum_{i=1}^{n} \left\{(y_i - \theta_i)^2 + (y_{n+i} - \theta_i - \beta)^2\right\}$.

4. $[\beta \mid b_j, j = 1, \cdots, m;\ c_j, j = 1, \cdots, r;\ \tau]$ has a normal distribution with

$$\text{mean} = \frac{\tau \sum_{i=1}^{n} l_i + w/W}{n\tau + 1/W}, \quad \text{variance} = \frac{1}{n\tau + 1/W},$$

where $l_i = y_{n+i} - \left(\sum_{j=1}^{m} b_j \delta_i^j + \sum_{j=1}^{r} c_j v_i^j\right)$.

Since our posterior density is rather complicated, it's impossible to get it directly. However, we can apply Gibbs sampler method to generate samples from the posterior distribution, because the model parameters $c_j, j = 1, \cdots, r$ and β have conditional distributions that can be easily sampled from, the parameters $b_j, j = 1, \cdots, m$ and τ have conditional distributions that can be obtained by latent variable technique through Gibbs sampler method.

4.4 Gibbs sampler

Monte Carlo Markov Chain (MCMC) methods are simulation techniques that generate samples from a given distribution. They are very useful in many statistical applications, especially when the joint distribution is hard to obtain while the full conditional distributions are easy to get. MCMC method is based on the following idea: given data $\mathcal{D}$, we can summarize the unknown model parameters, say $\boldsymbol{\phi}$, based on a set of random samples drawn from the posterior distribution $p(\boldsymbol{\phi}|\mathcal{D})$.

The Gibbs sampling algorithm is the best known and conceptually the simplest of MCMC methods. It was formally introduced by Geman & Geman (1984) and Gelfand & Smith (1990). It is widely applied into Bayesian analysis framework and can easily generate samples from an awkward posterior distribution, which has made it a great alternative to other sophisticated numerical or analytic approximation techniques.

The Gibbs sampling algorithm is essentially a Markovian updating scheme for extracting marginal distributions from the full conditional distribution. It requires all the full conditional distributions to be available for sampling. It is well known that specification of all full conditional distributions uniquely determines the full joint density (Besag 1974); i.e., for a collection of random variables $\theta_1, \theta_2, \cdots, \theta_k$, the joint density $p(\theta_1, \theta_2, \cdots, \theta_k)$ is uniquely determined by the full conditional densities $p(\theta_s|\theta_r, r \neq s), s = 1, 2, \cdots, k$. The Gibbs sampler proceeds as follows:

(1) Initialize an arbitrary starting set of values $\theta_1^{(0)}, \cdots, \theta_k^{(0)}$.

(2) Draw $\theta_1^{(1)}$ from $p(\theta_1|\theta_2^{(0)}, \cdots, \theta_k^{(0)})$, then $\theta_2^{(1)}$ from $p(\theta_2|\theta_1^{(1)}, \cdots, \theta_k^{(0)})$, and so on up to $\theta_k^{(1)}$ from $p(\theta_k|\theta_1^{(1)}, \cdots, \theta_{k-1}^{(1)})$.

(3) At t^{th} iteration, we would arrive at $(\theta_1^{(t)}, \theta_2^{(t)}, \cdots, \theta_k^{(t)})$, repeat step (2) until convergence is reached.

(4) After convergence is reached, we start to collect samples.

Geman and Geman showed that under mild conditions $\theta_s^{(t)} \xrightarrow{d} \theta_s \sim p(\theta_s)$ as $t \to \infty$. Thus, for t large enough we can treat $\theta_s^{(t)}$ as a realization from $p(\theta_s)$ density.

4.5 Latent variable technique

Latent variable technique (Damien and Walker, 2001) provides a convenient way to generate a sample from an unusual density. In our case, the conditional distribution of $[b_s \mid c_j, j = 1, \cdots, r; b_j, j \neq s; \tau; \beta]$ is proportional to

$$b_s^{\alpha_{s-1}} \exp\left\{-\tau \sum_{i=1}^{n} (\delta_i^s)^2 \left[b_s - \frac{\sum_{i=1}^{n} \delta_i^s l_i - \gamma/\tau}{2\sum_{i=1}^{n} (\delta_i^s)^2}\right]^2\right\},$$

which can be simplied into the following format

$$p(x) \propto x^{a-1} \exp\{-b(x-c)^2\},$$

where a, b, and c are constants. It's hard to generate a random sample directly from the above density. Hence, we define a latent random variable u, such that

$$p(x,u) \propto x^{a-1} I_{\{0 \le u \le e^{-b(x-c)^2}\}}$$

$p(x,u)$ is the joint distribution of x and u and $I_{\{t_1 \le x \le t_2\}}$ is an indicator variable, where

$$I_{\{t_1 \le x \le t_2\}} = \begin{cases} 1 & \text{if } t_1 \le x \le t_2 \\ 0 & \text{otherwise.} \end{cases}$$

The idea behind the latent variable is simple: first, we define a latent variable u, which together with x, construct a joint density of $p(x,u)$. There might be several choices of the joint density. The only requirement is that the full conditional distributions of $p(x|u)$ and $p(u|x)$ based on the chosen joint density are easy to obtain. Next, we apply Gibbs sampler method to generate a paired random sample of (x,u) from the fully conditional distributions. The x itself from the paired sample can be considered as a realization from the marginal distribution of $p(x)$.

In our case, $[u|x]$ has a uniform $\left(0, e^{b(x-c)^2}\right)$ distribution, and

$$p(x|u) \propto \begin{cases} x^{a-1} & \text{if } 0 \le u \le e^{-b(x-c)^2} \\ 0 & \text{otherwise.} \end{cases}$$

We can easily generate a sample u_1 from uniform (0,1) density, then $u_1 e^{b(x-c)^2}$ can be treated as a sample from $p(u|x)$. Also, since

$$0 \le u \le e^{-b(x-c)^2} \Longrightarrow s_1 \le x \le s_2$$

where $s_1 = c - \sqrt{\frac{-log(u)}{b}}$, $s_2 = c + \sqrt{\frac{-log(u)}{b}}$, we generate another sample u_2 from uniform (0,1) density, then $[u_2 s_2^a + (1-u_2)s_1^a]^{1/a}$ is a sample from $p(x|u)$.

4.6 Equal tailed posterior credible interval methods

In the fully Bayesian setting, once we specify the prior distributions $\pi(\boldsymbol{\phi})$, based on subjective information or prior knowledge, we can summarize all the inference about the model

parameter $\boldsymbol{\phi}$ from the posterior distribution $p(\boldsymbol{\phi}|\mathcal{D})$. One simple Bayes credible interval for ϕ_i based upon this posterior distribution can be defined as $\left(q_{\alpha/2}\left[p(\phi_i|\mathcal{D})\right], q_{1-\alpha/2}\left[p(\phi_i|\mathcal{D})\right]\right)$, where $q_{\alpha/2}\left[p(\phi_i|\mathcal{D})\right]$ and $q_{1-\alpha/2}\left[p(\phi_i|\mathcal{D})\right]$ are the lower and upper $\alpha/2$ percentiles of $p(\phi_i|\mathcal{D})$.

We apply Bayes credible interval to test the hypotheses

$$H_0 : \beta = 0 \quad vs. \quad H_1 : \beta \neq 0.$$

The procedures of Bayes credible interval can be summarized into three steps:

(1) Convergence loop: first, we generate B $(m + r + 2)$-tuples

$$(\hat{b}_j^{(i)}, j = 1, \cdots, m; \hat{c}_j^{(i)}, j = 1, \cdots, r; \hat{\tau}^{(i)}; \hat{\beta}^{(i)}), \quad i = 1, \cdots, B,$$

from the posterior distribution, where $B = 1000$, then examine their scatterplots to ensure that each of the parameter estimates converges within 1000 iterations.

(2) Sampling loop: after the process converges, we start to generate 20000 parameter estimates $(m + r + 2)$-tuples from the posterior distribution. By withdrawing only every 10^{th} iteration, the final sample can be considered as an independent sample with sample size $n = 2000$.

(3) Bayes credible interval: From the 2000 drawn independent $\hat{\beta^{(i)}}$ values, where $\hat{\beta}^{(i)}$ is the estimate of β from the $(10 \times i)^{th}$ sampling loop, we can find its 2.5% and 97.5% percentiles, denoted as $\hat{\beta}_{0.025}$ and $\hat{\beta}_{0.975}$, respectively. If this interval of $(\hat{\beta}_{0.025}, \hat{\beta}_{0.975})$ contains 0, then we can not reject the null hypothesis; otherwise, the null hypothesis is rejected.

4.7 Monitoring convergence of the sampler

Our posterior inference is based on a generated sample of 2000 iterations. These are found by running the Gibbs sampler for 21000 iterations in total. The first 1000 iterations are discarded as burn-in iterations, only after which the sampling algorithm is judged to have

converged. Then every tenth model visited in the last 20000 iterations is taken to be in the generated sample.

The plots in Figure 4.1 show the 2000 sampled parameter estimates of β (top), τ (middle) and c (bottom) when monotone restriction is applied in the model. Data are generated from a log underlying function $f(x_i) = 5log(x_i + 1)$, with true $\beta = 0$, $\gamma = 0.1$ and sample size $n = 40$.

The plots in Figure 4.2 show the 2000 sampled parameter estimates of β (top), τ (second) and c_1 (third) and c_2 (bottom) when convex restriction is applied in the model. Data are generated from a convex underlying function $f(x_i) = 2x_i + 1/x_i$, with true $\beta = 0$, $\gamma = 0.1$ and sample size $n = 40$.

From the plots, we can see that the Monte Carlo estimates of each parameter behave well. We also plot the histograms of these sampled parameter estimates (Figure 4.3 and Figure 4.4). The histograms of $\hat{\beta}$, $\hat{c}$'s have very nice bell shapes.

Durbin-Watson test statistic is used to detect first-order autocorrelation. It is defined as

$$D = \frac{\sum_{i=2}^{n}(t_i - t_{i-1})^2}{\sum_{i=1}^{n} t_i^2} \tag{4.7.1}$$

The distribution of the Durbin-Watson test statistic is symmetric about 2.00 and ranges from 0 to 4, where positive serial correlation results in a D near 0 and negative serial correlation results in a D close to 4. The value of Durbin-Watson statistic is close to 2 if the sample t_i's are uncorrelated.

We compute Durbin-Watson test statistic to test if the samples we have drawn from the posterior distribution are independent (Table 4.1). All the values of the Durbin-Watson test statistic are very close to 2. Hence we conclude that autocorrelation is not present, the samples we have obtained are independent samples.

We also compare the posterior mean estimates of β from Bayes shape-restricted regressions and the least-square estimate of β from shape-restricted regressions. 1000 data are generated from each of a log and a convex underlying function. For each data set, we apply both Bayes shape-restricted regression and shape-restricted regression, get the posterior mean

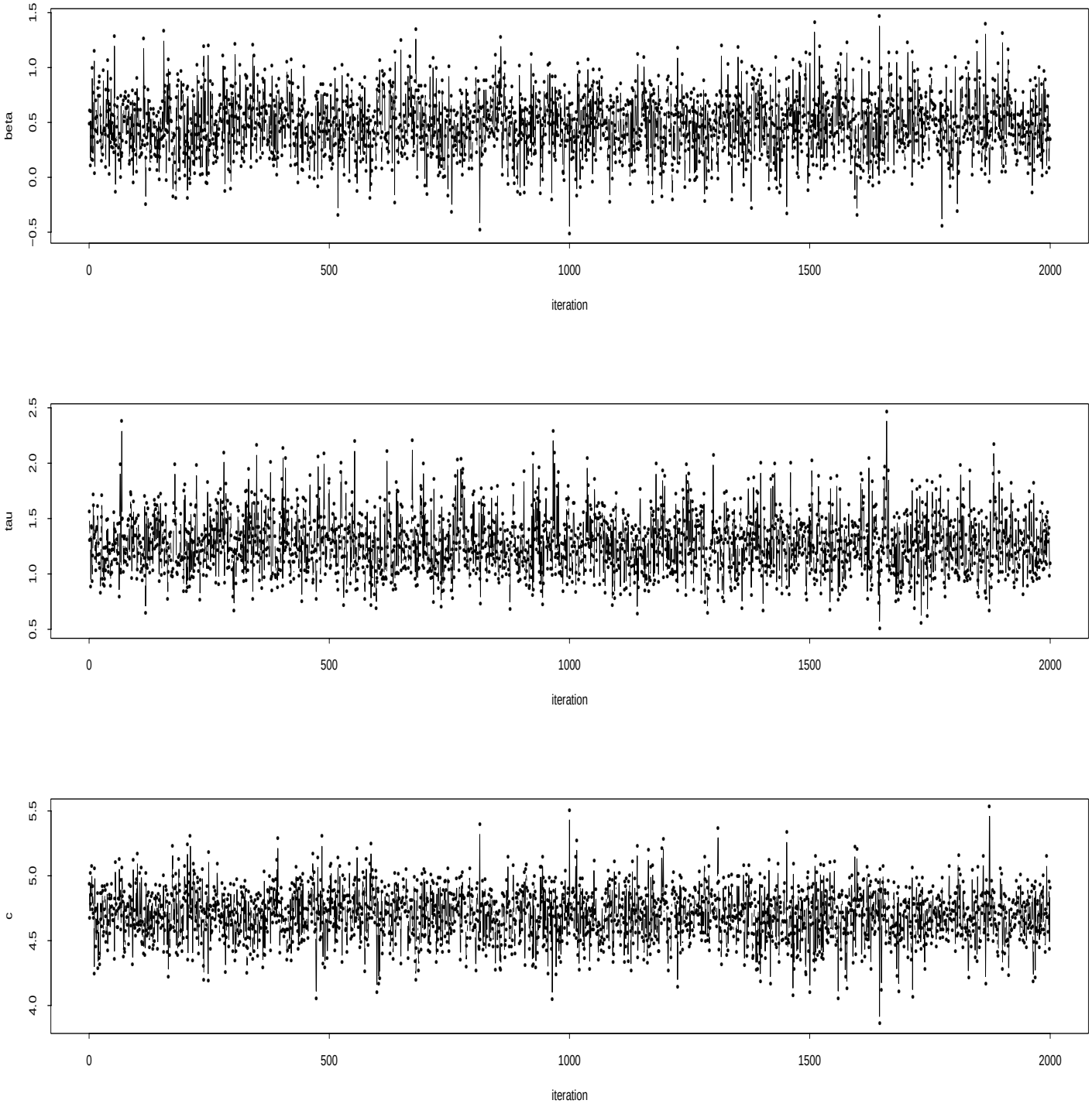

Figure 4.1: Plot of 2000 sampled parameter estimates of β (top), τ (middle) , c (bottom) from the Bayes monotone regression, using Gibbs sampler method.

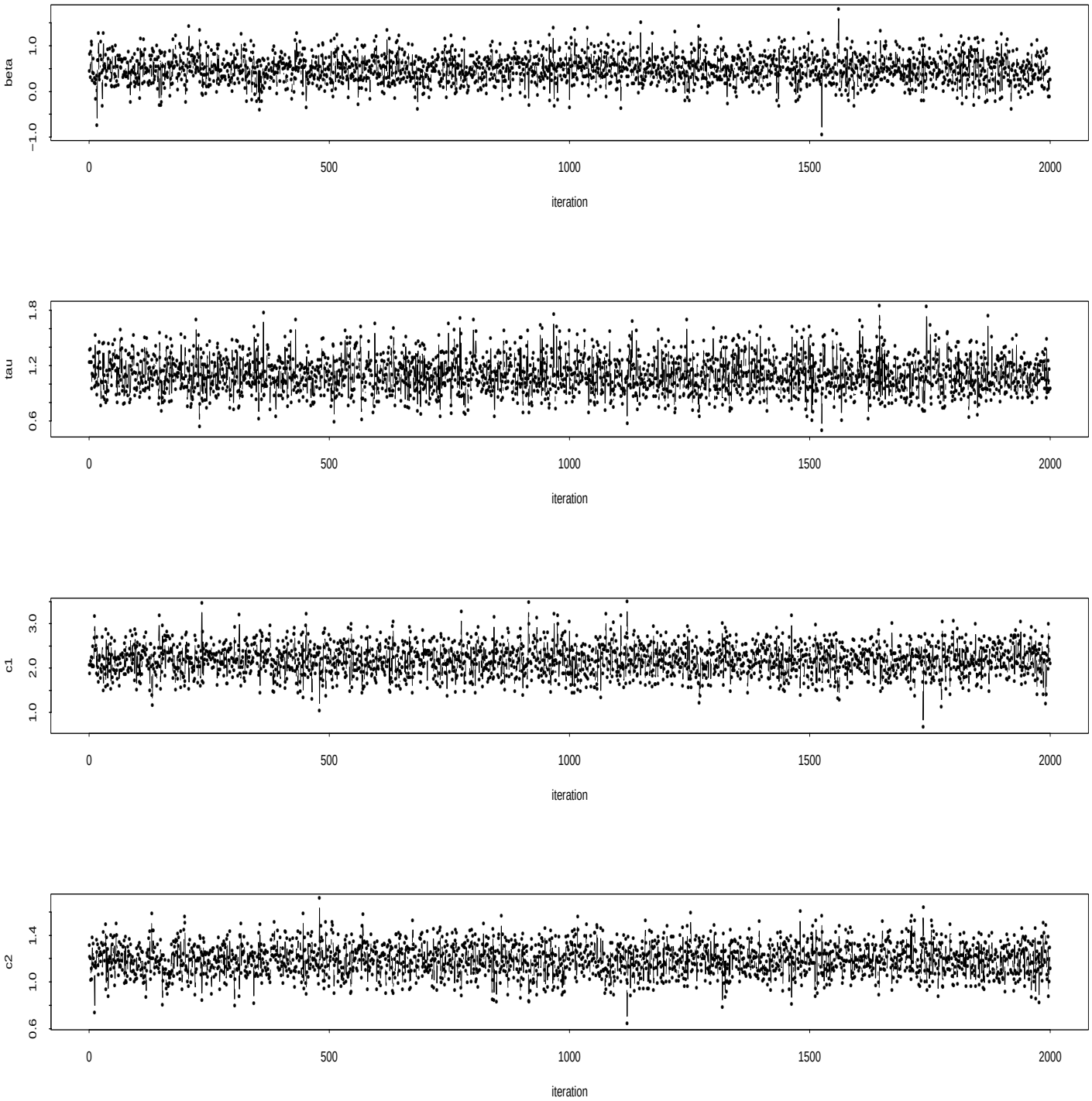

Figure 4.2: Plot of 2000 sampled parameter estimates of β (top), τ (second) , c_1 (third), c_2 (bottom) from the Bayes convex regression, using Gibbs sampler method.

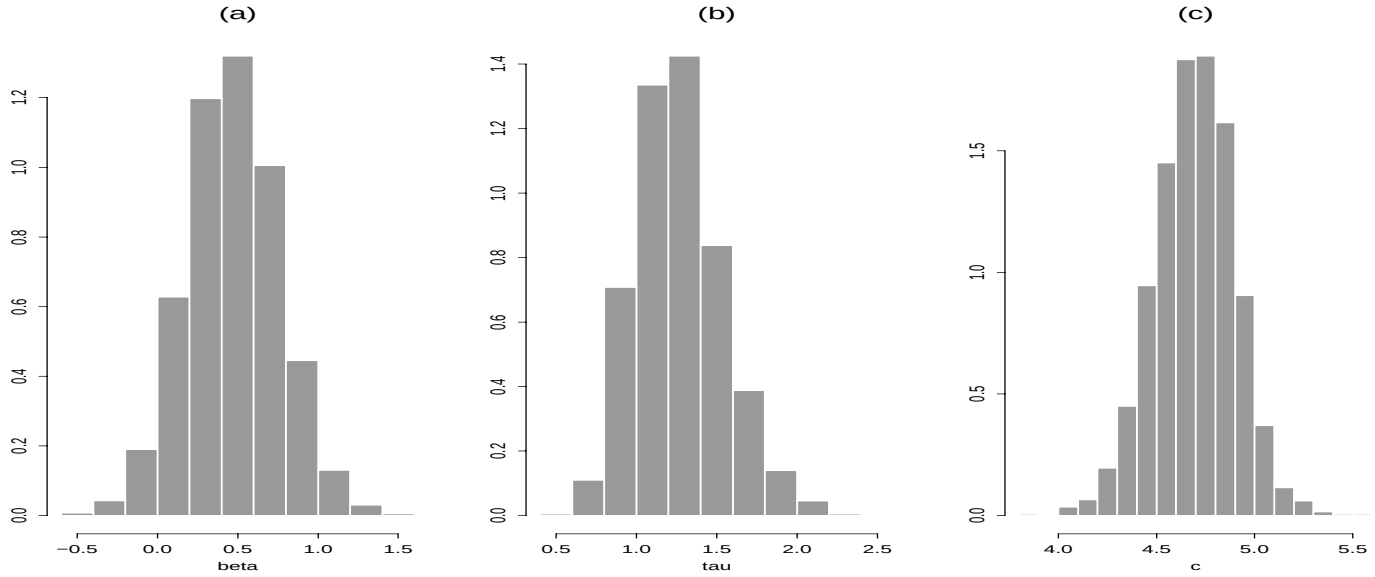

Figure 4.3: Histogram of the parameter estimates β (a), τ (b) , c (c) from the Bayes monotone regression.

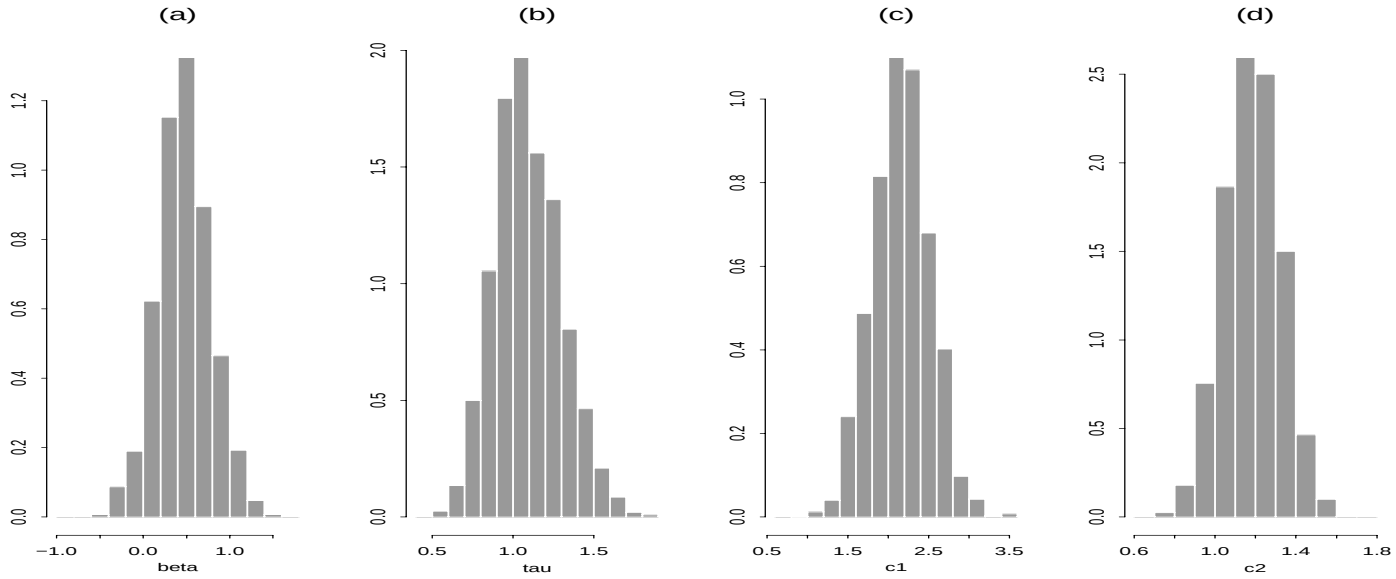

Figure 4.4: Histogram of the parameter estimates β (a), τ (b) , c_1 (c) and c_2 (d) from the Bayes convex regression.

Table 4.1: Durbin-Watson statistic for 1^{st} order autocorrelation, parameter estimates are from the Bayes monotone regression and the Bayes convex regression, set $\gamma = 0.1$.

Type of Fit	Parameter	Durbin-Watson D	1^{st} Order Autocorrelation
Bayes convex	β	1.987	0.006
	τ	2.088	-0.0044
	c_1	1.873	0.064
	c_2	1.843	0.078
Bayes monotone	β	2.005	-0.002
	τ	1.876	0.062
	c	2.080	-0.040

estimates of β and the least-square estimate of β, respectively. Figure 4.5 and Figure 4.6 show the histograms of posterior mean estimates of β and the least-square estimates of β. They have very close means and standard deviations.

4.8 Test size and power analysis

We apply Bayes credible interval to test the hypotheses

$$H_0 : \beta = 0 \quad vs. \quad H_1 : \beta \neq 0.$$

The test size and power behavior of this method is studied for different shape restrictions, underlying functions, γ's and sample sizes. As comparisons, the corresponding beta test from the same shape-restricted regression and the F-test from a parametric regression are also performed, and their test size and power are calculated.

For the log and linear underlying function, where $f(x_i) = 5log(x_i + 1)$ and $f(x_i) = x_i$, respectively, we fit Bayes monotone, monotone and linear regressions. Figure 4.7 shows the three different fits on a given data set, where the data are generated from the above log underlying function with true $\beta = 0$, with sample size 40. The γ are 0.01 (a), 0.1 (b) and 1

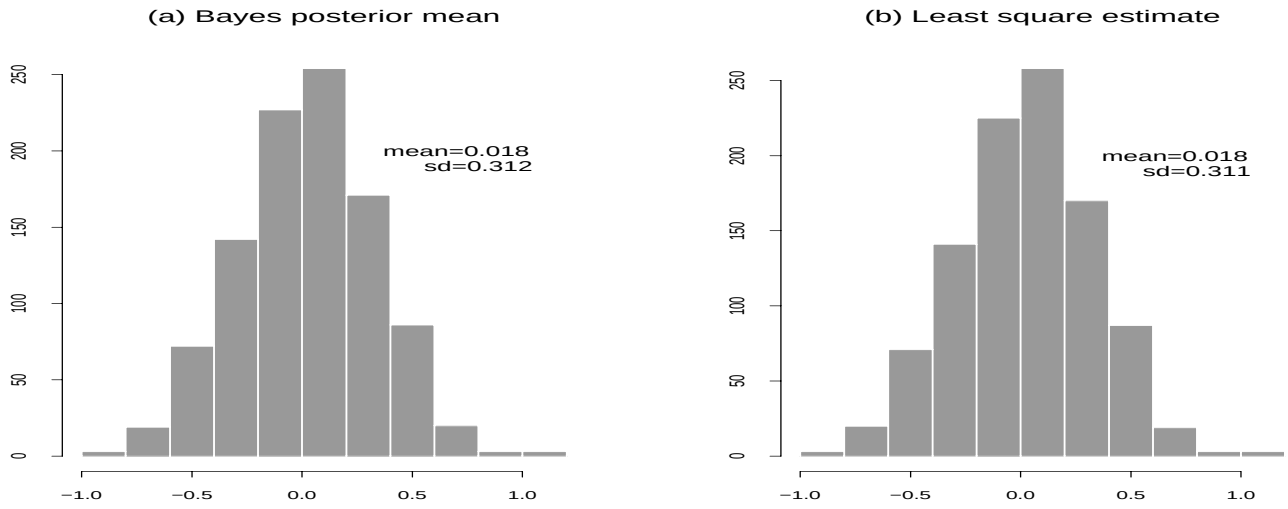

Figure 4.5: Comparison of posterior mean estimate of β from Bayes monotone regression with the least-square estimate of β from monotone regression, n=40, 1000 simulated datasets. The true β is 0.

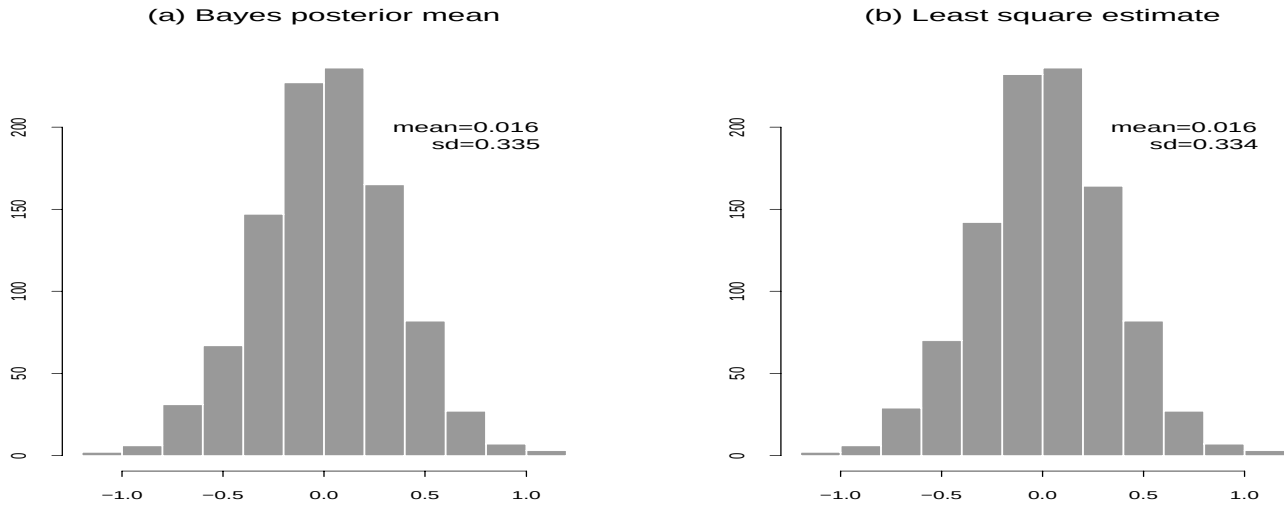

Figure 4.6: Comparison of posterior mean estimate of β from Bayes convex regression with the least-square estimate of β from convex regression, n=40, 1000 simulated datasets. The true β is 0.

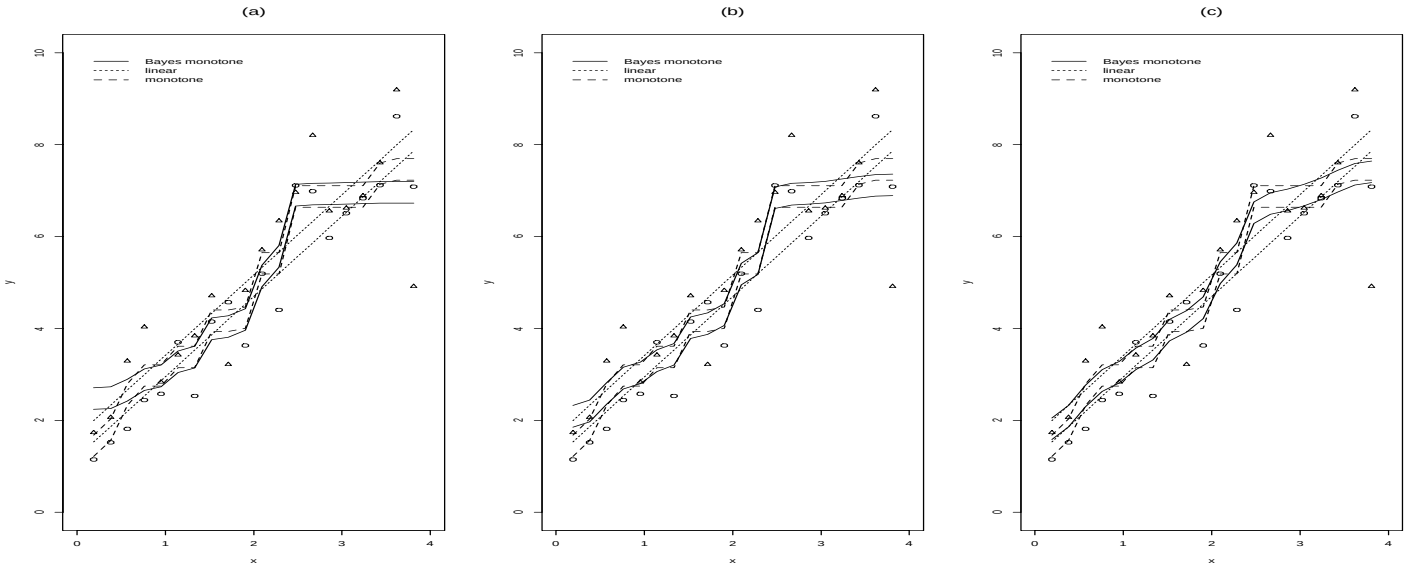

Figure 4.7: Comparisons of Bayes monotone, monotone, and linear regressions, sample size 40. The γ are 0.01 (a), 0.1 (b) and 1 (c) for Bayes monotone regression. The solid line is the Bayes monotone fit, the dashed line is the monotone fit, the dotted line is a simple linear fit.

(c) for Bayes monotone regression. The solid line is the fit from Bayes monotone regression, the dashed line is the fit from monotone regression, and the dotted line is the fit from a linear regression. The Bayes monotone fits are smoother than monotone fit, and as γ increases, the Bayes monotone fits are close to the simple linear fit.

For the convex and quadratic underlying function, where $f(x_i) = 2x_i + 1/x_i$ and $f(x_i) = x_i^2$, respectively, we fit Bayes convex, convex and quadratic regressions. Figure 4.8 shows the three different fits on a given data set, where the data are generated from the above convex underlying function with true $\beta = 0$, and sample size 40. The γ are 0.01 (a), 0.1 (b) and 1 (c) for Bayes convex regression. The solid line is the fit from Bayes convex regression, the dashed line is the fit from convex regression, and the dotted line is the fit from quadratic regression. For smaller γ, the fits of Bayes convex regression and convex regression are similar, except that the former is smoother. As γ increases, Bayes convex fit tends to be close to the quadratic fit.

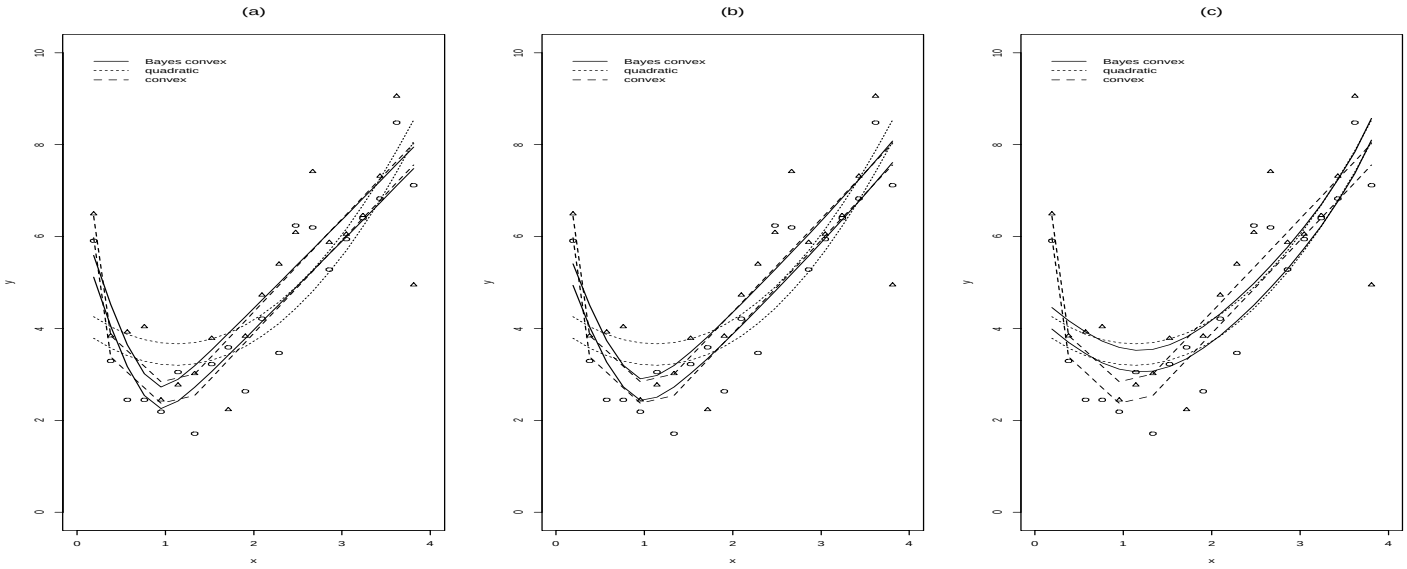

Figure 4.8: Comparisons of Bayes convex, convex and quadratic regressions, sample size 40. The γ are 0.01 (a), 0.1 (b) and 1 (c) for Bayes convex regression. The solid line is the Bayes convex fit, the dashed line is the convex fit, the dotted line is a quadratic fit.

For given log and linear underlying functions, We use three different methods, Bayes credible interval, Beta test, and F-test to perform the hypothesis tests. For Bayes credible interval method, we try three different γ values, 0.01, 0.1 and 1.0 and fit two parallel Bayes monotone curves; for F-test method, we fit two parallel straight lines to the data, then conduct an F-test; for Beta test method, we fit two parallel monotone curves to the data, then apply the beta test as mentioned in Chapter 3. For each sample size and method combination, 10000 simulations are generated. The test size and power of each method is studied and summaried in Table 4.2 and Table 4.3, respectively. For both log and linear underlying function, Bayes credible interval method gives reasonable test sizes, and it has the highest power in most of the cases, comparing with the other two methods. As γ increases, the power from Bayes credible interval method increases.

We perform the similar procedures to convex and quadratic underlying functions. For Bayes credible interval method, we fit two parallel convex curves to the data; for F-test

Table 4.2: Test sizes for Bayes credible interval method (monotone restriction), F-test and beta test (monotone restriction) when underlying functions are log and linear, respectively.

log regression function

Sample size	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1.0$	F-test	Beta test
20	0.031	0.053	0.065	0.039	0.031
40	0.038	0.060	0.066	0.036	0.039
60	0.045	0.053	0.062	0.035	0.044

linear regression function

Sample size	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1.0$	F-test	Beta test
20	0.043	0.058	0.064	0.051	0.040
40	0.047	0.058	0.063	0.050	0.045
60	0.047	0.056	0.056	0.050	0.046

Table 4.3: Power for Bayes credible interval method (monotone restriction), F-test and beta test (monotone restriction) when underlying functions are log and linear, respectively.

log regression function

Sample size	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1.0$	F-test	Beta test
20	0.478	0.575	0.596	0.425	0.512
40	0.846	0.874	0.883	0.830	0.834
60	0.960	0.970	0.974	0.955	0.959

linear regression function

Sample size	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1.0$	F-test	Beta test
20	0.544	0.611	0.626	0.560	0.488
40	0.857	0.877	0.885	0.868	0.850
60	0.967	0.978	0.974	0.967	0.963

method, we fit two parallel quadratic curves. The test size and power of each method is studied and summaried in Table 4.4 and Table 4.5, respectively. For each sample size, 10000 simulations are generated.

For the convex underlying function, F-test has very poor test size and lower power. Beta test has the best test size and good power. Bayes credible interval method has similar power and comparable test size to the Beta test.

For the quadratic underlying function, each of three methods has good test size. In most of the cases, Bayes credible interval method has best power comparing to the other two tests.

Table 4.4: Test size for Bayes credible interval method (convex restriction), F-test and beta test (convex restriction) when underlying functions are convex and quadratic, respectively.

convex regression function

Sample size	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1.0$	F-test	Beta test
20	0.049	0.058	0.041	0.044	0.055
40	0.042	0.040	0.036	0.025	0.052
60	0.043	0.049	0.037	0.010	0.051

quadratic regression function

Sample size	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1.0$	F-test	Beta test
20	0.055	0.059	0.058	0.050	0.053
40	0.048	0.052	0.049	0.050	0.049
60	0.045	0.058	0.055	0.050	0.049

From the above test size and power analysis, we conclude that the Bayes credible interval method can give reasonable results, comparing to the beta test and the F-test. The benefit of the Bayes credible interval method is that we do not need to formally construct a test statistic. All we need to do is to draw samples from the posterior distribution, then gain all the statistical inferences and results for hypotheses tests from the drawn samples. For an unbalanced data set, the semiparametric shape-restricted fit can be obtained, however, the frequentist test statistic is very hard to get, mainly because the error degree of freedom is hard to address. The Bayes credible interval method has no difficulty to deal with unbalanced data. The limitation of the Bayes credible interval method is that we need to know the prior

Table 4.5: Power for Bayes credible interval method (convex restriction), F-test and beta test (convex restriction) when underlying functions are convex and quadratic.

convex regression function

Sample size	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1.0$	F-test	Beta test
20	0.537	0.596	0.600	0.530	0.566
40	0.871	0.865	0.850	0.790	0.870
60	0.963	0.964	0.968	0.887	0.968

quadratic regression function

Sample size	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1.0$	F-test	Beta test
20	0.570	0.610	0.590	0.508	0.562
40	0.877	0.875	0.870	0.864	0.866
60	0.973	0.970	0.971	0.967	0.968

information, and we usually formulate the prior as "vague" if the statistical inferences are based solely on the given data.

Chapter 5

Bayes factor and hypotheses testing

Although the Bayes credible interval method can be used to test some hypotheses about the shape-restricted semiparametric ANCOVA model and achieve reasonable results. There are some cases when this method can not be applied. For example, if we have more than one categorical variables in the model, then it is hard to use Bayes credible interval method to test interactions among these categorical variables, since we can not do simultaneous inference. Or, in an even more general case, we want to check if our parallel model itself actually holds, i.e., if the the fits from each treatment level are really parallel to each other. In this chapter, Bayes factor is applied to solve the above problems.

5.1 Bayes factor

The Bayes factor is a practical tool of applied statistics. It offers an easy way of evaluating evidence in favor of each hypothesis. It also allows external information (the prior information) to be considered into the evaluation.

The Bayes factor was first introduced by Jeffreys (1935, 1961) in order to compare two competing scientific theories. In his approach, the Bayes factor was used as a methodology for quantifying the evidence in favor of a theory over the other. Although the Bayes factor is very useful, it has not gained much attention until recently, when high powered computers make computations feasible. Kass and Raftery (1994) reviewed and discussed the uses of Bayes factors in the context of five scientific applications.

For given data $\mathcal{D}$, suppose we consider two mutually exclusive and exhaustive hypotheses H_0 and H_1. Then $Pr(\mathcal{D}|H_0)$ and $Pr(\mathcal{D}|H_1)$ are the probability densities of the data according

to H_0 and H_1, respectively. Given the prior probabilities $\pi(H_0)$ and $\pi(H_1)$, where $\pi(H_1) = 1 - \pi(H_0)$, the posterior probabilities based on the data are $Pr(H_0|\mathcal{D})$ and $Pr(H_1|\mathcal{D}) = 1 - Pr(H_0|\mathcal{D})$. By Bayes' Theorem, we have

$$Pr(H_i|\mathcal{D}) = \frac{Pr(\mathcal{D}|H_i)\pi(H_i)}{Pr(\mathcal{D}|H_0)\pi(H_0) + Pr(\mathcal{D}|H_1)\pi(H_1)}, \quad i = 0, 1$$

Hence,

$$\frac{Pr(H_1|\mathcal{D})}{Pr(H_0|\mathcal{D})} = \frac{Pr(\mathcal{D}|H_1)\pi(H_1)}{Pr(\mathcal{D}|H_0)\pi(H_0)}.$$

The Bayes factor Bf_{10} is defined as

$$Bf_{10} = \frac{Pr(\mathcal{D}|H_1)}{Pr(\mathcal{D}|H_0)} \tag{5.1.1}$$

Since $\frac{Pr(H_1|\mathcal{D})}{Pr(H_0|\mathcal{D})}$ is the posterior odds, and $\frac{\pi(H_1)}{\pi(H_0)}$ is the prior odds, Bayes factor is the ratio of the posterior odds of H_1 to its prior odds. When the prior probability on the null is one-half, Bayes factor is the posterior odds against the null hypothesis.

When there are no free parameters in the two hypotheses H_0 and H_1, the Bayes factor Bf_{10} is simply the likelihood ratio. When unknown parameters $\boldsymbol{\phi}$ are present in the model, the probability densities $Pr(\mathcal{D}|H_i)$, $i = 0, 1$, sometimes called the marginal likelihood, or the integrated likelihood, can be obtained by integrating over the parameter space, such as

$$Pr(\mathcal{D}|H_i) = \int Pr(\mathcal{D}|\boldsymbol{\phi}_i, H_i)\pi(\boldsymbol{\phi}_i|H_i)d\boldsymbol{\phi}_i, \quad i = 0, 1 \tag{5.1.2}$$

where $\boldsymbol{\phi}_i$ is the free parameter vector under H_i, $\pi(\boldsymbol{\phi}_i|H_i)$ is the prior density of $\boldsymbol{\phi}_i$ under H_i, $Pr(\mathcal{D}|\boldsymbol{\phi}_i, H_i)$ is the probability density of the data under H_i and $\boldsymbol{\phi}_i$

In the Bayes factor setting, it's conceptually straightforward to decide between the null hypothesis H_0 and the alternative hypothesis H_1. We merely calculate the posterior probabilities $Pr(H_0|\mathcal{D})$ and $Pr(H_1|\mathcal{D})$, then make decisions based on them.

Bayes Factors are the dominant method of Bayesian model testing. They are the Bayesian analogues of likelihood ratio tests. The basic intuition is that prior and posterior information are combined in a ratio that provides evidence in favor of one model specification versus another.

Bayes Factors are very flexible, allowing multiple hypotheses to be compared simultaneously. Also, it allows easy comparison of non-nested models, and of irregular models. For some statistical models, it might be hard to derive the non-Bayesian significance tests. However, we can calculate Bayes factors instead, since they are conceptually simpler and computationally feasible.

Based on Jefferys' suggestions (1961, Appendix B), Kass and Raftery (1994) interpretated Bayes factor Bf_{10} into four categories (Table 5.1). These categories are not a calibration of the Bayes factor, but rather a rough descriptive statement about standards of evidence in scientific investigation. $2log(Bf_{10})$ is considered in these categories since it has the same score as the likelihood ratio statistics.

Table 5.1: Categories suggested by Kass and Raftery to interpret Bayes Factor (Bf_{10}).

$2log(Bf_{10})$	Bf_{10}	Evidence Against H_0
0 to 2	1 to 3	Not worth more than a bare mention
2 to 5	3 to 12	Positive
5 to 10	12 to 150	Strong
> 10	> 150	Decisive

5.2 Calculating Bayes factors

Unfortunately, although Bayes factors are rather intuitive, as a practical matter they are often quite difficult to calculate if the model is complicated.

In order to get the value of Bayes factor, we have to compute the marginal likelihood. For some simple cases, (5.1.2) might be evaluated analytically. DeGroot (1970) and Zellner (1971) showed the exact analytic evaluation of (5.1.2) for exponential family distribution with conjugate priors. But most often, numerical methods are needed to deal with the integral. In this section, we represent (5.1.2) simply as

$$I = \int P(\mathcal{D}|\boldsymbol{\phi}, H)\pi(\boldsymbol{\phi}|H)d\boldsymbol{\phi}. \tag{5.2.1}$$

5.2.1 Laplace's method

Laplace's method of approximation (De Bruijn, 1970 and Tierney & Kadane, 1986) assumes that the posterior density is highly peaked about its maximum $\tilde{\boldsymbol{\phi}}$, which is the posterior mode. Let $\tilde{l}(\boldsymbol{\phi}) = log\,[P(\mathcal{D}|\boldsymbol{\phi}, H)\pi(\boldsymbol{\phi}|H)]$, Expanding $\tilde{l}(\boldsymbol{\phi})$ as a quadratic about $\tilde{\boldsymbol{\phi}}$ and then exponentiating yields an approximation to $P(\mathcal{D}|\boldsymbol{\phi}, H)\pi(\boldsymbol{\phi}|H)$, which has a Normal density with mean $\tilde{\boldsymbol{\phi}}$ and covariance matrix $\tilde{\Sigma} = \left(-D^2\tilde{l}(\tilde{\boldsymbol{\phi}})\right)^{-1}$, where $D^2\tilde{l}(\tilde{\boldsymbol{\phi}})$ is the Hessian matrix of second derivatives. The approximation through Laplace's method yields

$$I \approx (2\pi)^{m/2}|\tilde{\Sigma}|^{1/2}P(\mathcal{D}|\tilde{\boldsymbol{\phi}}, H)\pi(\tilde{\boldsymbol{\phi}}|H) \equiv \hat{I}_1,$$

where m is the dimension of $\boldsymbol{\phi}$.

In general, this method provides adequate approximations for well-behaved problems of modest dimensionality (Kass and Raftery, 1994). Slate (1994) studied various parameterizations of exponential families and gave a detailed discussion of sample size required to obtain posterior normality, in which case the accuracy of Laplace's method is guaranteed. Jeffreys (1961), Lindley (1961), Leonard (1982), etc have applied this method in their papers.

We do not use this method because of the high dimension of our parameter space.

5.2.2 Simple Monte Carlo, importance sampling, and Gaussian quadrature

Simple Monte Carlo integration has been described by Raftery and Banfield (1990) among others. The simplest Monte Carlo integration estimate of (5.1.2) is

$$I \approx \frac{1}{m}\sum_{i=1}^{m} P(\mathcal{D}|\boldsymbol{\phi}^{(i)}, H) \equiv \hat{I}_1, \tag{5.2.2}$$

where $\boldsymbol{\phi}^{(i)}, i = 1, \cdots m$ is a sample from the prior distribution.

McCulloch and Rossi (1991) studied some particular cases and found the major difficulty with $\hat{I}_2$ is that most of the $\boldsymbol{\phi}^{(i)}$ have small likelihood values if the posterior is concentrated relative to the prior, so that the simulation process will be very inefficient. Thus the estimate is dominated by a few large values of the likelihood, and so the variance of $\hat{I}_2$ is large and its convergence to a Gaussian distribution is slow.

The precision of simple Monte Carlo integration can be improved by importance sampling. Geweke (1989) gave a general discussion of importance sampling, which consists of generating a sample $\boldsymbol{\phi}^{(i)}, i = 1, \cdots, m$ from an importance sampling function $\pi^*(\boldsymbol{\phi})$. Under general conditions, an estimate of I is

$$\hat{I}_3 = \frac{\sum_{i=1}^{m} w_i P(\mathcal{D}|\boldsymbol{\phi}^{(i)}, H)}{\sum_{i=1}^{m} w_i}, \tag{5.2.3}$$

where

$$w_i = \frac{\pi(\boldsymbol{\phi}^{(i)})}{\pi^*(\boldsymbol{\phi}^{(i)})} \tag{5.2.4}$$

A scheme based on adaptive Gaussian quadrature (Genz and Kass, 1993) is effective when the dimension of the parameter space is modest, roughly speaking, less than 15.

5.2.3 Simulating from the posterior

Several methods are available for simulating from the posterior distributions. The methods of direct simulation and rejection sampling are used for simple cases; Markov chain Monte Carlo methods, particularly the Gibbs sampler, are applied for complex cases. Newton and Raftery (1994) also presented the weighted likelihood bootstrap method.

For each of these methods, we can get a sample approximately drawn from the posterior density. Let the importance sampling function be

$$\pi^*(\boldsymbol{\phi}) = P(\boldsymbol{\phi}|\mathcal{D}, H) = \frac{P(\mathcal{D}|\boldsymbol{\phi}, H)\pi(\boldsymbol{\phi})}{P(\mathcal{D}|H)}. \tag{5.2.5}$$

Substituting (5.2.5) into (5.2.4) and (5.2.3) yields an estimate of $P(\mathcal{D}|H)$ (Newton and Raftery, 1994),

$$\hat{I}_4 = \left\{ \frac{1}{m} \sum_{i=1}^{m} P(\mathcal{D}|\boldsymbol{\phi}^{(i)}, H)^{-1} \right\}^{-1} \tag{5.2.6}$$

This converges almost surely to the correct value $P(\mathcal{D}|H)$ as $m \to \infty$. And it is very stable when outliers present. This method is very easy to calculate and often gives results that are accurate enough for interpretation on the logarithmic scale (Rosenkranz, 1992; Carlin and Chib, 1993; Raftery, 1994).

We apply the method of simulating from the posterior to calculate Bayes factors, since in our shape-restricted models, there are so many free parameters, which, in addition, have different prior densities. Simulations are utilized by Gibbs sampler and latent variable techniques.

5.3 Simulation study

We consider the semiparametric ANCOVA model with a k-level categorical variable,

$$y_i = f(x_i) + \beta_1 d_{1i} + \cdots + \beta_{k-1} d_{(k-1)i} + \sigma\varepsilon_i,$$

where $i = 1, \cdots, n$, f is any function from the shape-restricted family, ε_i's are independent random standard normal errors, β_1, $\cdots$, β_{k-1} and σ are unknown constants, d_{ji}'s are the indicator variables,

$$d_{ji} = \begin{cases} 1 & \text{if the } i^{th} \text{ observation is from group } j,\ j = 1, \cdots, k-1 \\ 0 & \text{otherwise.} \end{cases}$$

We are interested in the following hypotheses:

$$H_0\text{: there is no significant treatment effects, i.e., } \beta_1 = \cdots = \beta_{k-1} = 0$$

$$H_1\text{: at least one treatment level has significant effect.}$$

When we apply the Bayesian method, one important question is how to choose γ. Recall that γ is the scale parameter of the Gamma density, from which we generate the coefficients of the hinge vectors. The γ acts as a smoothing parameter; different γ might produce different fit. A very small γ tends to overfit the data (reproduce MLE) and a very large γ tends to underfit the data.

We choose a suitable γ by the method of ordinary cross-validation (see section 2.2.2):

(1) Generate a random sample with fixed sample size from the chosen underlying function.

(2) Fit a desired Bayes shape-restricted ANCOVA regression to the data set with a particular γ.

(3) Calculate the value of the cross-validation function $V(\gamma)$.

(4) Try a range of γ values, find the optimal γ that minimizes $V(\gamma)$.

(5) Apply this optimal γ in the simulation study, for this given underlying function, sample size and shape restriction.

We compare Bayes factor method with the traditional F-test (from parametric ANCOVA regression) and the F^s-test (from kernel smoothing ANCOVA regression, see section 2.2.1) through simulation studies. The smoothing parameter h in the kernel smoothing is also chosen by the method of ordinary cross-validation.

A total of 2000 samples are generated from each of the log and convex underlying functions, which have the form of $f(x_i) = 5log(x_i + 1)$ and $f(x_i) = 2x_i + 1/x_i$, respectively. Balanced designs with $k = 3$ treatment levels are studied, i.e., there are 3 observations at any given x and each of them corresponds to one treatment level.

For the log underlying function, we fit Bayes increasing concave regression, simple linear regression and kernel smoothing regression. For the convex underlying function, we fit Bayes convex regression, quadratic regression and kernel smoothing regression. Sample size was chosen to be 30.

We generated samples under both the null hypothesis and the alternative hypothesis ($\beta_2 = 1$ and $\beta_3 = 2$). Table 5.2 shows the results when data were generated under the null hypothesis. Table 5.3 shows the results when data were generated under the alternative hypothesis. The categories of Bf_{10} are according to Table 5.1.

Table 5.2: Data generated under null hypothesis

Underlying Function	$2log(Bf_{10})$				F^s-test P-value		F-test P-value	
	< 2	2 to 5	5 to 10	> 10	< 0.05	> 0.05	< 0.05	> 0.05
log	99.15%	0.65%	0.15%	0.05%	7.62%	92.38%	3.30%	96.70%
convex	85.40%	11.60%	2.70%	0.30%	8.60%	91.40%	4.30%	95.70%

Table 5.3: Data generated under alternative hypothesis

Underlying Function	$2log(Bf_{10})$				F^s-test P-value		F-test P-value	
	< 2	2 to 5	5 to 10	> 10	< 0.05	> 0.05	< 0.05	> 0.05
log	1.80%	11.95%	30.55%	55.70%	96.80%	3.20%	95.58%	4.42%
convex	1.30%	11.40%	47.40%	39.90%	90.88%	9.12%	96.90%	3.10%

The value $2log(Bf_{10}) = 5$ can be considered as a rough criterion to draw conclusions about the hypotheses tests. The simulation results show that Bayes factor method gives reasonable suggestions about the hypotheses, and it is less frequent to reject H_0 comparing with F^s-test and F-test. Actually, there is no reason to expect a P-value to be similar to the posterior probability that the null hypothesis is correct. There is a general feeling that Bayes factors are more conservative than P-values, mainly because when comparisons are made it becomes clear that a P-value of 0.05 can not represent much evidence against the null (see Edwards, Lindman and Savage, 1963; Berger and Mortera, 1991).

5.4 Applying Bayes factor in hypotheses testing

When Bayes factors are applied, the ideas of hypothesis testing about shape-restricted models are straightforward. Let's consider a simple case: suppose we have a semiparametric ANCOVA model with two categorical variables A and B. Treatment A has two levels and B has three levels, written as A_1, A_2, B_1, B_2, B_3, respectively. We can define three indicator variables according to them, such as,

$$d_{1i} = \begin{cases} 1 & \text{if the } i^{th} \text{ observation is in treatment level } A_1, \\ 0 & \text{otherwise.} \end{cases}$$

$$d_{2i} = \begin{cases} 1 & \text{if the } i^{th} \text{ observation is in treatment level } B_1, \\ 0 & \text{otherwise.} \end{cases}$$

$$d_{3i} = \begin{cases} 1 & \text{if the } i^{th} \text{ observation is in treatment level } B_2, \\ 0 & \text{otherwise.} \end{cases}$$

Then the model can be written as

$$y_i = f(x_i) + \beta_1 d_{1i} + \beta_2 d_{2i} + \beta_3 d_{3i} + \beta_4 d_{1i} d_{2i} + \beta_5 d_{1i} d_{3i} + \varepsilon_i,$$

where $i = 1, \cdots, n$, f is any function from the shape-restricted family, ε_i's are independent random standard normal errors, β_1, $\cdots$, β_5 and σ are unknown constants.

From the above model, the hypotheses about the categorical variables can be summarized as follows,

Hypotheses testing	simple expression
H_0: treatment A has no effect	$\beta_1 = 0$
H_0: treatment B has no effect	$\beta_2 = \beta_3 = 0$
H_0: no interactions between A and B	$\beta_4 = \beta_5 = 0$

For each of the hypotheses, we need to compute $P(\mathcal{D}|H_0)$ and $P(\mathcal{D}|H_1)$ from their corresponding posterior densities, then use Bayes factor Bf_{10} as a criteria to see if the data is in favor of the null hypothesis or against it.

We can also test if our assumption of parallel model holds by fitting six individual shape-restricted curves, and six parallel shape-restricted curves, then computing $P(\mathcal{D}|H_0)$ and $P(\mathcal{D}|H_1)$, finally obtaining Bayes factor Bf_{10} and drawing conclusions.

The Bayes factor itself does not consider the difference in dimensions of the parameter spaces under H_0 and H_1. However, it won't cause any trouble in our studies, since this difference is rather small compared with the high dimensions of our parameter spaces under both hypotheses. Also, in our studies, we have no preference about either hypothesis, i.e., the ratio of $\pi(H_1)$ to $\pi(H_0)$ is 1. Hence, the value of the Bayes factor equals the ratio of $Pr(H_1|\mathcal{D})$ to $Pr(H_0|\mathcal{D})$ and we can use the categories in Table 5.1 to draw conclusions about the hypotheses tests.

Let's now apply the Bayes factor method to explore some real data. The following sections consider three examples: senic data, mouthwash data, feet data.

5.5 Senic example revisited

In chapter 3, we have examined the senic data (see Appendix B) as an example of fitting an unbalanced shape-restricted ANCOVA model. The variables in the senic data set are census (average number of patients in hospital per day during study period), patients' infection risk (average estimated probability of acquiring infection in hospital) and geographic region (1=NE, 2=NC, 3=S, 4=W).

In this data set, our main interest is to see if the patients' infection risks vary across different geographic regions. Since there are four regions, we choose region 4 as the baseline and set three indicator variables d_1, d_2 and d_3 according to each of region 1, 2 and 3, i.e.,

$$d_{ij} = \begin{cases} 1 & \text{if the } j^{th} \text{ hospital is in region } i, \\ 0 & \text{otherwise,} \end{cases}$$

where $i = 1, 2, 3$ and $j = 1, \cdots, n_T$, n_T is the total sample size.

We assume that the relationship between infection risk and census is increasing concave, since the patients might have higher infection risk when there are more other patients around, and this increasing rate might slow down as census increases.

This semiparametric ANCOVA model can be written as

$$y_i = f(t_i) + \beta_1 d_{1i} + \beta_2 d_{2i} + \beta_3 d_{3i} + \sigma\epsilon_i$$

where t refers to census, y refers to the infection risk, β_1, β_2, β_3, σ are unknown constants, ϵ_i's are random standard normal errors, $i = 1, \cdots, n_T$.

We want to test the null hypothesis that the patients' infection risks are the same across the four geographic regions, i.e.,

$$H_0\text{: } \beta_1 = \beta_2 = \beta_3$$

$$H_1\text{: not } H_0$$

We arrange the data as follows:

(1) Let n_T denotes the total sample size, let n denotes number of non-duplicated t, let k refers to number of regions, here $n_T = 113$, $n = 98$ and $k = 4$.

(2) Define x as the distinct values of t and let w_i be the number of observations at each distinct value of x_i.

(3) Sort the data by x, then by region, from smallest to largest.

(4) Define z_{ij} as the j^{th} response at x_i, $j = 1, \cdots, w_i$.

(5) Define $d_{s(ij)}$ as an indicator to identify if z_{ij} is from region s, i.e.,

$$d_{s(ij)} = \begin{cases} 1 & \text{if in } z_{ij} \text{ is from region } s, \\ 0 & \text{otherwise,} \end{cases}$$

where $s = 1, 2, 3$.

5.5.1 Likelihood function, prior distribution, posterior distribution, and full conditional distributions

Let $\theta_i = f(x_i) = \sum_{j=1}^{n-1} b_j \delta_i^j + c$, $\tau = \sigma^{-2}$, the likelihood L for $\boldsymbol{\phi}$, is proportional to

$$L \quad \propto \quad \tau^{n_T/2} \prod_{i=1}^{n} \exp\left\{ -\frac{\tau}{2} \left[\sum_{j=1}^{w_i} (z_{ij} - \beta_1 d_{1(ij)} - \beta_2 d_{2(ij)} - \beta_3 d_{3(ij)} - \theta_i)^2 \right] \right\} \tag{5.5.1}$$

We apply Gamma(α_j, γ) prior on the edge coefficients $b_j, j = 1, \cdots, n-1$, normal $N(0, M)$ prior on the linear coefficients c, $N(0, W_j)$ prior on $\beta_j, j = 1, \cdots, k-1$, and Gamma (g_1, g_2) prior on τ, all independent, then the joint prior density π is proportional to

$$\prod_{j=1}^{n-1} b_j^{\alpha_j - 1} \exp(-b_j \gamma) \exp\left\{ \frac{c^2}{2M} \right\} \exp\left\{ -\sum_{j=1}^{k-1} \frac{\beta_j^2}{2W_j} \right\} \tau^{g_1 - 1} \exp\{-g_2 \tau\}.$$

Let $[a|b]$ denote the conditional distribution of a given b. The conditional distributions of the model parameters τ, b_j, $j = 1, \cdots, n-1$, c, and $\beta_j, j = 1, \cdots, k-1$, are as follows:

1. $[b_s \mid c, b_j, j \neq s, \tau, \beta_j, j = 1, \cdots, k-1]$ is from a distribution with a density proportional to

$$b_s^{\alpha_{s-1}} \exp\left\{ -\frac{\tau}{2} \sum_{i=1}^{n} \sum_{j=1}^{w_i} (\delta_i^s)^2 \, (b_s - r)^2 \right\}$$

where

$$r = \frac{\sum_{i=1}^{n} \delta_i^s \sum_{j=1}^{w_i} \left[z_{ij} - \beta_1 d_{1(ij)} - \beta_2 d_{2(ij)} - \beta_3 d_{3(ij)} - c - \sum_{j \neq s} b_j \right] - \gamma/\tau}{\sum_{i=1}^{n} w_i (\delta_i^s)^2}$$

2. $[c \mid b_j, j = 1, \cdots, n-1, \tau, \beta_j, j = 1, \cdots, k-1]$ has a normal distribution with

$$\text{mean} = \frac{\tau \sum_{i=1}^{n} \sum_{j=1}^{w_i} \left[z_{ij} - \beta_1 d_{1(ij)} - \beta_2 d_{2(ij)} - \beta_3 d_{3(ij)} - \sum_{l=1}^{n-1} b_l \delta_i^l \right]}{1/M + \tau n_T},$$

and

$$\text{variance} = \frac{1}{1/M + \tau n_T}$$

3. $[\tau \mid b_j, j = 1, \cdots, n-1, c, \beta_j, j = 1, \cdots, k-1]$ has a Gamma distribution with parameters $(n_T/2 + g_1, \text{sse}/2 + g_2)$, the density is proportional to

$$\tau^{(n^T/2 + g_1 - 1)} \exp\left\{-\tau(sse/2 + g_2)\right\},$$

where

$$\text{sse} = \sum_{i=1}^{n} \sum_{j=1}^{w_i} \left[z_{ij} - \beta_1 d_{1(ij)} - \beta_2 d_{2(ij)} - \beta_3 d_{3(ij)} - \sum_{l=1}^{n-1} b_l \delta_i^l - c)^2 \right]$$

4. $[\beta_1 \mid \beta_2, \beta_3, b_j, j = 1, \cdots, n-1, c, \tau]$ has a normal distribution with

$$\text{mean} = \frac{\tau \sum_{i=1}^{n} \sum_{j=1}^{w_i} d_{1(ij)} \left[z_{ij} - \beta_2 d_{2(ij)} - \beta_3 d_{3(ij)} - \sum_{l=1}^{n-1} b_l \delta_i^l - c \right]}{1/W_1 + \tau \sum_{i=1}^{n} \sum_{j=1}^{w_i} d_{1(ij)}^2},$$

and

$$\text{variance} = \frac{1}{1/W_1 + \tau \sum_{i=1}^{n} \sum_{j=1}^{w_i} d_{1(ij)}^2}.$$

5. $[\beta_2 \mid \beta_1, \beta_3, b_j, j = 1, \cdots, n-1, c, \tau]$ has a normal distribution with

$$\text{mean} = \frac{\tau \sum_{i=1}^{n} \sum_{j=1}^{w_i} d_{2(ij)} \left[z_{ij} - \beta_1 d_{1(ij)} - \beta_3 d_{3(ij)} - \sum_{l=1}^{n-1} b_l \delta_i^l - c \right]}{1/W_2 + \tau \sum_{i=1}^{n} \sum_{j=1}^{w_i} d_{2(ij)}^2},$$

and

$$\text{variance} = \frac{1}{1/W_2 + \tau \sum_{i=1}^{n} \sum_{j=1}^{w_i} d_{2(ij)}^2}.$$

6. $[\beta_3 \mid \beta_1, \beta_2, b_j, j = 1, \cdots, n-1, c, \tau]$ has a normal distribution with

$$\text{mean} = \frac{\tau \sum_{i=1}^{n} \sum_{j=1}^{w_i} d_{3(ij)} \left[z_{ij} - \beta_1 d_{1(ij)} - \beta_2 d_{2(ij)} - \sum_{l=1}^{n-1} b_l \delta_i^l - c \right]}{1/W_3 + \tau \sum_{i=1}^{n} \sum_{j=1}^{w_i} d_{3(ij)}^2},$$

and

$$\text{variance} = \frac{1}{1/W_3 + \tau \sum_{i=1}^{n} \sum_{j=1}^{w_i} d_{3(ij)}^2}.$$

Since this data set contains 113 observations, it is hard to get the best γ according to the extensive computation time it might require. Instead, we try 3 different γ values, 0.01, 0.1 and 1. The results of the parameter estimates and the hypotheses tests show that our posterior probabilities are insensitive to γ.

Table 5.4: Senic data: parameter estimates from Bayes parallel increasing concave regression.

γ	Parameters	2.5% percentile	50% percentile	97.5% percentile	mean	SD
0.01	β_1	-0.8665	-0.1385	0.6061	-0.1337	0.2923
	β_2	-0.6672	-0.5331	-0.3697	-0.5219	0.3831
	β_3	-1.0536	-0.9007	-0.7314	-0.8938	0.3048
	τ	0.6527	0.7583	0.8823	0.7600	0.0583
	c	4.3351	4.7618	5.2275	4.7655	0.2225
0.1	β_1	-0.8326	-0.1398	0.6244	-0.1196	0.2914
	β_2	-0.6372	-0.5349	-0.3581	-0.5137	0.3799
	β_3	-1.0348	-0.9246	-0.7272	-0.8951	0.3039
	τ	0.6568	0.7589	0.8867	0.7618	0.0587
	c	4.3367	4.7630	5.2257	4.7725	0.2218
1	β_1	-0.8208	-0.1101	0.6518	-0.1021	0.2908
	β_2	-0.6406	-0.5226	-0.3286	-0.5058	0.3779
	β_3	-1.0417	-0.9068	-0.7308	-0.8974	0.3044
	τ	0.6623	0.7648	0.8902	0.7682	0.0591
	c	4.3509	4.7835	5.2334	4.7848	0.2217

Table 5.4 lists the parameter estimates from Bayes increasing concave regression. For each parameter, 2.5% percentile, 50% percentile, 97.5% percentile, mean and standard deviation (SD) are calculated. The $\beta_1, \beta_2, \beta_3$ are the coefficients of the indicator variables according to region 1, 2 and 3; c is the coefficient of the $\boldsymbol{v}^1$ vector, where $\boldsymbol{v}^1 = (1, \cdots, 1)'$; $\tau = 1/\sigma^2$.

The results show that region 4 has the highest infection risk, region 1 has the second highest infection risk, region 3 has the lowest infection risk.

Figure 5.1 shows the fitted Bayes parallel increasing concave regressions, with $\gamma = 0.01$ (a), $\gamma = 0.1$ (b), $\gamma = 1$ (c). The x axis is the census and the y axis is the patients' infection risk. The fits are very similar for different γ.

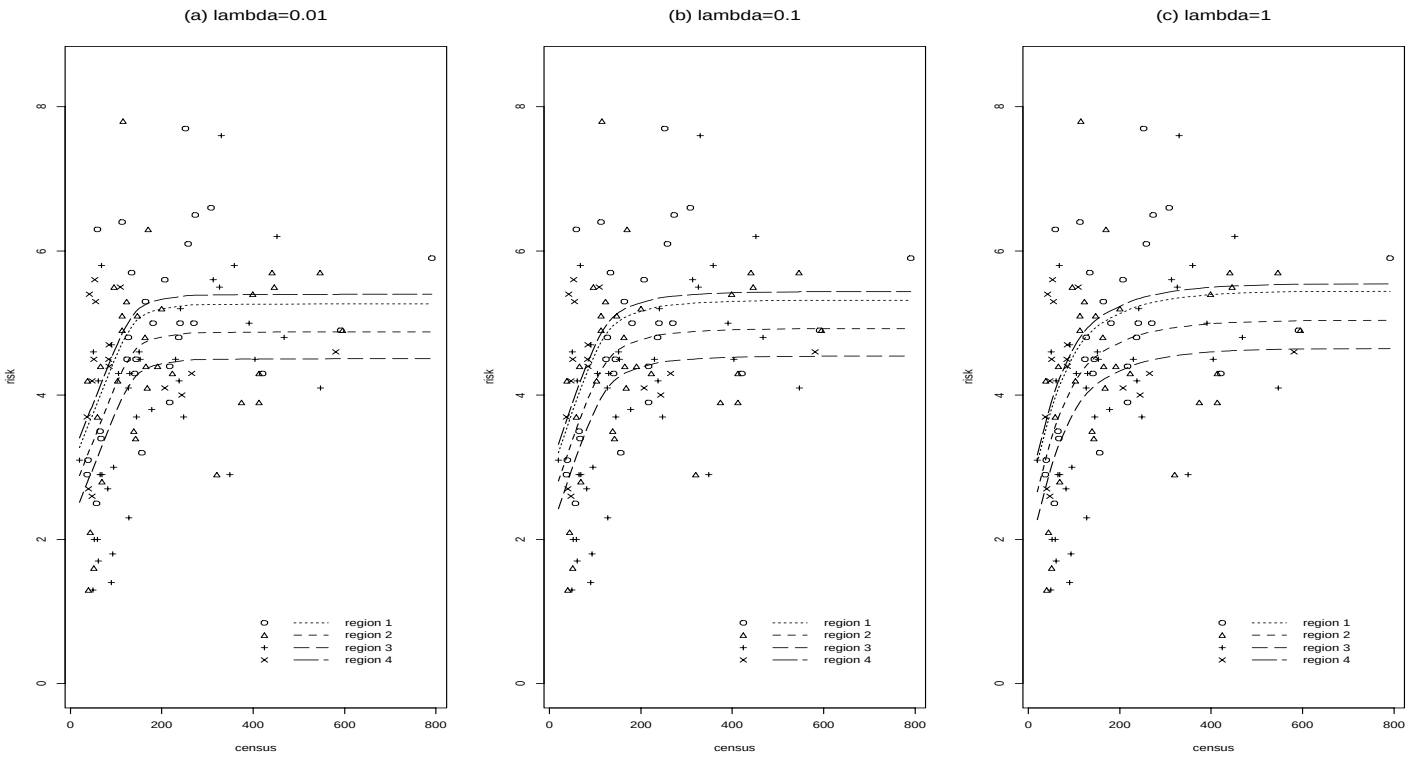

Figure 5.1: Senic data: Bayes parallel increasing concave fit, one categorical variable *region* is included in the model, γ is chosen to be 0.01 (a), 0.1 (b), 1 (c).

Since the variable *region* has 4 levels, we can go further to test the contrasts among these levels. Table 5.5 shows the results of contrasts by Bayes factor. Also Bayes credible interval method is applied whenever possible. These two methods give results that agree. The null hypothesis that there are no region effects is rejected with $2log(Bf_{10}) = 7.4012$ for $\gamma = 0.01$ and $2log(Bf_{10}) = 7.5444$ for $\gamma = 1$. Also, the patients' infection risks are significantly different in regions 3 and 4, with $2log(Bf_{10}) = 10.0390$ for $\gamma = 0.01$ and $2log(Bf_{10}) = 10.0587$ for $\gamma = 1$. The patients' infection risks in region 1 and 4 are also significantly different.

Table 5.5: Senic data: hypotheses tests about categorical variable *region*.

Null Hypotheses	Indicator Expression	γ	Bf_{10}	$2log(Bf_{10})$	Credible Interval
No region difference	$\beta_1 = \beta_2 = \beta_3 = 0$	0.01	40.4706	7.4012	NA
		1	43.4766	7.5444	NA
region 1 and 2	$\beta_1 = \beta_2$	0.01	1.8827	1.2654	Accept
		1	2.0759	1.4608	Accept
region 1 and 3	$\beta_1 = \beta_3$	0.01	24.9240	6.4317	Reject
		1	30.9096	6.8621	Reject
region 1 and 4	$\beta_1 = 0$	0.01	0.7489	-0.5782	Accept
		1	0.8382	-0.3531	Accept
region 2 and 3	$\beta_2 = \beta_3$	0.01	9.7876	4.5622	Accept
		1	11.3387	4.8565	Accept
region 2 and 4	$\beta_2 = 0$	0.01	4.3040	2.9190	Accept
		1	4.5888	3.0473	Accept
region 3 and 4	$\beta_3 = 0$	0.01	151.3331	10.0390	Reject
		1	152.8318	10.0587	Reject

Some people consider the medical school affiliation as another important factor that would affect the patients' infection risk in a hospital. Since one would feel that if there exists a medical school affiliation, then this hospital might have a higher medical standard and the patients' infection risk from this hospital might be lower than those hospitals which do not have a medical school affiliation. A variable *medsch* indicates whether there is a medical school affiliation ($medsch = 1$) or not ($medsch = 0$). After adding this variable, our model now is

$$y_i = f(t_i) + \beta_1 d_{1i} + \beta_2 d_{2i} + \beta_3 d_{3i} + \beta_4 d_{4i} + \beta_5 d_{1i} d_{4i} + \beta_6 d_{2i} d_{4i} + \beta_7 d_{3i} d_{4i} + \sigma\epsilon_i$$

we set a new indicator d_{4i} as

$$d_{4i} = \begin{cases} 1 & \text{if the } i^{th} \text{ hospital has a medical school affiliation,} \\ 0 & \text{otherwise.} \end{cases}$$

β_4, β_5, β_6, β_7 are unknown parameters. All other parameters are defined as before.

We call it the full model and set the case $region = 4$ and $medsch = 0$ as baseline. Figure 5.3 shows the fit for baseline from the above full model based on Bayes increasing concave regression, with γ 0.01 (a), 0.1 (b), 1 (c). Again, the fits are very similar for different γ.

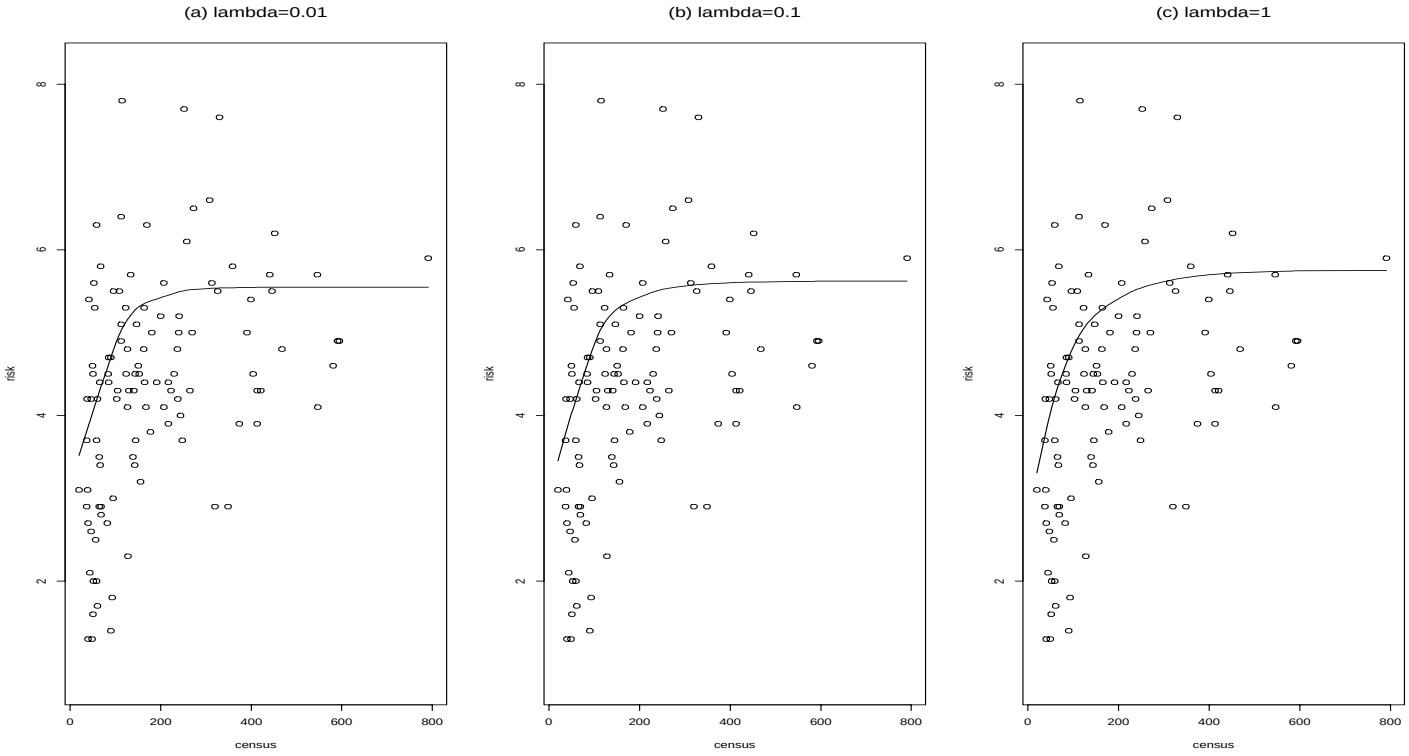

Figure 5.2: Senic data: Bayes increasing concave fit, two categorical variables $region$ and $medsch$ are included in the model. The solid line is for baseline ($region = 4$ and $medsch = 0$). The γ is chosen to be 0.01 (a), 0.1 (b), 1 (c).

Table 5.6 lists the parameter estimates from Bayes parallel increasing concave regressions, when two categorical variables $region$ and $medsch$ are included in the model.

When there are interaction terms in a model, we need to perform the hypotheses tests about these interaction terms first. The null hypothesis we are testing now is that there is no significant interactions between region and medical school affiliation. Table 5.7 shows the results of the hypotheses test. All the values of Bayes factor conclude that the null hypothesis is strongly rejected. Hence, we can not go further to test the marginal effects of geographic region and medical school affiliation.

Table 5.6: Senic data: parameter estimates from Bayes increasing concave regression, with two categorical variables *region* and *medsch*.

γ	Parameters	2.5% percentile	50% percentile	97.5% percentile	Mean	SD
0.01	β_1	-1.1156	-0.3342	0.4676	-0.3386	0.4139
	β_2	-1.3589	-0.5875	0.2096	-0.5911	0.4025
	β_3	-1.9134	-1.1710	-0.4323	-1.1644	0.3813
	β_4	-3.0186	-1.2093	0.4465	-1.2353	0.8984
	β_5	-0.2519	1.7117	3.7239	1.7044	1.0265
	β_6	-1.0200	0.8928	2.8739	0.9082	0.9987
	β_7	0.6917	2.7047	4.9256	2.7255	1.0928
	τ	0.6664	0.7870	0.9159	0.7878	0.0631
	c	4.3771	5.0444	5.6737	5.0382	0.3311
0.1	β_1	-1.1362	-0.3481	0.4356	-0.3457	0.3975
	β_2	-1.3697	-0.6012	0.1682	-0.5980	0.3946
	β_3	-1.9013	-1.1763	-0.4497	-1.1727	0.3688
	β_4	-3.0413	-1.2694	0.4291	-1.2781	0.8963
	β_5	-0.2663	1.7283	3.7570	1.7289	1.0282
	β_6	-0.9982	0.8879	2.8489	0.8998	0.9860
	β_7	0.6645	2.6998	4.8981	2.7243	1.0882
	τ	0.6661	0.7919	0.9200	0.7913	0.0644
	c	4.4339	5.0481	5.6955	5.0502	0.3220
1	β_1	-1.1046	-0.3381	0.4536	-0.3365	0.4048
	β_2	-1.2895	-0.5777	0.2075	-0.5709	0.3861
	β_3	-1.9319	-1.1931	-0.4451	-1.1838	0.3780
	β_4	-3.1023	-1.3112	0.2998	-1.3491	0.8897
	β_5	-0.1819	1.7258	3.7902	1.7397	1.0185
	β_6	-1.0411	0.8130	2.7269	0.8432	0.9756
	β_7	0.7313	2.7044	4.8981	2.7257	1.0795
	τ	0.6732	0.7992	0.9265	0.7992	0.0643
	c	4.4085	5.0641	5.7020	5.0608	0.3272

Table 5.7: Senic data: hypotheses test for interactions between region and medical school affiliation.

γ	Bf_{10}	$2log(Bf_{10})$
0.01	24.1563	6.3691
0.1	25.9058	6.5089
1	19.7773	5.9691

Figure 5.3 shows the *medsch* by *region* profile plot of the mean infection risk at average census level. From the plot, it seems that the medical school affiliation results in higher infection risks in regions 1, 2 and 3, with the biggest gap happens in region 3. In region 4, the existence of the medical school affiliation does not have much effect on the infection risk. The results are different from what we get before, when only one variable *region* is included in the model, and it is due to the significant interactions between these two categorical variables.

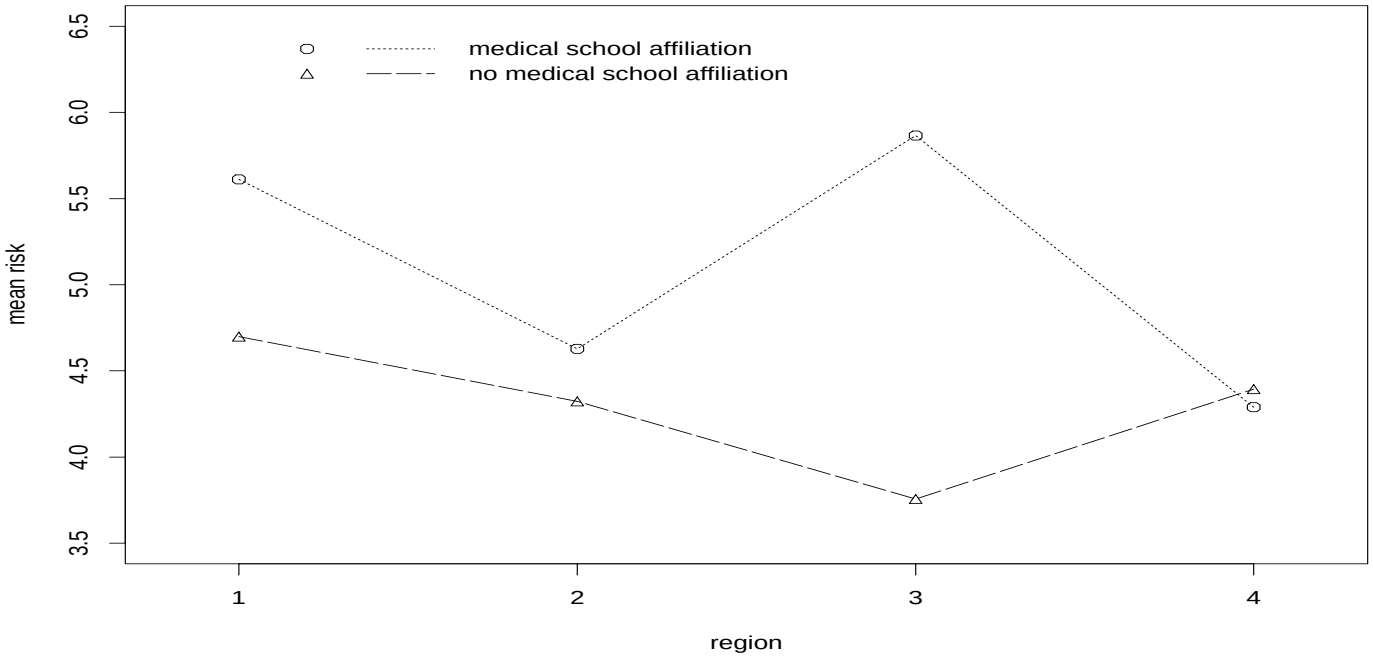

Figure 5.3: Senic data: *medsch* by *region* interaction plot, Bayes increasing concave fit, $\gamma = 0.1$.

5.6 Mouthwash example

The mouthwash data are from an experiment (Platt, 1985) to determine whether a regular program of mouthwash with a common brand of analgesic is effective in treating a type of gum disease. A total of 30 volunteers with varying degrees of gum disease were randomized

to two groups of 15 each, a control group using only a water rinse for mouthwash and an experimental treatment group. Baseline measurements were taken followed by weekly measurements over the three weeks of the study. Table 5.8 contains the baseline and week-three data on one of the variables, SBI, a measurement indicating gum shrinkage for which lower numbers indicate better health. Variable *Group* indicates whether the subject receives control ($Group = 0$) or treatment ($Group = 1$).

Table 5.8: Mouthwash data

Week 3 SBI	Baseline SBI	Group
0.39	0.25	1
0.19	0.25	0
0.30	0.30	1
0.15	0.33	0
0.14	0.34	0
0.15	0.38	1
0.17	0.40	0
0.19	0.45	1
0.18	0.46	1
0.33	0.48	0
0.29	0.54	0
0.45	0.55	1
0.41	0.57	1
0.29	0.59	1
0.09	0.60	0
0.65	0.63	1
0.45	0.63	1
0.15	0.63	1
0.44	0.65	1
0.51	0.65	0
0.50	0.66	1
0.18	0.69	1
0.54	0.71	0
0.47	0.71	0
0.51	0.75	1
0.42	0.99	0
0.42	0.99	0
0.69	1.32	0
0.57	1.42	0
0.31	1.72	0

The initial hypothesis was that the treatment group would show greater improvement than the control group, where the improvement is measured by difference between the baseline and week-three SBI ($\Delta = SBI_{baseline} - SBI_{week3}$). The easiest way to compare the differences in the two groups is a simple two-sample t-test, using Δ as the response variable. The t statistic is -2.08 on 28 degree of freedom, with a P-value 0.0468. This actually suggests that control group ($\bar{\Delta} = 0.4107$) gains greater improvement than the treatment group ($\bar{\Delta} = 0.1967$).

If we examine the data carefully, we might notice that there are 4 subjects from the control group with very high baseline SBI values. Hence, a more appropriate analysis should account for the baseline measurements as a covariate.

The model can be written as

$$y_i = f(x_i) + \beta d_i + \sigma\varepsilon_i,$$

where $i = 1, \cdots, n$, f is a function of x_i, ε_i's are independent random standard normal errors, σ and β are unknown constants, d is an indicator variable, where

$$d_i = \begin{cases} 1 & \text{if the } i^{th} \text{ subject is in the treatment group,} \\ 0 & \text{otherwise.} \end{cases}$$

Speckman (1988) used a kernel smoothing technique to analyze this data set, and the estimated treatment effect was $\hat{\beta} = 0.040$ with $F = 0.59$ on 1 and 24.7 degrees of freedom. We follow Speckman's method, but use a different kernel smoothing matrix as (3.2.1) , the estimated treatment effect we get is $\hat{\beta} = 0.045$ with F=0.65 on 1 and 24.9 degree of freedom, slightly different from Speckman's. Figure 5.4 shows our fitted parallel kernel smoothing splines. The response variable is week-three SBI, and the covariate is baseline SBI. The approximate F-test (F^s-test) shows no significant difference between control and treatment groups.

Although the kernel smoothing fits the data quite well, the shape of the curve is questionable. There is no reason to believe that the positive association between baseline SBI and

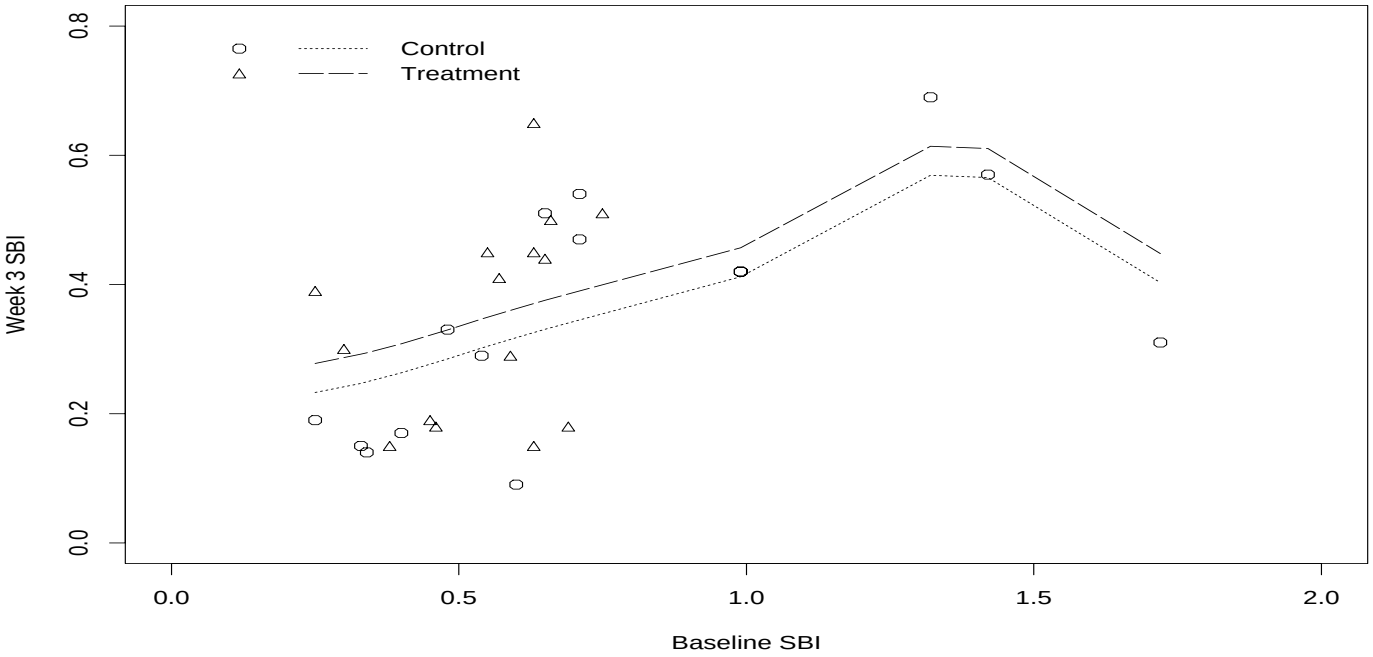

Figure 5.4: Mouthwash data: kernel smoothing fit.

week-three SBI will change its direction when baseline SBI reaches to a fixed point. Intuitively, week-three SBI and baseline SBI should be positively related. Thus, an increasing concave regression may be a better choice in this case.

We use the mouthwash data to illustrate the way of choosing suitable γ value, where γ is the scale parameter of a Gamma density from which we generate the coefficients of the edge vectors. The Bayes parallel increasing concave regression method is applied to the data set. Figure 5.5 shows the plot of the ordinary cross-validation function $V(\gamma)$ versus γ. The function $V(\gamma)$ has its minimum around $\gamma = 2.0$, but there are only tiny changes of $V(\gamma)$ as γ varies from 0 to 4, which means that our posterior probability is insensitive to the values of γ. Because the method of ordinary cross-validation might take lots of computing time, in addition that the Gibbs sampler itself is time-consuming, it is hard to apply this method on

a large data set. However, this insensitivity of the posterior probability to γ allows us to use a moderate γ instead of the "best" γ (in the sense of minimizing $V(\gamma)$) for a large data set.

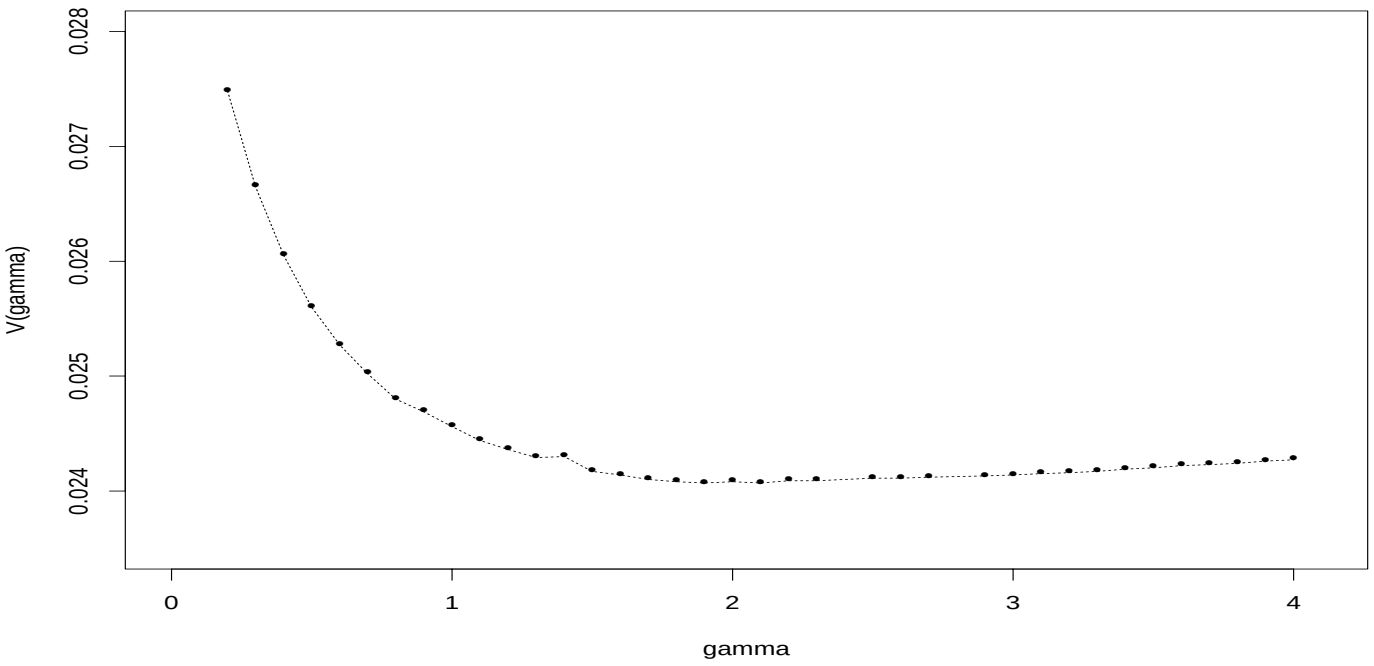

Figure 5.5: Mouthwash data: cv plot.

Table 5.9 lists the parameter estimates from Bayes increasing concave regression, with γ=2.0. For each parameter, 2.5% percentile, 50% percentile, 97.5% percentile, mean and standard deviation (SD) are calculated. Here, β is the coefficient of the indicator variable d; c is the coefficient of the $\boldsymbol{v}^1$ vector, where $\boldsymbol{v}^1 = (1, \cdots, 1)'$; and $\tau = 1/\sigma^2$. The mean and median of $\hat{\beta}$ is 0.0117 and 0.0127, respectively; note that it is less than that of kernel smoothing ($\hat{\beta} = 0.045$). Again, the control group has a slightly lower week 3 SBI measurement than the treatment group, and we want to know if this difference is statistically significant.

Figure 5.6(a) shows the Bayes parallel increasing concave fits to the mouthwash data, where the dotted line is the fit for the control group and the dashed line is the fit for the treatment group.

Table 5.9: Mouthwash data: parameter estimates from Bayes parallel increasing concave regression.

Parameters	2.5% percentile	50% percentile	97.5% percentile	Mean	SD
β	-0.2012	0.0127	0.2232	0.0117	0.1082
τ	7.0796	11.5127	17.0978	11.6110	2.5435
c	0.1921	0.3435	0.4915	0.3418	0.0762

In addition, we want to test if our assumption of parallel increasing concave model holds, i.e, if there is significant interaction between the categorical variable and the covariate. In this case, the two hypotheses we are comparing are

$$H_0\text{: } y_i = f(x_i) + \beta d_i + \varepsilon_i,$$
$$H_1\text{: } y_i = f_j(x_i) + \varepsilon_i$$

where $f_j, j = 1, 2$ are the increasing concave functions, according to each of the control group and the treatment group, β, d and ε are the same as defined before.

The usual statistical inference can not be applied in this case. However, Bayes factor provides a simple solution for this case. We apply an individual Bayes increasing concave regression to both the control group and the treatment group, through Gibbs sampler method and draw the parameter estimates from their corresponding posterior densities. For each set of the sampled parameter estimates $(\hat{\tau}^c, \hat{c}^c, \hat{b}_1^c, \cdots, \hat{b}_{n_1-1}^c)$ and $(\hat{\tau}^t, \hat{c}^t, \hat{b}_1^t, \cdots, \hat{b}_{n_2-1}^t)$, where the subscripts c and t refer to the control group and the treatment group, n_1 counts the number of distinct baseline SBI measurements in the control group and n_2 counts the number of distinct baseline SBI measurements in the treatment group. We calculate the new τ as

$$\hat{\tau} = \frac{n_1 + n_2 - 2}{\frac{n_1-1}{\hat{\tau}^c} + \frac{n_2-1}{\hat{\tau}^t}},$$

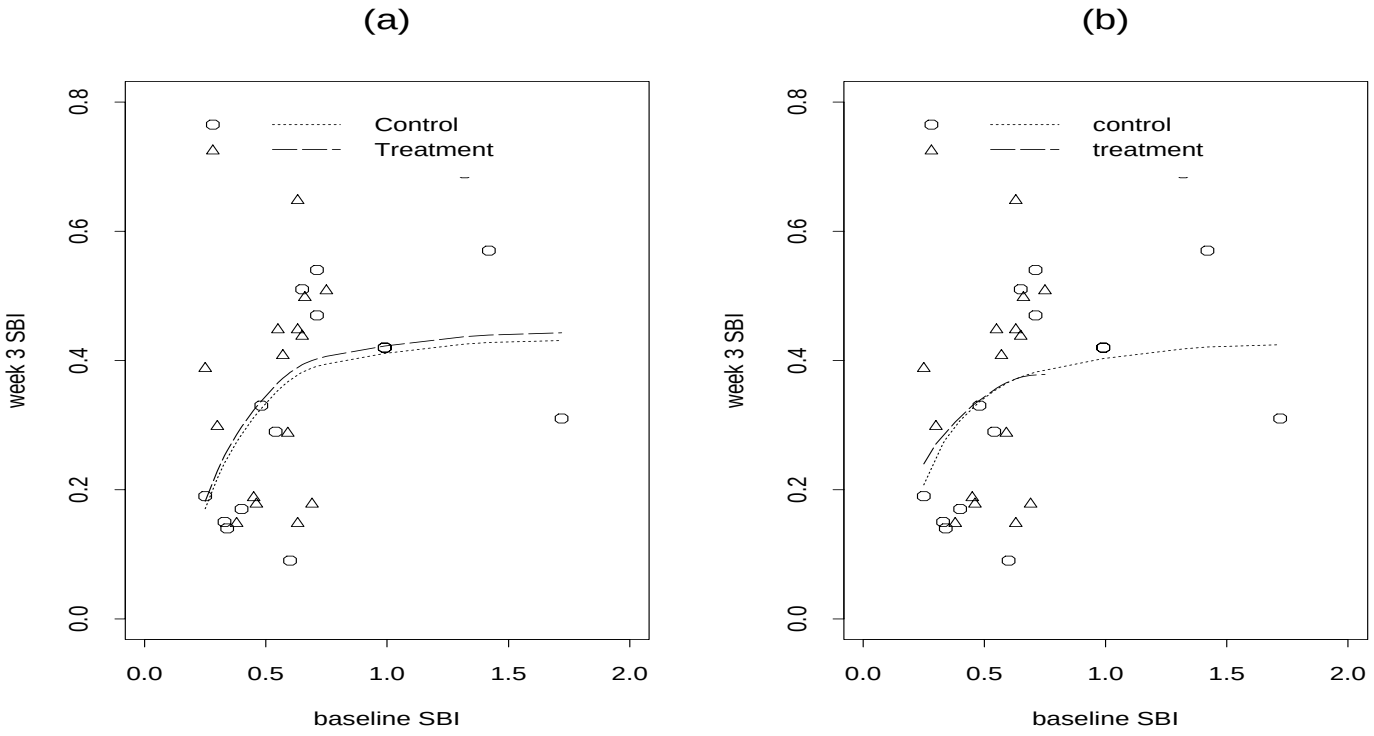

Figure 5.6: Mouthwash data: Bayes parallel increasing concave fit, $\gamma = 2$ (a); Bayes individual increasing concave fit (b). The dotted line is the fit for the control group, and the dashed line is the fit for the treatment group.

This is derived from the expression of the pooled sample variance

$$s^2_{pool} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

where s_1 and s_2 are the standard deviations of group one and two. Thus, τ can be considered as pooled τ of τ^c and τ^t. We then use $(\hat{\tau}, \hat{c}^c, \hat{b}^c_1, \cdots, \hat{b}^c_{n_1-1})$ and $(\hat{\tau}, \hat{c}^t, \hat{b}^t_1, \cdots, \hat{b}^t_{n_2-1})$ to calculate the estimate of $P_r(\mathcal{D}|H_1)$, $P_r(\mathcal{D}|H_0)$ is calculated as usual. Figure 5.6(b) shows two individual Bayes increasing concave fits to the mouthwash data, where the dotted line refers to the fit for control group, dashed line refers to the fit for the treatment group.

The results of hypotheses testings by Bayes factor method are listed in Table 5.10. For testing the hypothesis if there are significant differences between the control group and the

treatment group in the measurements of subjects' week 3 SBI. Bayes factor shows that the data is strongly in favor of the null hypothesis ($2logBf_{10} = -1.1581$). Bayes credible interval is also applied, as a comparison, where "Accept" means that the interval of 2.5% and 97.5% percentiles of $\hat{\beta}$ drawn from the posterior density contains 0. Both Bayes factor and Bayes credible interval give same result.

Table 5.10: Mouthwash data: applying Bayes factor in hypotheses testings

Hypothesis	Bf_{10}	$2log(Bf_{10})$	Credible Interval
H_0: $\beta = 0$	0.5604	-1.1581	Accept
H_0: parallel curves	0.0031	-11.5472	NA

For testing the hypothesis of no interaction between baseline SBI and groups, Bayes factor method again concludes that the data is strongly in favor of H_0 ($2logBf_{10} = -11.5472$), i.e., our parallel model holds.

5.7 Feet example revisited

The feet data (see Appendix A) contains three variables, variable *length* and *width* measure the length and width of the feet. The variable *gender* indicates whether the student is a boy (*gender*='B') or a girl (*gender*='G'). In chapter 3, we fit both parallel increasing concave regression and parallel monotone regression to fit the data. But the hypotheses tests cannot be done by then since the distribution of the test statistic is hard to determine. Now, let's apply Bayes factors to perform the hypothesis tests.

We fit the feet data by Bayes parallel increasing concave regression, also the Bayes parallel monotone regression as a simple comparison. The γ is chosen to be 0.3 for Bayes increasing concave regression and 2.7 for Bayes monotone regression, by the method of ordinary cross-validation. Figure 5.7 shows the plot of the function $V(\gamma)$ versus γ, where $V(\gamma)$ reaches its minimum around $\gamma = 0.3$ under the increasing concave constraint and $\gamma = 2.7$ under the monotone constraint.

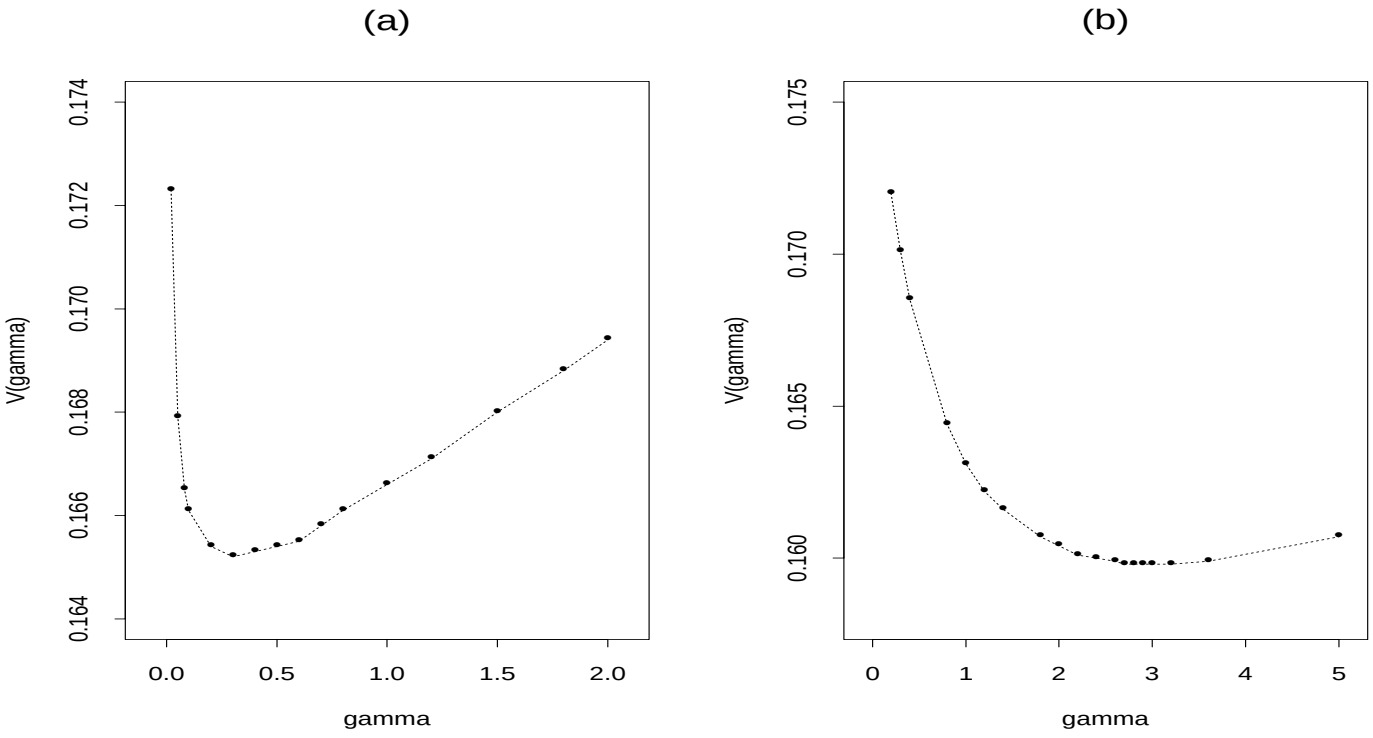

Figure 5.7: Feet data: plot of the OCV function $V(\gamma)$ vs. γ, (a) Bayes increasing concave regression is applied, (b) Bayes monotone regression is applied.

The model can be written as

$$y_i = f(x_i) + \beta d_i + \sigma\varepsilon_i,$$

where $i = 1, \cdots, n$, f is any function belongs to the increasing concave family (for increasing concave regression) or the monotone family (for monotone regression), ε_i's are independent random standard normal errors, σ and β are unknown constants, d is an indicator variable, where

$$d_i = \begin{cases} 1 & \text{if the } i^{th} \text{ student is a boy,} \\ 0 & \text{otherwise.} \end{cases}$$

Figure 5.8(b) shows the Bayes monotone fit. It has abrupt jumps and we perfer a smoother fit. Compared to the monotone fits, the Bayes increasing concave fit is much smoother (Figure 5.8 (a)).

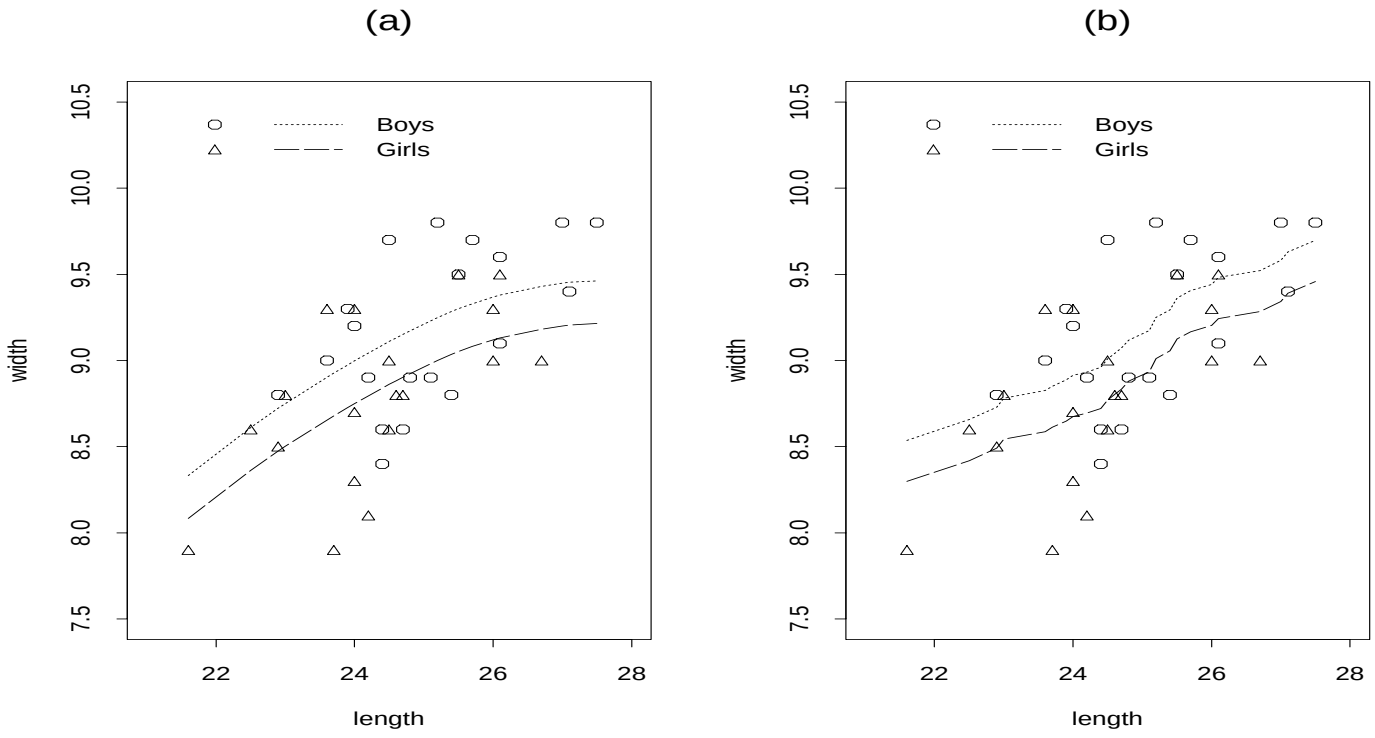

Figure 5.8: Feet data: Bayes parallel increasing concave fit, $\gamma = 0.3$ (a); Bayes parallel monotone fit (b), $\gamma = 2.7$.

Table 5.11 lists the parameter estimates from Bayes parallel increasing concave regression and Bayes parallel monotone regression. For each parameter, 2.5% percentile, 50% percentile, 97.5% percentile, mean and standard deviation (SD) are calculated. The c is the coefficient of the $\boldsymbol{v}^1$ vector, where $\boldsymbol{v}^1 = (1, \cdots, 1)'$; β is the coefficient of the indicator variable d; and $\tau = 1/\sigma^2$. Both fits give negative posterior mean for β, with $\hat{\beta} = -0.2485$ from Bayes increasing concave fit and $\hat{\beta} = -0.2384$ from Bayes monotone fit, boys seem to have wider feet than girls when their feet are of the same length.

The results of the hypothesis tests by Bayes factor method are listed in Table 5.12. For testing the hypothesis of no difference in the feet width of boys and girls, the values of Bayes factor are $2log(Bf_{10}) = 2.2615$ from increasing concave regression and $2log(Bf_{10}) =$

Table 5.11: Feet data: parameter estimates from posterior distribution.

Fit	Parameters	2.5% percentile	50% percentile	97.5% percentile	Mean	SD
Increasing concave	β	-0.5347	-0.2496	0.0612	-0.2485	0.1528
	τ	3.4127	5.0371	7.0222	5.0801	0.9276
	c	8.9073	9.1141	9.3208	9.1141	0.1055
Monotone	β	-0.5176	-0.2419	0.0657	-0.2384	0.1476
	τ	3.5756	5.2445	7.3149	5.2836	0.9599
	c	8.9200	9.1261	9.3242	9.1264	0.1029

Table 5.12: Feet data: Applying Bayes factor in hypotheses testings.

Type of Fit	Hypothesis	Bf_{10}	$2log(Bf_{10})$	Credible Interval
Increasing concave	H_0: $d = 0$	3.0979	2.2615	Accept
	H_0: parallel curves	0.1142	-4.3400	NA
Monotone	H_0: $d = 0$	4.0842	2.8143	Accept
	H_0: parallel curves	0.1582	-3.6874	NA

2.8143 from monotone regression, both show that there is no strong evidence against H_0, i.e., the difference is not statistically significant. Same results are obtained from Bayes credible interval.

For testing the null hypothesis of two parallel shape-restricted curves, the values of Bayes factor are $2log(Bf_{10}) = -4.3400$ from Bayes increasing concave regression, and $2log(Bf_{10}) = -3.6874$ from Bayes monotone regression, both show that the data is strongly in favor of H_0. Hence, we can conclude that our assumption of the parallel model does hold. Figure 5.9 shows the two individual Bayes increasing concave fit and the two individual Bayes monotone fit.

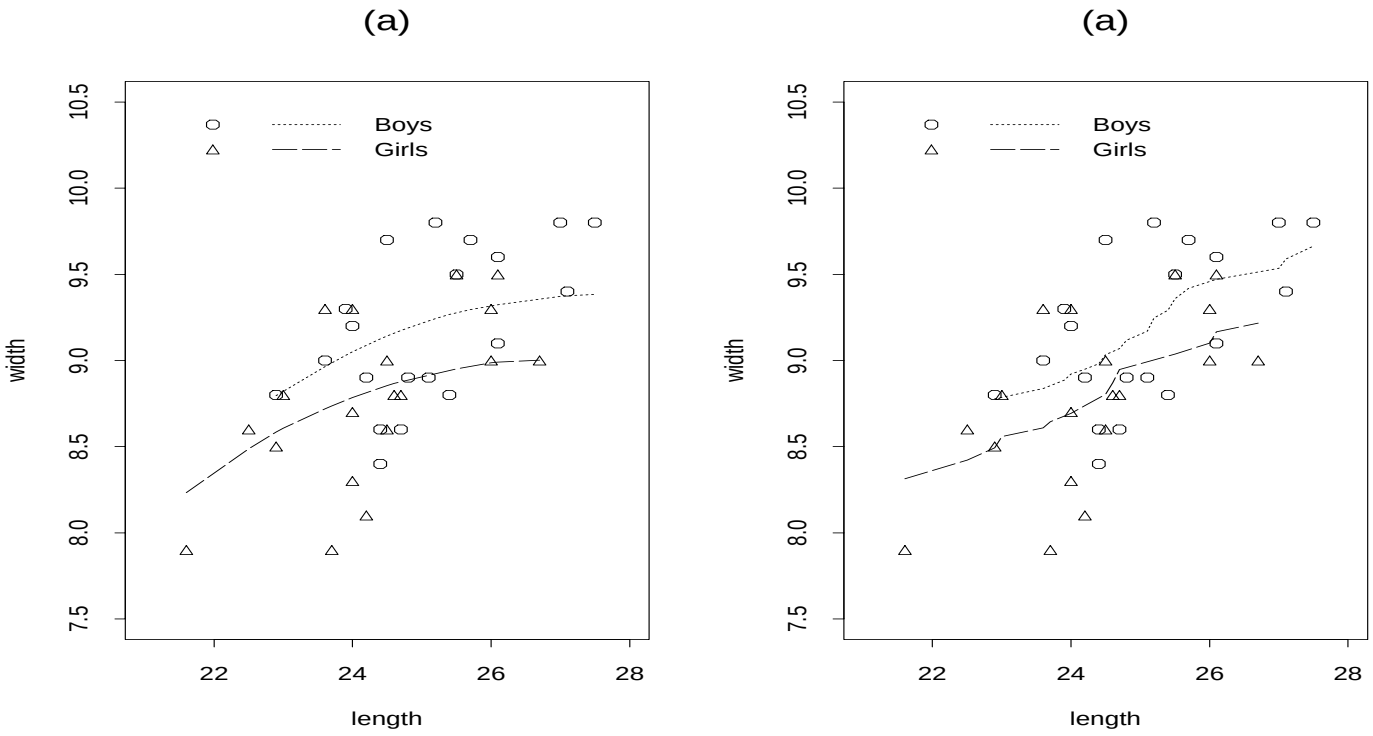

Figure 5.9: Feet data: two individual Bayes increasing concave fit (a); two individual Bayes monotone fit (b).

5.8 Summary

In this dissertation, we have focused on the shape-restricted ANCOVA model, obtained least-squares estimates for the shape-restricted ANCOVA regression. When our main interest is the categorical variables, we conduct a beta test statistic for the balanced shape-restricted ANCOVA model and its test size and power behaviors are studied and compared with related tests.

For the unbalanced shape-restricted ANCOVA, the distribution of the test statistic is hard to derive, hence we apply Bayesian methods to perform hypothesis tests and obtain all the information based on samples drawn from the posterior distribution. Two Bayesian analysis methods (the Bayes credible interval method and Bayes factor) are used. The Bayes

credible interval method is applied for simple hypotheses, such as contrasts within the levels of one categorical variable; Bayes factor is a general method that can deal with any kind of hypotheses, such as the test of interactions between two categorical variables with more than two levels each, and the test of interactions between categorical variables and covariates. We have examined and compared the test size and power of the Bayes credible interval method and the Bayes factor method, and the reasonable results ensure the feasibility of these two methods. We have presented three real world examples to demonstrate how to apply the above two methods and how easily they can be used to perform the hypotheses tests about the shape-restricted ANCOVA models.

The benefit of the shape-restricted ANCOVA regressions is that they can provide flexible fit to the data, while parametric ANCOVA regressions require the fixed forms of the underlying functions. Although other nonparametric methods, such as kernel smoothing technique, can be applied to ANCOVA model and obtain flexible fit, they might waste some important information about the true underlying functions. The limitation of the shape-restricted ANCOVA model is that the inference for the unbalanced data is hard to derive, mainly because of the difficulty in deriving the error degrees of freedom.

As future work, we will try to apply randomization tests on the unbalanced data set, where a randomization test is a permutation test which is based on randomization. The test is carried out in the following manner. A test statistic is computed for the experimental data, then the data are permuted repeatedly in a manner consistent with the random assignment procedure, and the test statistic is computed for each of the resulting data permutations. These data permutations, including the one representing the obtained results, constitute the reference set for determining significance. The proportion of data permutations in the reference set with test statistic values greater than or equal to the value for the experimentally obtained results is the P-value.

In addition, we can add order restrictions on the group differences between curves. The comparison of the shape-restricted ANCOVA regressions and nonlinear regressions is also an interesting issue we want to address later.

Bibliography

[1] Allen, D. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16**, 125-127.

[2] Berger, J. O. and Mortera, J. (1991). Interpreting the stars in precise hypothesis testing. *International Statistical Review*, **59**, 337-353.

[3] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice system. *J. R. Statist. Soc. B*, **36**, 192-236.

[4] Carlin, B. P. and Chib, S. (1993). Bayesian model choice via Markov chain Monte Carlo. *Research Report 93-006, Division of Biostatistics, University of Minnesota.*

[5] Carlin B. P. and Polson N. G. (1991). Inference for nonconjugate Bayesian models using the Gibbs sampler. *Canadian J. Statist.*, **19**, 399-405.

[6] Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist.*, **16**, 136-146.

[7] Damien, P. and Walker, S. G. (2001). Sampling truncated normal, beta, and gamma densities. *Journal of Computational and Graphical Statistics*, **10**, 206-215.

[8] De Bruijn, N. G. (1970). *Asymptotic methods in analysis.* Amsterdam: North-Holland.

[9] DeGroot, M. H. (1970). *Optimal statistical decisions.* New York: McGraw-Hill.

[10] Denby, L. (1986). Smooth regression functions. *Statistical Research Report*, **26**, Murray Hill: AT&T Bell Laboratories.

[11] Dykstra, R. L. (1983). An algorithm for restricted least squares regression. *J. Amer. Statist. Assoc.*, **78**, 837-842.

[12] Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psych. Rev.*, **70**, 193-242.

[13] Engle, R., Granger, C., Rice, J. and Weiss, A. (1986). Nonparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.*, **81**, 310-320.

[14] Fraser, D. A. S. and Massam, H. (1989). A mixed primal-dual bases algorithm for regression under inequality constraints. Application to convex regression. *Scand. J. Statist.*, **16**, 65-74.

[15] Gasser, Th. and Müller, H. G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for curve estimation*, Gasser, Th. and Rosenblatt, M. (eds), 23-68. Berlin: Springer.

[16] Gelfand, A. E, Hills, S. E., Racine-Poon, A., Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampler. *J. Amer. Statist. Assoc.*, **85**, 972-985.

[17] Gelfand, A. E. and Smith, A. F. M (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, **85**, 398-409.

[18] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs Distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-741.

[19] Genz, A. and Kass, R. E. (1993). Subregion adaptive integration of functions having a dominant peak. *Technical report, Department of Statistics, Carnegie Mellon University.*

[20] Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, **57**, 1317-1340.

[21] Green, P. (1985). Linear models for field trials, smoothing and cross-validation. *Biometrika*, **72**, 527-537.

[22] Green, P., Jennison, C. and Seheult, A. (1985). Analysis of field experiments by least squares smoothing. *J. R. Statist. Soc. B*, **47**, 299-315.

[23] Härdle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.*, **12**, 1465-1481.

[24] Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statist. Sci.*, **1**, 297-310.

[25] Heckman, N. (1986). Spline smoothing in a partly linear model. *J. R. Statist. Soc. B*, **48**, 244-248.

[26] Hildreth, C. (1954). Point estimates of ordinates of concave functions. *J. Amer. Statist. Assoc.*, **49**, 598-619.

[27] Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proc. Camb. Phil. Soc.*, **31**, 203-222.

[28] Jeffreys, H. (1961). *Theory of Probability*. Third edition. Oxford University Press.

[29] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. statist. Assoc.*, **90**, 773-795.

[30] Leonard, T. (1982). Comment on "A simple predictive density function". *J. Amer. Statist. Assoc.*, **77**, 657-658.

[31] Lindley, D. V. (1961). The use of prior probability distributions in statistical inference and decisions. Fourth Berkeley Symposium. In *Proc. of Fourth Berkeley Symposium on Math. Stat. and Prob.*, Berkeley: U. of California Press, 453-468.

[32] McCulloch, R. E. and Rossi, P. E. (1991). A Bayesian approach to testing the arbitrage pricing theory. *J. Econometrics*, **49**, 141-168.

[33] Meyer, M. C. (1999t). An algorithm for projections onto convex cones with applications to nonparametric regression and quadratic programming. *Department of Statistics Technical Report, the University of Georgia.*

[34] Meyer, M. C. (1999). An extension of the mixed primal-dual bases algorithm to the case of more constraints than dimemsions. *J. Statist. Plan. Infer.*, **81**, 13-31.

[35] Meyer, M. C. (2003). A test for linear versus convex regression function using shape-restricted regression. *Biometrika*, **90**, 223-232.

[36] Meyer, M. C., and Laud, P. W. (2005). A Bayesian approach to shape-restricted regression. *Department of Statistics Technical Report, the University of Georgia, to appear.*

[37] Meyer, M. C., and Woodroofe, M. (2000). On the degrees of freedom in shape-restricted regression. *Ann. Statist.*, **28**, 1083-1104.

[38] Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap. *J. R. Statist. Soc. B*, **56**, 3-48.

[39] Platt, R. (1985). Personal communication with Paul Speckman.

[40] Pshenichny, B. N. & Danilin, M. Y. (1978). *Numerical methods in extremal problems.* Mir Publishers, Moscow.

[41] Raftery, A. E. (1994). Hypothesis testing and model selection with posterior simulation. In *Practical Markov Chain Monte Carlo*, London: Chapman and Hall.

[42] Raftery, A. E. and Banfield, J. D. (1990). Stopping the Gibbs sampler, the use of morphology and other issues in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**, 32-43.

[43] Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.*, **12**, 1215-1230.

[44] Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. John Wiley & Sons, New York.

[45] Rockafellar, R. T. (1970). *Convex analysis*. Princeton university press. New Jersey.

[46] Rosenkranz, S. (1992). The Bayes factor for model evaluation in a hierarchical Poisson model for area counts. *Ph. D. dissertation, Department of Biostatistics, University of Washington.*

[47] Shiau, J., Wahba, G. and Johnson, D. R. (1986). Partial spline models for the inclusion of tropopause and frontal boundary information in otherwise smooth two and three dimensional objective analysis. *J. Atmos. Ocean. Technol.*, **3**, 714-725.

[48] Slate, E. (1994). Parameterizations for natural exponential families with quadratic variance functions. *J. Amer. Statist. Assoc.*, **89**, 1471-1482.

[49] Speckman, P. (1988). Kernel smoothing in partial linear models. *J. R. Statist. Soc. B*, **50**, 413-436.

[50] Speckman, P., Chiu, J-Er, Hewett, J. E. and Bertelson, S. E. (2001). A one-sided test adjusting for covariates by ranking residuals following smoothing. *Department of Statistics, University of Missouri-Columbia, Technical Report.*

[51] Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.*, **81**, 82-86.

[52] Wahba, G. (1971). *Spline models for observational data*. Captial city press, Montpelier, Vermont.

[53] Wahba, G. (1984). Partial spline models for the semiparametric estimation of functions of several variables. In *Analyses for Time Series, Japan-US Joint Sem.*, 319-329. Tokyo: Institute of Statistical Mathematics.

[54] Wahba, G. and Wold, S. (1975). A completely automatic French curve. *Commun. Statist.*, **4**, 1-17.

[55] Wilhelmsen, D. R. (1976). A nearest point algorithm for convex polyhedral cones and applications to positive linear approximation. *Math. Comp.*, **30**, 48-57.

[56] Wu, C. F. (1982). Some algorithms for concave and isotonic regression. *Stud. Management Sci.*, **19**, 105-116.

[57] Young, S. G. and Bowman, A. W. (1995). Non-Parametric Analysis of covariance. *Biometrics*, **51**, 920-931.

[58] Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*. New York: Wiley.

Appendix A

Feet data

Table A.1: Feet data

ID	Length	Width	Gender
1	21.6	7.9	G
2	22.5	8.6	G
3	22.9	8.8	B
4	22.9	8.5	G
5	23	8.8	G
6	23.6	9	B
7	23.6	9.3	G
8	23.7	7.9	G
9	23.9	9.3	B
10	24	9.2	B
11	24	8.3	G
12	24	8.7	G
13	24	9.3	G
14	24.2	8.9	B
15	24.2	8.1	G
16	24.4	8.4	B
17	24.4	8.6	B
18	24.5	9.7	B
19	24.5	8.6	G
20	24.5	9	G
21	24.6	8.8	G
22	24.7	8.6	B
23	24.7	8.8	G
24	24.8	8.9	B
25	25.1	8.9	B
26	25.2	9.8	B
27	25.4	8.8	B
28	25.5	9.5	B
29	25.5	9.5	G
30	25.7	9.7	B
31	26	9	G
32	26	9.3	G
33	26.1	9.1	B
34	26.1	9.6	B
35	26.1	9.5	G
36	26.7	9	G
37	27	9.8	B
38	27.1	9.4	B
39	27.5	9.8	B

Appendix B

Senic data

Table B.1: Senic data

ID	Patients' infection risk	Region	Medical school affiliation	Census
1	3.1	3	0	20
2	2.9	1	0	37
3	3.7	4	0	37
4	4.2	2	0	38
5	3.1	1	0	39
6	1.3	2	0	40
7	2.7	4	0	40
8	5.4	4	0	42
9	2.1	2	0	44
10	4.2	4	0	47
11	2.6	4	0	47
12	1.3	3	0	49
13	4.6	3	0	50
14	1.6	2	0	51
15	4.5	4	0	51
16	2	3	0	52
17	5.6	4	0	53
18	5.3	4	0	55
19	2.5	1	0	57
20	6.3	1	0	59
21	3.7	2	0	59
22	2	3	0	59
23	4.2	3	0	61
24	1.7	3	0	61
25	3.5	1	0	65
26	2.9	3	0	65
27	4.4	2	0	66
28	3.4	1	0	67
29	5.8	3	0	68
30	2.8	2	0	69
31	2.9	3	0	69
32	2.7	3	0	82
33	4.5	4	0	84
34	4.7	4	0	85
35	4.4	4	0	85
36	4.7	3	0	90
37	1.4	3	0	90
38	1.8	3	0	93
39	3	3	0	95
40	5.5	2	0	96

ID	Patients' infection risk	Region	Medical school affiliation	Census
41	4.2	2	0	103
42	4.3	3	0	105
43	5.5	4	0	109
44	6.4	1	0	113
45	4.9	2	0	113
46	5.1	2	0	113
47	7.8	2	0	115
48	5.3	2	0	123
49	4.5	1	0	124
50	4.8	1	0	127
51	4.1	3	0	127
52	2.3	3	0	128
53	4.3	3	0	130
54	5.7	1	0	134
55	3.5	2	0	139
56	4.3	1	0	141
57	3.4	2	0	143
58	4.5	1	1	144
59	3.7	3	0	145
60	5.1	2	0	147
61	4.6	3	0	151
62	4.5	3	0	153
63	3.2	1	0	156
64	4.8	2	0	163
65	5.3	1	0	164
66	4.4	2	0	165
67	4.1	2	0	168
68	6.3	2	0	170
69	3.8	3	0	178
70	5	1	0	181
71	4.4	2	0	191
72	5.2	2	0	200
73	5.6	1	0	207
74	4.1	4	0	207
75	4.4	1	0	217
76	3.9	1	0	217
77	4.3	2	0	223
78	4.5	3	0	230
79	4.8	1	0	237
80	4.2	3	0	238

ID	Patients' infection risk	Region	Medical school affiliation	Census
81	5	1	1	240
82	5.2	3	0	241
83	4	4	1	244
84	3.7	3	0	248
85	7.7	1	1	252
86	6.1	1	0	258
87	4.3	4	0	265
88	5	1	0	270
89	6.5	1	0	273
90	6.6	1	0	308
91	5.6	3	1	313
92	2.9	2	1	320
93	5.5	3	0	326
94	7.6	3	0	330
95	2.9	3	0	349
96	5.8	3	1	359
97	3.9	2	1	374
98	5	3	0	391
99	5.4	2	1	399
100	4.5	3	0	404
101	3.9	2	1	413
102	4.3	2	0	413
103	4.3	1	0	421
104	5.7	2	1	441
105	5.5	2	0	446
106	6.2	3	1	452
107	4.8	3	0	468
108	5.7	2	1	546
109	4.1	3	0	547
110	4.6	4	1	581
111	4.9	1	1	591
112	4.9	2	1	595
113	5.9	1	1	791