5-2012

Bayesian regression analysis.

Sara Evans
University of Louisville

BAYESIAN REGRESSION ANALYSIS

By

Sara Evans
M.A., University of Louisville, 2012

A Thesis
Submitted to the Faculty of the
College of Arts and Sciences of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Master of Arts

Department of Mathematics
University of Louisville
Louisville, KY

May 2012

BAYESIAN REGRESSION ANALYSIS

Submitted by

Sara Evans

A Thesis Approved on

April 10, 2012
(Date)

by the Following Reading and Examination Committee:

Prasanna Sahoo, Thesis Director

Ryan Gill

David Brown

DEDICATION

This thesis is dedicated to my late grandmother

Mrs. Marjorie Thomas Marbury (1923-2012)

and grandfather

Mr. Robert C. Marbury

who always believed I could succeed.

ABSTRACT

BAYESIAN REGRESSION ANALYSIS

Sara Evans

May 12, 2012

Regression analysis is a statistical method used to relate a variable of interest, typically y (the dependent variable), to a set of independent variables, usually, $x_1, x_2, \ldots, x_n$. The goal is to build a model that assists statisticians in describing, controlling, and predicting the dependent variable based on the independent variable(s). There are many types of regression analysis: Simple and Multiple Linear Regression, Nonlinear Regression, and Bayesian Regression Analysis to name a few. Here we will explore simple and multiple linear regression and Bayesian linear regression. For years, the most widely used method of regression analysis has been the Frequentist methods, or simple and multiple regression. However, with the advancements of computers and computing tools such as WinBUGS, Bayesian methods have become more widely accepted. With the use of WinBUGS, we can utilize a Markov Chain Monte Carlo (MCMC) method called Gibbs Sampling to simplify the increasingly difficult calculations. Given that Bayesian regression analysis is a relatively "new" method, it is not without faults. Many in the statistical community find that the use of Bayesian techniques is not a satisfactory method since the choice of the prior distribution is purely a guessing game and varies from statistician to statistician. In this thesis, an example is presented using both Frequentist and Bayesian methods and a comparison is made between the two. As computers

become more advanced, the use of Bayesian regression analysis may become more widely accepted as the method of choice for regression analyses as it allows for the interpretation of a "probability as a measure of degree of belief concerning actual data observed." [5]

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
FREQUENTIST REGRESSION ANALYSIS

1.1 Introduction

Frequentist statistics became widely used in the first half of the 20th century, though the method of least-squares has been around since the early 1800s. Discovered independently by Carl Friedrich Gauss in Germany and Adrien Marie Legendre in France, the method of least-squares is the most frequently used regression method to date. In its early days, it was applied to astronomic and geodetic data and was originally published in an appendix to a book published by Legendre on determining the orbits of comets. Today, in many undergraduate statistics courses, students may only be exposed to the "frequentist" approach to statistics. [9]

1.2 Simple Linear Regression

We begin with simple linear regression, the "simplest" of the regression analyses. It is defined to be the least squares estimator of a regression model with a single explanatory (independent) variable in which a straight line is fit through n points so that the sum of squared residuals (SSR), $\sum e_i^2$, is minimized. That is, the distance between the regression line and the data points is minimal.

We can think of the model as similar to the slope-intercept form of a line

with an error (residual) term. The simple linear regression model is given by

$$y_i = \alpha + \beta x_i + e_i$$

for $i = 1, 2, \ldots, n$, where y_i is the response (dependent) variable, x_i is the explanatory/predictor (independent) variable, α and β are the unknown parameters that are to be estimated, and e_i is the residual term. The e_i term is independent and identically distributed (iid) with a normal distribution with mean 0 and unknown variance σ^2. Along with α and β, σ^2 can be estimated.

1.2.1 Estimating the Unknown Parameters α and β

The goal in estimating the unknown parameters is to find a line that "best" fits the data. To do this we use an estimated regression line

$$Y = \hat{\alpha} + \hat{\beta} X,$$

where $\hat{\alpha}$ and $\hat{\beta}$ are estimates of α and β, so that we may look at the size of the residuals

$$\hat{e}_i = y_i - (\hat{\alpha} + \hat{\beta} x_i).$$

It is important to note any estimates will be denoted with a "hat" symbol, similar to the estimates for α and β.

Similar to the residual, e_i, the estimated residual, $\hat{e}_i$, is the vertical distance between the estimated regression line and the data point (x_i, y_i). The idea is to choose $\hat{\alpha}$ and $\hat{\beta}$ so that the residuals are "small." While there are several methods with which to minimize the residuals, here we will use the method of least squares.

We measure the overall size of the residuals by $\sum \hat{e}_i^2$. The least squares estimates of α and β, $\hat{\alpha}$ and $\hat{\beta}$, give the least SSR.

To estimate α and β, first find the derivative of $\sum e_i^2$ with respect to α and

β, respectively.

$$\frac{\partial}{\partial \alpha} \sum_{i=1}^{n} e_i^2 = 2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)(-1)$$
$$\frac{\partial}{\partial \beta} \sum_{i=1}^{n} e_i^2 = 2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)(-x_i)$$

Setting each derivative to zero gives us

$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0 \tag{1.1}$$
$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) x_i = 0 \tag{1.2}$$

From (1.1), we get

$$\sum_{i=1}^{n} y_i = n\alpha + \beta \sum_{i=1}^{n} x_i$$
$$n\overline{y} = n\alpha + n\beta\overline{x}$$
$$\overline{y} = \alpha + \beta\overline{x} \tag{1.3}$$

where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ denote the averages of y_i and x_i, respectively. From (1.2), we see that

$$\sum_{i=1}^{n} x_i y_i = \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2.$$

We can rewrite this equation as

$$\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) + n\overline{x}\,\overline{y} = n\alpha\overline{x} + \beta \sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x}) + n\beta\overline{x}^2.$$

Define

$$S_{xy} := \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$
$$S_{xx} := \sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x}).$$

Then

$$S_{xy} + n\overline{x}\,\overline{y} = n\alpha\overline{x} + \beta S_{xx} + n\beta\overline{x}^2.$$

Using (1.3), we have

$$S_{xy} + n\overline{x}\,\overline{y} = n\overline{x}(\overline{y} - \beta\overline{x}) + \beta S_{xx} + n\beta\overline{x}^2$$

$$S_{xy} + n\overline{x}\,\overline{y} = n\overline{x}\,\overline{y} - n\beta\overline{x}^2 + \beta S_{xx} + n\beta\overline{x}^2$$

$$S_{xy} = \beta S_{xx}$$

$$\beta = \frac{S_{xy}}{S_{xx}}$$

Hence

$$\alpha = \overline{y} - \frac{S_{xy}}{S_{xx}}\,\overline{x}.$$

Thus the estimates of α and β are given by

$$\hat{\alpha} = \overline{y} - \frac{S_{xy}}{S_{xx}}\,\overline{x} \tag{1.4}$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}. \tag{1.5}$$

[12]

The random variable (RV) Y given $X = x$ will be denoted by Y_x. When choosing, in succession, values $x_1, x_2, \ldots, x_n$ for x, a sequence

$$Y_{x_1}, Y_{x_2}, \ldots, Y_{x_n}$$

of RVs is obtained. For convenience, we will denote this sequence of RVs as

$$Y_1, Y_2, \ldots, Y_n.$$

To do statistical analysis, we make the following assumptions:

1. $E(Y_x) = \alpha + \beta x$ so that $y_i = E(Y_i) = \alpha + \beta x_i$;

2. $Y_1, Y_2, \ldots, Y_n$ are independent RVs;

3. Each Y_i, for $i = 1, 2, \ldots, n$, has the same variance, σ^2.

[10]

THEOREM 1.1. *Under the above assumptions, the least squares estimators, $\hat{\alpha}$ and $\hat{\beta}$, of the linear model $E(Y \mid x) = \alpha + \beta x$ are unbiased.*

Proof.

$$\begin{aligned}
E(\widehat{\beta}) &= E\left(\frac{S_{xY}}{S_{xx}}\right) = \frac{1}{S_{xx}} E\left(\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})\right) \\
&= \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \overline{x})E(Y_i - \overline{Y}) \\
&= \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \overline{x})E(Y_i) - \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \overline{x})E(\overline{Y}) = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \overline{x})E(Y_i) \\
&= \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \overline{x})(\alpha + \beta\, x_i) \\
&= \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \overline{x})\alpha + \frac{1}{S_{xx}}\beta\sum_{i=1}^{n}(x_i - \overline{x})x_i \\
&= \frac{1}{S_{xx}}\beta\sum_{i=1}^{n}(x_i - \overline{x})x_i = \frac{1}{S_{xx}}\beta\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x}) \\
&= \frac{1}{S_{xx}}\beta S_{xx} = \beta
\end{aligned}$$

Hence, $E(\widehat{\beta}) = \beta$.

$$\begin{aligned}
E(\widehat{\alpha}) &= E\left(\overline{Y} - \frac{S_{xY}}{S_{xx}}\overline{x}\right) \\
&= E(\overline{Y} - \widehat{\beta}\overline{x}) \\
&= E(\overline{Y}) - E(\widehat{\beta})\overline{x} \\
&= E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) - \beta\overline{x} \\
&= \frac{1}{n}\sum_{i=1}^{n} E(Y_i) - \beta\overline{x} \\
&= \frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta\, x_i) - \beta\overline{x} \\
&= \alpha + \beta\overline{x} - \beta\overline{x} \\
&= \alpha
\end{aligned}$$

Hence $E(\widehat{\alpha}) = \alpha$.

Thus, $\hat{\beta}$ and $\hat{\alpha}$ are unbiased estimators for β and α, respectively. [12] □

1.2.2 Hypothesis Testing

In a regression model if $\beta = 0$, then the response variable Y and the explanatory variable X are not related. Hence we test this relationship between Y and X using a hypothesis test where the null hypothesis is $H_o : \beta = 0$ and the alternative hypothesis is $H_a : \beta \neq 0$.

The null and alternative hypotheses allow for the comparison of the full model

$$Y = \alpha + \beta X + e$$

with the reduced model

$$Y = \alpha + e$$

to determine whether the relationship between X and Y is significant.

To test the hypothesis that $\beta = 0$, use the least squares estimator $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ and then estimate σ^2. This estimate can be found using the maximum likelihood method and is given by

$$\hat{\sigma}^2 = \frac{1}{n}\left[S_{YY} - \frac{S_{xy}}{S_{xx}}S_{xy}\right].$$

THEOREM 1.2. *An unbiased estimator s^2 of σ^2 is given by*

$$s^2 = \frac{SSR}{n-2} = \frac{n\hat{\sigma}^2}{n-2}.$$

Proof. First, it should be noted that $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$ and that $E(\chi^2(n-2)) = n-2$.

$$\begin{aligned} E(s^2) &= E\left(\frac{n\hat{\sigma}^2}{n-2}\right) \\ &= \frac{\sigma^2}{n-2}E\left(\frac{n\hat{\sigma}^2}{\sigma^2}\right) \\ &= \frac{\sigma^2}{n-2}E(\chi^2(n-2)) \\ &= \frac{\sigma^2}{n-2}n-2 \\ &= \sigma^2 \end{aligned}$$

Hence, $E(s^2) = \sigma^2$ and thus s^2 is an unbiased estimator of σ^2.[12] □

We can calculate the standard deviation of $\hat{\beta}$ using the formula

$$SD(\hat{\beta}) = \frac{\sigma}{\sqrt{S_{xx}}}.$$

Here, σ is the standard deviation of the population of errors. We really wish to use the estimate of the standard deviation in subsequent steps so we must replace σ with an unbiased estimate, s, to obtain

$$\widehat{SD(\hat{\beta})} = \frac{s}{\sqrt{S_{xx}}}.$$

Next, we evaluate the test statistic

$$\begin{aligned} |t| &= \left|\frac{(\hat{\beta}-\beta)}{\sqrt{\frac{s^2}{S_{xx}}}}\right| \\ &= \frac{\hat{\beta}-\beta}{\hat{\sigma}}\sqrt{\frac{(n-2)S_{xx}}{n}} \end{aligned} \tag{1.6}$$

and compare it to $t_{\gamma/2}(n-2)$, the Student's $t-$distribution with $n-2$ degrees of freedom at a given significance level γ. [5]

The hypothesis test at a given significance level, $100(1-\gamma)\%$, is then to

"Reject $H_o : \beta = 0$ if $|t| > t_{\gamma/2}(n-2)$."

Rejecting the null hypothesis does not necessarily imply that we should accept the alternative hypothesis. However, it does indicate that there is a significant relation between the explanatory and response variables X and Y, respectively.

1.2.3 Confidence Interval

In frequentist statistics, it is often necessary to give a range of values with which we are $100(1-\gamma)\%$ confident that β falls within that range. From (1.6), we can determine a confidence interval for β; that is, let $\theta = \hat{\sigma}\sqrt{\frac{n}{(n-2)S_{xx}}}$. Then the $100(1-\gamma)\%$ confidence interval is given by

$$P\left(\hat{\beta} - \theta t_{\gamma/2} < \beta < \hat{\beta} + \theta t_{\gamma/2}\right) = 1 - \gamma.$$

In interval notation, we can express this as

$$\left[\hat{\beta} - \hat{\sigma}\sqrt{\frac{n}{(n-2)S_{xx}}}t_{\gamma/2}\,,\ \hat{\beta} + \hat{\sigma}\sqrt{\frac{n}{(n-2)S_{xx}}}t_{\gamma/2}\right].$$

1.3 Multiple Linear Regression

Multiple regression occurs when there are two or more explanatory variables present. This means the model, in terms of RVs, is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e.$$

In terms of observed data, the model is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i$$

for $i = 1, 2, \ldots, n$. As is similar to simple regression, x_i, y_i and e_i are the explanatory, response, and residual variables, respectively. $\beta_0, \beta_1, \ldots, \beta_p$ are the unknown parameters that are to be estimated. We note that, when $p = 1$, the multiple linear

regression reduces to simple linear regression and it is a generalization of the simple linear regression.

To assist with calculations, we introduce some matrix notation. We use bold face letters to denote matrices and vectors. We use upper case letters for matrices and lower case letters for vectors to distinguish between the two. Let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \text{ and } \qquad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}.$$

[5] Using this notation, the observed data model can now be expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}.$$

1.3.1 Estimating the Unknown Parameters Vector $\boldsymbol{\beta}$

We define the estimates of the parameters $\beta_0, \beta_1, \ldots, \beta_p$ as $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ where these estimates minimize the least squares residuals $\sum \hat{e}_i^2$, where $\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip})$. When written out, formulas for these estimates can become complex for each individual parameter. However, by introducing matrix notation, the formula becomes compact. In fact, the formula for the the least squares estimates is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \tag{1.7}$$

where $\mathbf{X}^{\mathrm{T}}$ is the transpose of the matrix $\mathbf{X}$. The objective in using this formula is to minimize the Euclidean length of the difference

$$||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}||.$$

Suppose

$$\hat{\boldsymbol{\beta}}^{\star} = \begin{bmatrix} \beta_0^{\star} \\ \beta_1^{\star} \\ \vdots \\ \beta_n^{\star} \end{bmatrix}$$

is a minimizing vector. Then the line $y = \beta_0^{\star} + \beta_1^{\star} x_{i1} + \cdots + \beta_p^{\star} x_{ip}$ is referred to as the least squares line fit to the data. To see how to find $\hat{\boldsymbol{\beta}}$, fix a vector $\mathbf{y}$ in $\mathbb{R}^n$. As $\hat{\boldsymbol{\beta}}$ varies, the vectors $\mathbf{X}\hat{\boldsymbol{\beta}}$ form a subspace of $\mathbb{R}^n$. If we wish $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ to have minimum length, then it must be orthogonal to the column space of the matrix $\mathbf{X}$. Let $\hat{\boldsymbol{\beta}}^{\star}$ be such a vector so that $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\star}$ is orthogonal to the column space. Then, the inner product of $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\star}$ with $\mathbf{X}\hat{\boldsymbol{\beta}}$ is zero for any vector $\hat{\boldsymbol{\beta}}$. Thus,

$$(\mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\star}) = 0$$

for all vectors $\hat{\boldsymbol{\beta}}$. Recall that $(\mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}} = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}$. Then, using a bit of algebra,

$$\hat{\boldsymbol{\beta}}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}}^{\star}) = 0$$

for all vectors $\hat{\boldsymbol{\beta}}$. The fixed vector $\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}}^{\star}$ is orthogonal to every vector $\hat{\boldsymbol{\beta}}$ if it is the zero vector. That, is,

$$\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}}^{\star}.$$

Since $\mathbf{X}$ is a $n \times p$ matrix and $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is a $p \times p$ matrix, and if $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible, then $\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}}^{\star}$ has a unique solution $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\star}$ [4]. Namely,

$$\hat{\boldsymbol{\beta}}^{\star} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

1.3.2 Hypothesis Testing

Similar to Simple Linear Regression. we wish to test the significance of the explanatory variables and whether these variables hold significant explanatory information. Our goal is to compare the full model, $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + e$, with the reduced model, $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q + e$ where $q < p$, to test whether there is a significant difference between the two models. As one would expect, the null hypothesis is given by

$$H_0 : \beta_{q+1} = \cdots = \beta_p = 0.$$

Rather than comparing the individual parameters and gathering data from each, it may be simpler to compare the models instead. The size of the residual e_i tells the suitability of the model. "The smaller the residual, the better the model fits the data" [5]. Let the sum of squared residuals be denoted by SSR. Then SSR_{full} and SSR_{reduced} can be used to compare the reduced model against the full model. The test statistic for the null hypothesis is given by

$$F = \frac{SSR_{\text{reduced}} - SSR_{\text{full}}}{(p-q)\hat{\sigma}^2}$$

where $\hat{\sigma}^2$ is an estimate of the distribution of the random errors. In order for $\hat{\sigma}^2$ to be an unbiased estimate of σ^2, define

$$\hat{\sigma}^2 = \frac{SSR_{\text{full}}}{n-(p+1)}$$

where $n-(p+1)$ is used rather than $n-1$. The use of $p+1$ comes from the fact that we must estimate $p+1$ unknown parameters $\beta_0, \beta_1, \ldots, \beta_p$ in order to form the residuals $\hat{e}_i$. If we assume the random errors are distributed normally, then the test statistic F has an F distribution with $p-q$ and $n-(p+1)$ degrees of freedom at significance level γ, denoted by $F_\gamma(p-q, n-(p+1))$, when we fail to reject the null.

The hypothesis test is then to

"Reject $H_o : \beta_{q+1} = \cdots = \beta_p = 0$ if $F > F_\gamma(p-q, n-(p+1))$."

Rejecting the null hypothesis indicates that there is a significant relation between the response variable $\mathbf{y}$ and the explanatory variable $\mathbf{X}$.

1.3.3 Confidence Interval

We can compute the $100(1-\gamma)\%$ confidence interval for multiple linear regression models in a way similar to that discussed in Section 1.2.3. For a certain $\beta_j, j = 0, 1, \ldots, n$, we have the $100(1-\gamma)\%$ confidence interval (see [3]) as

$$P\left(\hat{\beta}_j - t_{\gamma/2}(n-(p+1))\sqrt{C_{jj}} < \beta_j < \hat{\beta}_j + t_{\gamma/2}(n-(p+1))\sqrt{C_{jj}}\right) = 1-\gamma,$$

where C_{jj} is the diagonal element of C, a symmetric variance-covariance matrix of the estimated regression coefficients defined by

$$C = \hat{\sigma}^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1},$$

that represents the variance of $\hat{\beta}_j$. We use $t_{\gamma/2}(n-(p+1))$ for the confidence interval rather than $F_{\gamma}(p-q, n-(p+1))$ since we are finding the interval for a certain β_j. In interval notation, we can express this as

$$\left[\hat{\beta}_j - t_{\gamma/2}(n-(p+1))\sqrt{C_{jj}}\,,\ \hat{\beta}_j + t_{\gamma/2}(n-(p+1))\sqrt{C_{jj}}\right].$$

CHAPTER 2
BAYESIAN REGRESSION ANALYSIS

2.1 Introduction

In 1764, over a year after his death, the theorem appropriately named for its founder, Thomas Bayes, was published. It is this theorem that lends its name to the modern Bayesian approach to data analysis. However, what we know as Bayesian analysis today has not always had a "clear run" since 1764. In fact, due to their inability to handle prior probabilities properly, Bayes' methods lacked respect in the 19th century. Without the aid of powerful computers, Bayesian analysis, though not completely forgotten, was widely unused. Thanks to new computational methods and the accessibility of more powerful computers in the late 20th century, Bayesian analysis has become widely popular. "The subsequent explosion of interest in Bayesian statistics has led not only to extensive research in Bayesian methodology but also to the use of Bayesian methods to address pressing questions in diverse application areas such as astrophysics, weather forecasting, health care policy, and criminal justice." [8]

2.2 Bayes' Theorem

Before we introduce Bayesian Regression Analysis, it is important to first state Bayes' Theorem, the idea behind Bayesian Analysis. Bayes' Theorem is also known as Bayes' rule.

Bayes' rule was introduced by Reverend Thomas Bayes as a means for calculating conditional probabilities. Let S be a sample space and let A and B be two events in S. Denote the probability of A occurring by $P(A)$ and similarly for B, $P(B)$, where $P(A) \neq 0$ and $P(B) \neq 0$. The probability of an event A occurring given B is denoted by $P(A \mid B)$. The conditional probability formulas, $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ and $P(B \mid A) = \frac{P(A \cap B)}{P(A)}$, are helpful in deriving Bayes' rule. Substituting $P(A \cap B) = P(B \mid A)P(A)$ gives us

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \tag{2.1}$$

We refer to (2.1) as Bayes' rule.

In the case where A is a set of j mutually exclusive events, A_j, we use the law of total probability in the discrete case to calculate $P(B)$:

$$P(B) = \sum_j P(B \mid A_j)P(A_j).$$

So, given that event B has occurred, the posterior probability of any of these j events occurring is

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_j P(B \mid A_j)P(A_j)}.$$

What is meant by prior probability and posterior probability? The prior probability is an initial probability obtained prior to any additional information being obtained, denoted by $P(A)$. The posterior probability is a probability value that has been revised using additional information obtained later. We denote the posterior by $P(A \mid B)$. We refer to the additional information obtained as the likelihood and marginal likelihood denoted by $P(B \mid A)$ and $P(B)$, respectively. Thus, Bayes' rule can be written as

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}.$$

Bayes' rule can best be summed up by the following quote presented in a 2000 article in the Economist on the Bayesian Approach:

The essence of the Bayesian approach is to provide a mathematical rule explaining how you should change your existing beliefs in the light of new evidence. In other words, it allows scientists to combine new data with their existing knowledge or expertise. The canonical example is to imagine that a precocious newborn observes his first sunset, and wonders whether the sun will rise again or not. He assigns equal prior probabilities to both possible outcomes, and represents this by placing one white and one black marble into a bag. The following day, when the sunrises, the child places another white marble in the bag. The probability that a marble plucked randomly from the bag will be white (i.e., the child's degree of belief in future sunrises) has thus gone from a half to two-thirds. After sunrise the next day, the child adds another white marble, and the probability (and thus the degree of belief) goes from two-thirds to three-quarters. And so on. Gradually, the initial belief that the sun is just as likely as not to rise each morning is modified to become a near-certainty that the sun will always rise. [2]

2.3 Priors

The most controversial element of Bayesian Analysis is the use of a prior distribution. Criticized for introducing subjective information, the use of a prior is purely an educated guess and can vary from one scientist to another.

There are two types of priors. The first type, called the Conjugate Prior, occurs when the posterior distribution has the same form as the prior distribution. The second type, called the Noninformative Prior, is used when we have very little knowledge or information about the prior distribution. The noninformative prior is used to "conform to the Bayesian model in a correct parametric form." [13].

2.3.1 Conjugate Priors

To have a better understanding of conjugate priors, we use the following example.

Suppose X is a random variable with a beta distribution, that is, $X \sim \text{BETA}(a, b)$, where a and b are known. Then X has probability distribution function

$$f(x \mid \theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}, \tag{2.2}$$

where $0 < \theta < 1, a, b > 1$. Let $\Theta \sim \text{BIN}(n, \theta)$, where n is known. Then the distribution of Θ is given by

$$h(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x},$$

where $x = 0, 1, 2, \ldots, n$. Bayes' Rule states that

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{marginal likelihood}}.$$

Since the marginal likelihood is a constant we have that

$$k(\theta \mid x) \propto h(\theta) \times f(x \mid \theta),$$

or

$$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$

Therefore,

$$\begin{aligned} k(\theta \mid x) &\propto \theta^x (1-\theta)^{n-x} \theta^{a-1} (1-\theta)^{b-1} \\ &\propto \theta^{x+a-1} (1-\theta)^{n+b-x-1} \end{aligned}$$

Let $A = x + a$ and $B = n + b - x$. Then

$$k(\theta \mid x) \propto \theta^{A-1} (1-\theta)^{B-1}. \tag{2.3}$$

Prior Distribution	Likelihood	Posterior Distribution
Univariate Models		
Beta	Binomial	Beta
Gamma	Poisson	Gamma
Gamma	Exponential	Gamma
Normal	Normal (known variance, unknown mean)	Normal
Multivariate Models		
Dirichlet	Multinomial	Dirichlet
Multivariate Normal	Multivariate Normal (known variance matrix)	Multivariate Normal

Table 2.1 – Commonly used conjugate priors [13]

We call $k(\theta \mid x)$ in (2.3) the conjugate prior since its distribution is similar to the prior distribution in (2.2).

The conjugate prior simplifies the mathematics involved in Bayesian data analysis. Some commonly used conjugate priors are listed in Table 2.1.

2.3.2 Noninformative Prior

In the event that we have very little knowledge of the data set,we opt to use a noninformative prior. The question then becomes,"How do I choose the prior?" The answer is quite simple. The conservative, and most likely easiest, approach to choosing priors is to assume very little is known about a parameter and to use a noninformative prior. The most common choice is a Uniform prior, "a flat distribution that assigns the same probability to every value of the parameter." [14] There arise a few problems, however, if the parameter is infinite in range and lacks

the ability to be described by a "proper prior." This problem is not necessarily serious if the likelihood has a bounded range such that the function is nonzero.

Rather than using a Uniform prior we may opt to use a BETA(1,1) prior, which has the same result as a Uniform prior would. In fact, the BETA(1,1) prior is a special case of the Uniform prior. In the event that we have a multivariate model, we may choose to use the Dirichlet model with prior parameters set to 1. This is the multivariate version of the BETA(1,1) model.

2.4 Estimating the Regression Line

2.4.1 Simple Linear Regression

Suppose we wish to estimate the regression line for a data set involving the random variables x_i and y_i, where x_i denotes the ith observation on the independent variable and y_i denotes the ith observation on the dependent variable. We may use the model

$$y_i \mid x_i = \alpha + \beta x_i + e_i$$

where, α and β are unknown parameters. Assume, in order to obtain a third unknown parameter, that the e_i are independent and normally distributed. That is, $e_i \sim \mathrm{N}(0, \sigma^2)$. Now we have the unknown parameters α, β, and σ^2.

Let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

and

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

Then, the likelihood function becomes

$$f(\mathbf{x}, \mathbf{y} \mid \alpha, \beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right\}$$
$$\propto \frac{1}{\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right\}, \tag{2.4}$$

where we ignore the proportionality constant $(2\pi)^{-\frac{n}{2}}$ in equation (2.4).

Choose a noninformative prior density, $h(\alpha, \beta, \sigma)$, so that

$$h(\alpha, \beta, \sigma) = h_1(\alpha)h_2(\beta)h_3(\sigma), \tag{2.5}$$

where $h_1(\alpha) \propto$ constant, $h_2(\beta) \propto$ constant, and $h_3(\sigma) \propto \frac{1}{\sigma}$.

Now that we have the likelihood function and the prior density we can use Bayes' Theorem, $k(\alpha, \beta, \sigma \mid \mathbf{x}, \mathbf{y}) = h(\alpha, \beta, \sigma) f(\mathbf{x}, \mathbf{y} \mid \alpha, \beta, \sigma)$, to find the posterior density. That is, multiply equations (2.4) and (2.5) to get

$$k(\alpha, \beta, \sigma \mid \mathbf{x}, \mathbf{y}) \propto \frac{1}{\sigma}\frac{1}{\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right\}$$
$$= \frac{1}{\sigma^{n+1}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right\} \tag{2.6}$$

We take the estimates of α and β to be $E(\hat{\alpha})$ and $E(\hat{\beta})$ under the posterior distribution. Since these estimates are unbiased, that is, $E(\hat{\alpha}) = \alpha$ and $E(\hat{\beta}) = \beta$, then they are the same as the least-squares estimates discussed in Theorem 1.1. That is,

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}, \quad \hat{\beta} = \frac{S_{xy}}{S_{xx}},$$

where $\overline{x} = \frac{1}{n}\sum x_i, \overline{y} = \frac{1}{n}\sum y_i, S_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y})$, and $S_{xx} = \sum(x_i - \overline{x})^2$. An unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-2}\sum(y_i - \hat{\alpha} - \hat{\beta}x_i)^2.$$

Before continuing in our calculations, it is first important to note that

$$\begin{aligned}\sum_{i=1}^{n}(\alpha-\hat{\alpha})(y_i-\hat{\alpha}-\hat{\beta}x_i) &= \sum_{i=1}^{n}(\alpha-\hat{\alpha})(y_i-\overline{y}+\hat{\beta}\overline{x}-\hat{\beta}x_i)\\ &= (\alpha-\hat{\alpha})(n\overline{y}-n\overline{y}+n\hat{\beta}\overline{x}-n\hat{\beta}\overline{x})\\ &= (\alpha-\hat{\alpha})0\\ &= 0\end{aligned}$$

Hence,

$$\sum_{i=1}^{n}(\alpha-\hat{\alpha})(y_i-\hat{\alpha}-\hat{\beta}x_i)=0. \tag{2.7}$$

We also note that, given the result of (2.7), we have that

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)(\beta-\hat{\beta}x_i) &= \beta\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)-\hat{\beta}\sum_{i=1}^{n}x_i(y_i-\hat{\alpha}-\hat{\beta}x_i)\\ &= 0-\hat{\beta}\sum_{i=1}^{n}x_i(y_i-\hat{\alpha}-\hat{\beta}x_i)x_i \qquad \text{(by 2.7)}\\ &= -\hat{\beta}\sum_{i=1}^{n}(y_i-\overline{y}+\hat{\beta}\overline{x}-\hat{\beta}x_i)x_i\\ &= -\hat{\beta}\left[\sum_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}-\hat{\beta}\sum_{i=1}^{n}(x_i-\overline{x})x_i\right]\\ &= -\hat{\beta}\left[\sum_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}-\hat{\beta}\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{)}\right]\\ &= -\hat{\beta}\left[\sum_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}-\hat{\beta}S_{xx}\right]\\ &= -\hat{\beta}\left[\sum_{i=1}^{n}x_i-y_i-n\overline{x}\,\overline{y}-\frac{S_{xy}}{S_{xx}}S_{xx}\right]\\ &= -\hat{\beta}\left[\sum_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}-\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})\right]\\ &= -\hat{\beta}\left[\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})-\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})\right]\\ &= 0\end{aligned}$$

Hence,

$$\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)(\beta - \hat{\beta})x_i = 0 \tag{2.8}$$

Rewrite $\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$ to get

$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 = \sum_{i=1}^{n}\Big[(y_i - \hat{\alpha} - \hat{\beta}x_i) - (\alpha - \hat{\alpha}) - (\beta - \hat{\beta})x_i\Big]^2.$$

Evaluating, we see that

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 &= \sum_{i=1}^{n}\Big[(y_i - \hat{\alpha} - \hat{\beta}x_i) - (\alpha - \hat{\alpha}) - (\beta - \hat{\beta})x_i\Big]^2 \\
&= \sum_{i=1}^{n}\Big[(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 + (\alpha - \hat{\alpha})^2 + (\beta - \hat{\beta})^2 x_i^2 \\
&\quad +2(\alpha - \hat{\alpha})(\beta - \hat{\beta})x_i \\
&\quad -2(y_i - \hat{\alpha} - \hat{\beta}x_i)(\alpha - \hat{\alpha}) \\
&\quad -2(y_i - \hat{\alpha} - \hat{\beta}x_i)(\beta - \hat{\beta})x_i\Big] \\
&= \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 + n(\alpha - \hat{\alpha})^2 + \sum_{i=1}^{n}(\beta - \hat{\beta})^2 x_i^2 \\
&\quad + 2(\alpha - \hat{\alpha})(\beta - \hat{\beta})\sum_{i=1}^{n} x_i \\
&= (n-2)s^2 + n(\alpha - \hat{\alpha})^2 + (\beta - \hat{\beta})^2 \sum_{i=1}^{n} x_i^2 \\
&\quad + 2(\alpha - \hat{\alpha})(\beta - \hat{\beta})\sum_{i=1}^{n} x_i
\end{aligned}$$

Let $\ell = \frac{1}{2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$ be a nonnegative constant. Integrating $k(\alpha, \beta, \sigma \,|\, \mathbf{x}, \mathbf{y})$ with respect to σ, we have

$$k_1(\alpha, \beta \,|\, \mathbf{x}, \mathbf{y}) = \int_0^{\infty} \frac{1}{\sigma^{n+1}} e^{-\frac{\ell}{\sigma^2}}\, d\sigma$$

where $\ell > 0$. Let $z = \ell\sigma^{-2}$. Then $d\sigma = -\frac{1}{2\ell}\sigma^3\, dz$. Then

$$\begin{aligned}
\int_0^\infty \frac{1}{\sigma^{n+1}} e^{-\frac{\ell}{\sigma^2}}\, d\sigma &= \int_0^\infty \frac{\sigma^3}{\sigma^{n+1}} \left(-\frac{1}{2}\right)\left(\frac{1}{\ell}\right) e^{-z}\, dz \\
&= \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}-1} \left(-\frac{1}{2}\right)\left(\frac{1}{\ell}\right) e^{-z}\, dz \\
&= \int_0^\infty z^{\frac{n}{2}-1} \left(\frac{1}{\ell}\right)^{\frac{n}{2}-1} \left(\frac{1}{\ell}\right)\left(-\frac{1}{2}\right) e^{-z}\, dz \\
&= \left(-\frac{1}{2}\right)\left(\frac{1}{\ell}\right)^{\frac{n}{2}} \int_0^\infty z^{\frac{n}{2}-1} e^{-z}\, dz \\
&= \left(-\frac{1}{2}\right) \ell^{-\frac{n}{2}} \Gamma\left(\frac{n}{2}\right) \\
&\propto \ell^{-\frac{n}{2}}
\end{aligned}$$

Thus we have,

$$\begin{aligned}
k_1(\alpha, \beta \,|\, \mathbf{x}, \mathbf{y}) = \Bigg[&(n-2)s^2 + n(\alpha - \hat{\alpha})^2 + (\beta - \hat{\beta})^2 \sum_{i=1}^n x_i^2 \\
&+2(\alpha - \hat{\alpha})(\beta - \hat{\beta}) \sum_{i=1}^n x_i \Bigg]^{-\frac{n}{2}}. \qquad (2.9)
\end{aligned}$$

Together, both α and β follow a bivariate Student's $t-$distribution. The marginal distribution of α follows a univariate Student's $t-$posterior distribution given by

$$(\alpha - \hat{\alpha}) \left[\frac{S_{xx}}{\frac{s^2}{n}\sum x_i^2}\right]^{\frac{1}{2}} \sim t(n-2). \qquad (2.10)$$

The marginal distribution of β follows a univariate Student's $t-$ posterior distribution given by

$$\frac{\beta - \hat{\beta}}{\left(\frac{s}{[S_{xx}]^{\frac{1}{2}}}\right)} \sim t(n-2). \qquad (2.11)$$

Unlike frequentist statistics, in Bayesian data analysis, we use a credibility statement instead of a confidence interval. In (2.10), let $\xi = \left[\frac{S_{xx}}{\frac{s^2}{n}\sum x_i^2}\right]^{\frac{1}{2}}$. Then the credibility statement is

$$P\left(\hat{\alpha} - \frac{z_\gamma}{\xi} \leq \alpha \leq \hat{\alpha} + \frac{z_\gamma}{\xi}\right) = 1 - 2\gamma,$$

where z_γ is the $\gamma-$fractile point of the Student's t-distribution. Similarly, in (2.11), let $\delta = \frac{s}{[S_{xx}]^{\frac{1}{2}}}$. Then the credibility statement is

$$P\left(\hat{\beta} - \delta z_\gamma \leq \beta \leq \hat{\beta} + \delta z_\gamma\right) = 1 - 2\gamma.$$

Suppose we are asked to predict the outcome of $y^\star$ given a new observation $x^\star$. From our previous calculations, the model for the new point is

$$y^\star \mid x^\star = \alpha + \beta x^\star + e,$$

where $e \sim \mathrm{N}(0, \sigma^2)$. In order to make a prediction, we calculate a predictive density, given that the parameters have already been estimated.

DEFINITION 2.1. *Let $X_1, \ldots, X_n$ be a set of random variables with a density $f(x_1, \ldots, x_n \mid \theta)$. For the prior, $h(\theta)$, we obtain the posterior*

$$k(\theta \mid x_1, \ldots, x_n) \propto \prod_{i=1}^{n} f(x_i \mid \theta) h(\theta).$$

If we wish to predict a new observation $y^\star$, we utilize a predictive density for $y^\star$ given by

$$p(y^\star \mid x_1^\star, \ldots, x_n^\star) = \int f(y^\star \mid \theta) k(\theta \mid x_1^\star, \ldots, x_n^\star)\, d\theta.$$

[11]

Thus, the predictive density of $y^\star$ is given by

$$\begin{aligned} p(y^\star \mid x^\star) &= \iiint f(y^\star \mid x^\star, \alpha, \beta, \sigma) k(\alpha, \beta, \sigma \mid \mathbf{x}, \mathbf{y})\, d\alpha\, d\beta\, d\sigma \\ &= \iiint \frac{1}{\sigma^{2n+1}} \exp\left\{ -\frac{1}{2\sigma^2} \left[\sum_{i=1}^{n} (y^\star - \alpha - \beta x^\star)^2 \right.\right. \\ &\quad \left.\left. + \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \right] \right\} d\alpha\, d\beta\, d\sigma. \end{aligned} \tag{2.12}$$

Using similar techniques as above to calculate the two-dimensional posterior density, we have that

$$\frac{y^\star - \hat{\alpha} - \hat{\beta} x^\star}{\hat{\sigma}\left[1 + \frac{1}{n} + \frac{(x^\star - \overline{x})^2}{S_{xx}}\right]^{\frac{1}{2}}} \sim t(n-2). \tag{2.13}$$

2.4.2 Multiple Linear Regression

The word "multiple" indicates that there is more than one independent variable in the regression model. As seen in Section 2.4.1, the calculations can become increasingly difficult. Similar to Section 1.3, we utilize matrix notation to simplify our calculations. However, this does not aid in completely simplifying complicated calculations. Instead, with the use of a computing tool, such as R in combination with WinBUGS (Windows version of Bayesian inference Using Gibbs Sampling), we can utilize Markov Chain Monte Carlo (MCMC) methods to assist in these calculations. We outline this technique in Section 2.5.

2.5 Markov Chain Monte Carlo (MCMC) Methods

As the number of variables increases in our models, the more difficult it becomes to evaluate and analyze the solution of a posterior distribution. Here is where the MCMC Methods become quite useful. MCMC techniques simulate the posterior so that it can be analyzed. The results can then be used to draw inferences about the models and parameters. There are many MCMC algorithms with which to choose.

2.5.1 Gibbs Sampling

Gibbs Sampling is one such algorithm and is especially useful in applications of Bayesian analysis. The Gibbs sampler is a technique that generates random variables indirectly from a distribution without having to calculate the density. Thus, we are able to create a sequence of easier calculations while avoiding the much more difficult ones.

The main idea of the Gibbs sampler is to fix all values of the random variables,

save for one. In other words, we consider univariate conditional distributions.

Suppose we wish to obtain the marginal density,
$f(x) = \int \cdots \int f(x, y_1, \ldots, y_n)\, dy_1 \cdots dy_n$, of a given joint density $f(x, y_1, y_2, \ldots, y_n)$ in order to obtain the mean or variance. As we are often programmed to do, the first natural instinct would be to calculate $f(x)$ and obtain the desired information. In many instances, however, the calculations of $f(x)$ can become quite complicated and an alternative method is needed. The Gibbs sampler, the alternative method, generates a random sample $X_1, \ldots, X_n \sim f(x)$ without requiring $f(x)$. Any characteristics drawn from $f(x)$ can be calculated to the desired accuracy with a large enough simulation sample.

As an example, consider a pair of random variables (X, Y). Rather than sampling from $f(x)$, the Gibbs sampler will generate a sampling from the conditional distributions $f(x \mid y)$ and $f(y \mid x)$. Given the initial value $Y_0' = y_0'$, we can obtain a "Gibbs sequence" [6] of random variables

$$Y_0', X_0', Y_1', X_1', Y_2', X_2', \ldots, Y_k', X_k' \tag{2.14}$$

by alternately generating values from

$$\begin{aligned} X_j' &\sim f(x \mid Y_j' = y_j') \\ Y_j' &\sim f(y \mid X_j' = x_j') \end{aligned} \tag{2.15}$$

We call the generation of (2.14) the Gibbs sampler. It turns out that, as $k \to \infty$, the distribution of X_k' tends to $f(x)$. For a large k, the value $X_k' = x_k'$ is a point from $f(x)$.

For the frequentist statistician, the Gibbs sampler can be used to calculate the likelihood functions and characteristics of likelihood estimators. For the Bayesian statistician, the main use of the Gibbs sampler is to generate the posterior distributions.

As the dimension of a problem increases, the Gibbs sampler becomes more useful since it allows us to avoid "prohibitively difficult" integrals in high dimensions.

CHAPTER 3
AN EXAMPLE

As seen in Chapters 1 and 2, the Frequentist and Bayesian data analysis methods are quite different. In this chapter we will look at a simple linear regression example as well as a multiple linear regression example.

3.1 Simple Linear Regression

Consider a set of fifteen women between the ages of 30 and 39 in Table 3.1. For each subject, her height and mass was recorded.

A plot of the data points, where x is the height of the women in meters and y is the mass of the women in kilograms, is shown in Figure 3.1 . The data set is linear so we may apply the simple linear regression model

$$y_i = \alpha + \beta x_i + e_i.$$

Table 3.1 – Average mass of women age 30-39 as a function of their height [1]

Height meters(m) x	Mass kilograms(kg) y
1.47	52.21
1.50	53.12
1.52	54.48
1.55	55.84
1.57	57.20
1.60	58.57
1.63	59.93
1.65	61.29
1.68	63.11
1.70	64.47
1.73	66.28
1.75	68.10
1.78	69.92
1.80	72.19
1.83	74.46

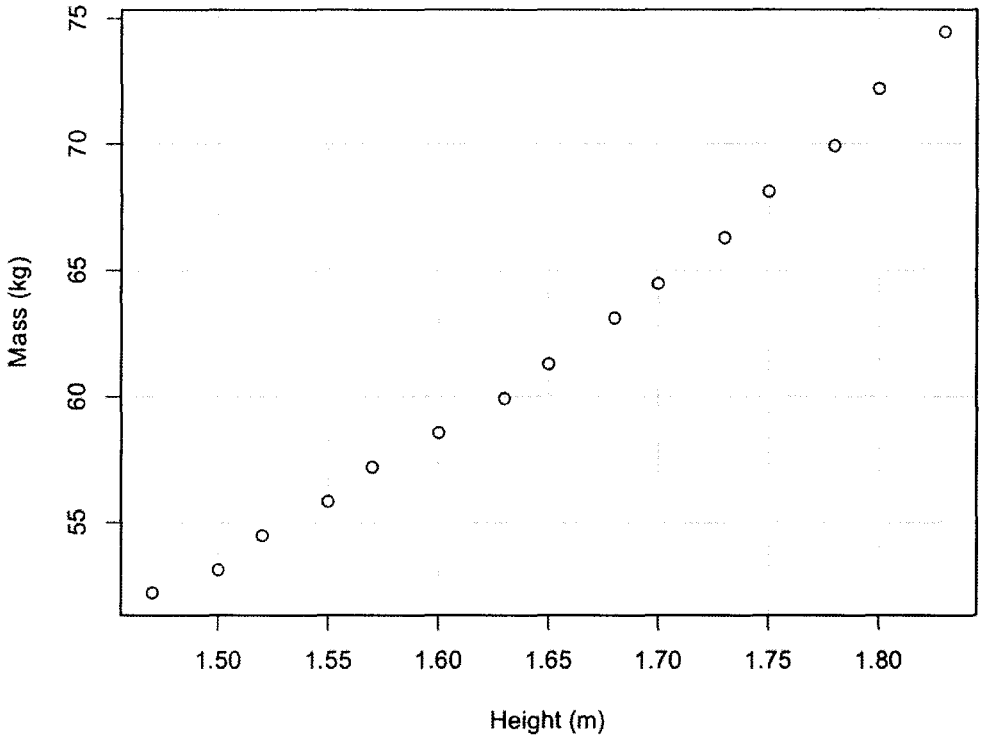

Figure 3.1 – Plot of the average mass of women age 30-39 as a function of their height data

3.1.1 Frequentist Methods

To perform the necessary calculations, we use R, a statistical computing software [15], for assistance. We obtain the following:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{15} x_i = 1.650667 \tag{3.1}$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{15} y_i = 62.078 \tag{3.2}$$

$$S_{xx} = \sum_{i=1}^{15}(x_i - \overline{x})^2 = 0.1826933 \tag{3.3}$$

$$S_{xy} = \sum_{i=1}^{15} = (x_i - \overline{x})(y_i - \overline{y}) = 11.19402 \tag{3.4}$$

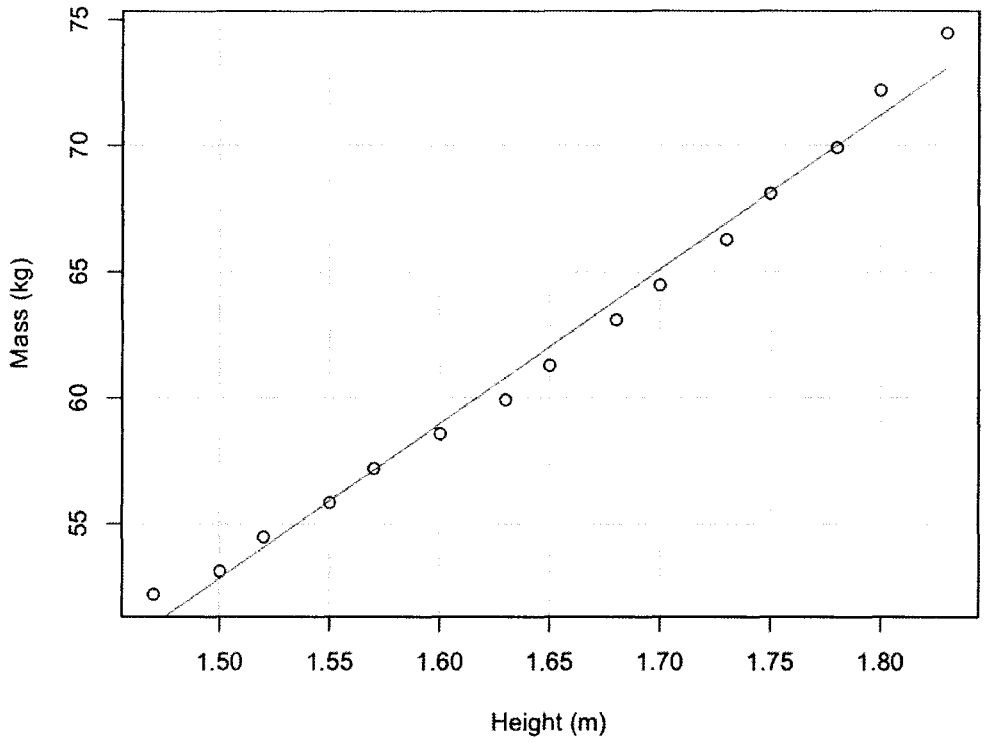

Figure 3.2 – Plot of the average mass of women age 30-39 as a function of their height data with estimated regression line

Substituting these values into the equations (1.4) and (1.5), we have that

$$\hat{\alpha} = \overline{y} - \frac{S_{xy}}{S_{xx}}\overline{x} = -39.06196 \tag{3.5}$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = 61.27219 \tag{3.6}$$

Thus, the regression line can be estimated by

$$\hat{y}_i = -39.06196 + 61.27219x_i. \tag{3.7}$$

We can see in Figure 3.2 that the sum of the squared residuals (SSR) is minimal. Given a large data set, we can use R to perform these calculations as seen in Figure 3.3.

Given a 95% significance level, we "Reject $H_0 : \beta = 0$ if $|t| > 2.160$." In order to calculate the confidence interval, we need to first calculate s. That is,

$$s = \sqrt{\frac{SSR}{13}} = 0.7590763. \tag{3.8}$$

```
Call:
lm(formula = yvar ~ xvar, data = Wikipedia)

Residuals:
     Min       1Q   Median       3Q      Max
-0.88171 -0.64484 -0.06993  0.34095  1.39385

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -39.062      2.938  -13.29 6.05e-09 ***
xvar          61.272      1.776   34.50 3.60e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7591 on 13 degrees of freedom
Multiple R-squared: 0.9892, Adjusted R-squared: 0.9884
F-statistic:  1190 on 1 and 13 DF,  p-value: 3.604e-14
```

Figure 3.3 – R output of Average mass of women data

Thus, with confidence level 95%, β lies in the interval

$$[57.436, 65.108].$$

3.1.2 Bayesian Methods

Given that we have no prior knowledge of the data in Table 3.1, we will use a uniform prior distribution

$$h(\alpha, \beta, \sigma) \propto \frac{1}{\sigma} \tag{3.9}$$

Thus, we can use the likelihood function in equation (2.4).

Since the estimates of the parameters are unbiased, we can use the least squares estimates found in (3.5) and (3.6). As well, we can use the values found in (3.1) and (3.8). The marginal posterior densities are given in Figures 3.4 and 3.5. The posterior means are approximately equal to the least-squares estimates of the parameters and are given by

$$E(\alpha \,|\, \mathbf{y}, \mathbf{x}) = -39.05, \qquad E(\beta \,|\, \mathbf{y}, \mathbf{x}) = 61.26. \tag{3.10}$$

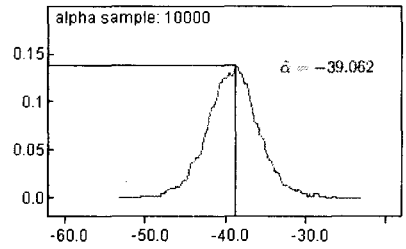

Figure 3.4 – Posterior density of α.

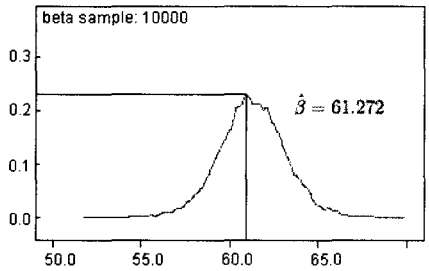

Figure 3.5 – Posterior density of β.

We compute these values using the 3-step method seen in Figure 3.7 to obtain the data in Figures 3.4, 3.5, and 3.6 using WinBUGS [7].

Given γ, the credibility statement for α and β is as follows:

$$P\left(-39.062 - \frac{z_\gamma}{0.340} \leq \alpha \leq -39.062 + \frac{z_\gamma}{0.340}\right) = 1 - 2\gamma \tag{3.11}$$

$$P\left(61.272 - 1.776z_\gamma \leq \beta \leq 61.272 + 1.776z_\gamma\right) = 1 - 2\gamma. \tag{3.12}$$

Unlike frequentist methods, the credibility statements can be used to predict the true values of α and β.

3.2 Multiple Linear Regression

node	mean	sd	2.5%	median	97.5%	sample
alpha	-39.05	3.205	-45.35	-39.02	-32.67	10000
beta	61.26	1.937	57.42	61.25	65.06	10000

Figure 3.6 - WinBUGS posterior data of parameters α and β.

```
1) Program code
model
{
for(i in 1:n)
{y[i] ~ dnorm(mu[i],tau)
mu[i ]<- alpha+beta*x[i]
}
alpha ~ dnorm(0,0.000000000001)
beta ~ dnorm(0,0.000000000001)
tau ~ dgamma(0.001,0.001)
sigma <- 1/sqrt(tau)
}
2. Data
list(x=c(1.47,1.50,1.52,1.55,1.57,
1.60,1.63,1.65,1.68,1.70,1.73,
1.75,1.78,1.80,1.83),
y=c(52.21,53.12,54.48,55.84,
57.20,58.57,59.93,61.29,63.11,
64.47,66.28,68.10,69.92,72.19,74.46),
n=15)
3. Initial values
list(alpha = 0, beta = 0, tau = 1)
```

Figure 3.7 – WinBUGS data input.

Consider a set of seventeen samples taken by a chemical analyst who expects the yield to be affected by two factors, x_1 and x_2. Table 3.2 lists the values recorded. A plot of the data points, where x_1 and x_2 are the two Factors and y is the yield, is shown in Figure 3.8. The data set is linear so we may apply the regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i.$$

Table 3.2 – Chemical Analysis Data [3]

Factor 1	Factor 2	Yield
x_1	x_2	y
41.9	29.1	251.3
43.4	29.3	251.3
43.9	29.5	248.3
44.5	29.7	267.5
47.3	29.9	273.0
47.5	30.3	276.5
47.9	30.5	270.3
50.2	30.7	274.9
52.8	30.8	285.0
53.2	30.9	290.0
56.7	31.5	297.0
57.0	31.7	302.5
63.5	31.9	304.5
65.3	32.0	309.3
71.1	32.1	321.7
77.0	32.5	330.7
77.8	32.9	349.0

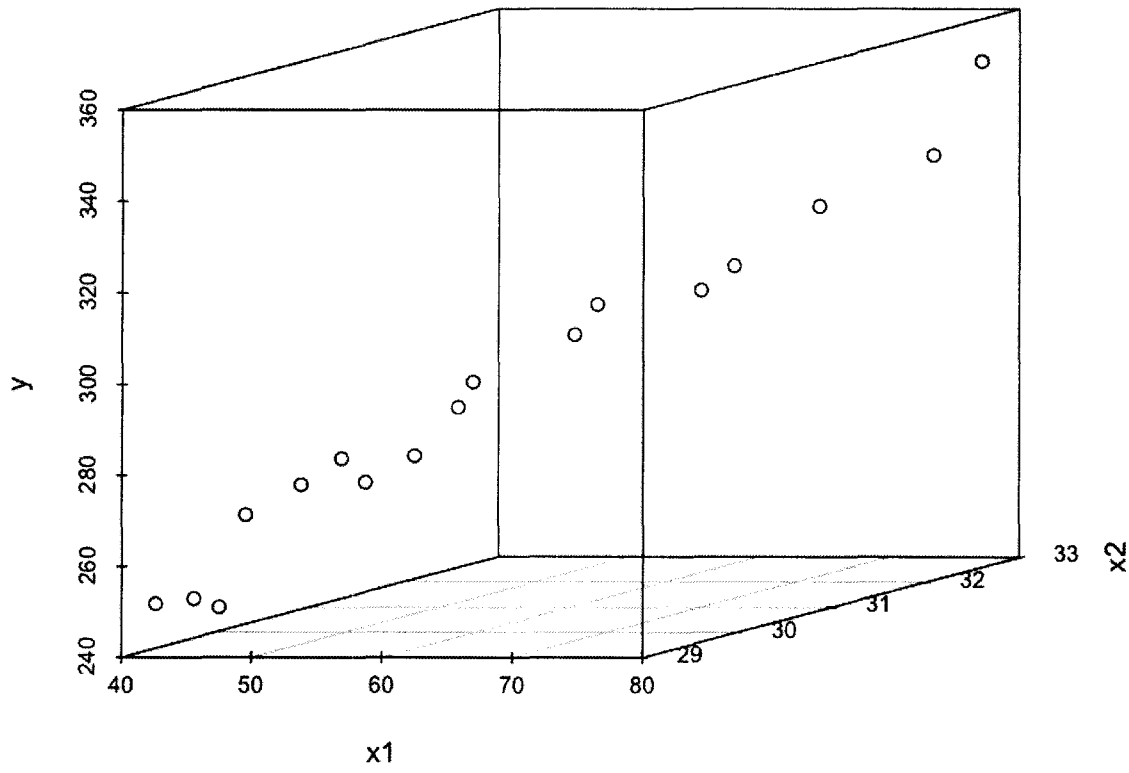

Figure 3.8 – Plot of chemical analysis data.

3.2.1 Frequentist methods

To simplify the calculations we use matrix notation. Let

$$\mathbf{X} = \begin{bmatrix} 1 & 41.9 & 29.1 \\ 1 & 43.4 & 29.3 \\ 1 & 43.9 & 29.5 \\ 1 & 44.5 & 29.7 \\ 1 & 47.3 & 29.9 \\ 1 & 47.5 & 30.3 \\ 1 & 47.9 & 30.5 \\ 1 & 50.2 & 30.7 \\ 1 & 52.8 & 30.8 \\ 1 & 53.2 & 30.9 \\ 1 & 56.7 & 31.5 \\ 1 & 57.0 & 31.7 \\ 1 & 63.5 & 31.9 \\ 1 & 65.3 & 32.0 \\ 1 & 71.1 & 32.1 \\ 1 & 77.0 & 32.5 \\ 1 & 77.8 & 32.9 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} 251.3 \\ 251.3 \\ 248.3 \\ 267.5 \\ 273.0 \\ 276.5 \\ 270.3 \\ 274.9 \\ 285.0 \\ 290.0 \\ 297.0 \\ 302.5 \\ 304.5 \\ 309.3 \\ 321.7 \\ 330.7 \\ 349.0 \end{bmatrix}$$

Then, the estimates of the parameters can be written as

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$$

Thus, using equation (1.7), we find that

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 336.521 & 1.228 & -13.089 \\ 1.228 & 0.005 & -0.049 \\ -13.089 & -0.049 & 0.511 \end{bmatrix} \begin{bmatrix} 4902.8 \\ 276614.4 \\ 152021.1 \end{bmatrix}$$
$$= \begin{bmatrix} -153.512 \\ 1.239 \\ 12.082 \end{bmatrix}.$$

Therefore,

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} -153.512 \\ 1.239 \\ 12.082 \end{bmatrix}. \tag{3.13}$$

We can see in Figure 3.9 that the SSR is minimal. Given a large data set, we can use R to perform these calculations as seen in Figure 3.10.

To test the significance of the parameter $\beta_1 = 0$ we perform a hypothesis test. We obtain the following values:

$$SSR_{full} = \sum_{i=1}^{17} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 = 423.374$$
$$SSR_{reduced} = \sum_{i=1}^{17} (y_i - \hat{\beta}_0 - \hat{\beta}_2 x_{i2})^2 = 83697.15$$
$$\hat{\sigma}^2 = \frac{SSR_{full}}{n-3} = 32.567.$$

Thus, the value of the test statistic F is

$$F = \frac{SSR_{reduced} - SSR_{full}}{\hat{\sigma}^2} = 2556.98 \tag{3.14}$$

Assuming normality of the distribution of the random errors, $F \sim F(1, 13)$. Therefore, with a 95% confidence level, we "Reject $H_0 : \beta_1 = 0$ if $F > 4.67$." In this case, we reject the null hypothesis, which indicates there is a significant relation between y_i and x_{i1}.

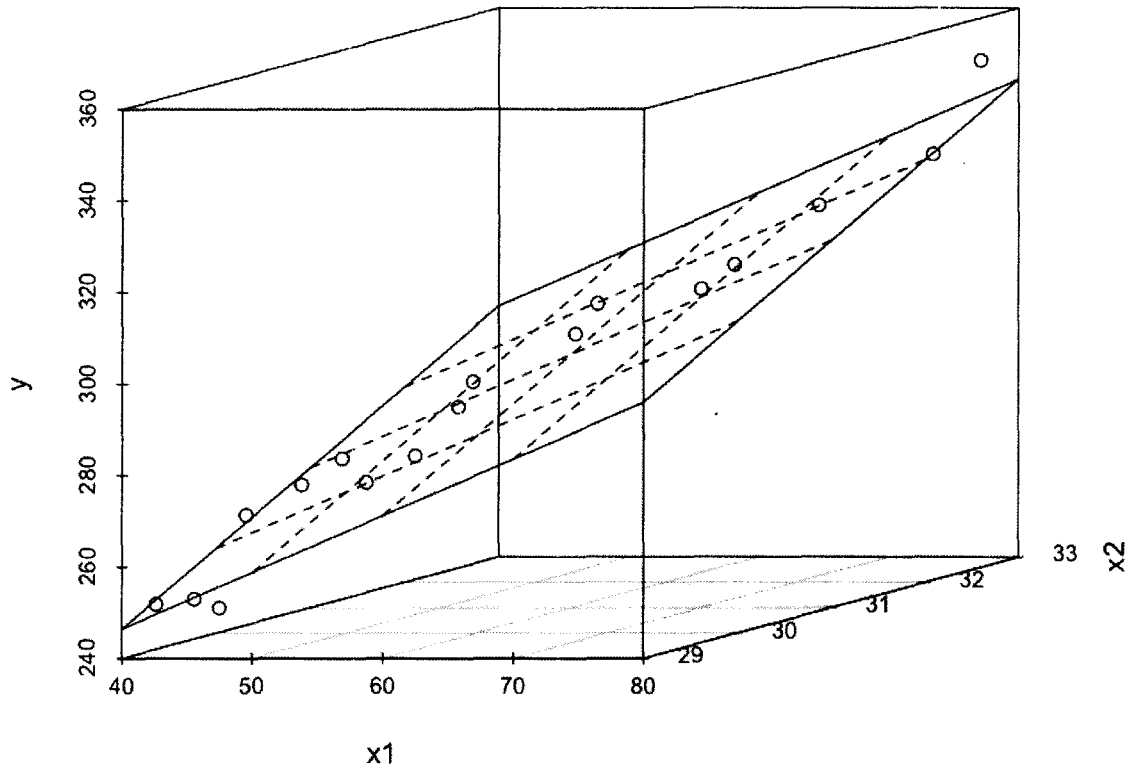

Figure 3.9 – Plot of chemical analysis data with estimated regression plane.

To construct the confidence interval, first we need to compute $C = \hat{\sigma}^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ as seen in (3.15).

$$C = \begin{bmatrix} 10959.571 & 40.002 & -426.274 \\ 40.002 & 0.168 & -1.595 \\ -426.274 & -1.595 & 16.652 \end{bmatrix} \tag{3.15}$$

Thus $C_{11} = 10959.571$. The 95% confidence interval is given by

$$[-224.887, 227.365]. \tag{3.16}$$

3.2.2 Bayesian methods

To simplify our calculations, we utililize WinBUGS and the program code in Figure 3.11. Since we have no prior knowledge of the data, we will use a uniform

```
Call:
lm(formula = y ~ x1 + x2, data = Chemical)

Residuals:
   Min     1Q Median     3Q    Max
-8.998 -4.035 -0.318  4.267  8.630

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -153.5117   100.8799  -1.522  0.15034
x1             1.2387     0.3946   3.139  0.00724 **
x2            12.0824     3.9323   3.073  0.00827 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.499 on 14 degrees of freedom
Multiple R-squared: 0.968,  Adjusted R-squared: 0.9635
F-statistic: 211.9 on 2 and 14 DF,  p-value: 3.419e-11
```

Figure 3.10 – R output of chemical analysis data

prior. In addition, we assume that the parameters have a normal distribution. After performing several iterations, we obtain the summary in Figure 3.12. As calculated in the frequentist methods, we find that the estimates for $\hat{\boldsymbol{\beta}}$ are given by (3.13).

We find that the marginal posteriors densities are given in Figures 3.13, 3.14, and 3.15. The posterior means are approximately equal to the least-squares estimates of the parameters and are given by

$$E(\beta_0 \,|\, \mathbf{y}, \mathbf{X}) = -153.4, \quad E(\beta_1 \,|\, \mathbf{y}, \mathbf{X}) = 1.238, \quad E(\beta_2 \,|\, \mathbf{y}, \mathbf{X}) = 12.08. \qquad (3.17)$$

```
1) Program code
model
{
for(i in 1:n)
{y[i] ~ dnorm(mu[i], tau)
mu[i]<- beta0 + beta1*x1[i] + beta2*x2[i]
}
beta0 ~ dnorm(0,0.000000000001)
beta1 ~ dnorm(0,0.000000000001)
beta2 ~ dnorm(0,0.000000000001)
tau ~ dgamma(0.001,0.001)
sigma<-1/sqrt(tau)
}
2. Data
list(x1=c(41.9,43.4,43.9,44.5,47.3,47.5,
47.9,50.2,52.8,53.2,56.7,57.0,63.5,65.3,
71.1,77.0,77.8),
x2=c(29.1,29.3,29.5,29.7,29.9,30.3,30.5,
30.7,30.8,30.9,31.5,31.7,31.9,32.0,32.1,32.5,32.9),
y=c(251.3,251.3,248.3,267.5,273.0,276.5,
270.3,274.9,285.0,290.0,297.0,302.5,304.5,
309.3,321.7,330.7,349.0),
n=17)
3. Initial values
list(beta0=0, beta1=0, beta2=0, tau=1)
```

Figure 3.11 – WinBUGS program code

node	mean	sd	2.5%	median	97.5%	sample
beta0	-153.4	109.4	-370.4	-154.2	65.91	20000
beta1	1.238	0.4291	0.3879	1.238	2.086	20000
beta2	12.08	4.268	3.55	12.08	20.38	20000

Figure 3.12 – Data summary in WinBUGS

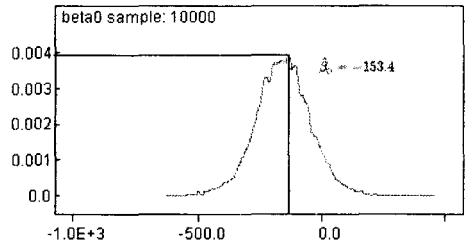

Figure 3.13 – Posterior density of β_0

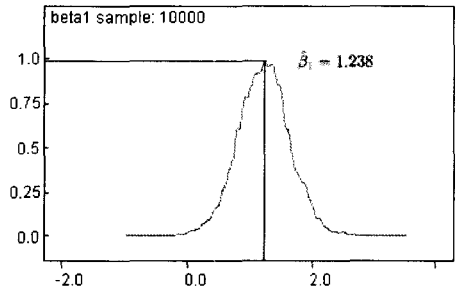

Figure 3.14 – Posterior density of β_1

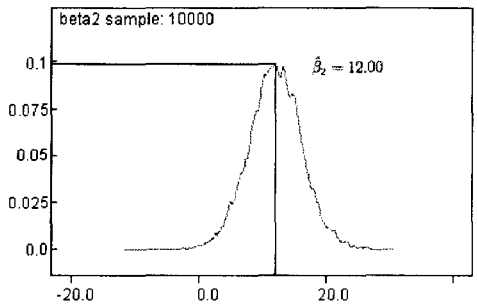

Figure 3.15 – Posterior density of β_2

CHAPTER 4
FREQUENTIST v. BAYESIAN

In the early 20th century, Frequentist methods of data analysis became more and more popular, while Bayesian methods were tabled until proper advancements had been made to handle the increasing difficulty of the calculations. In the late 20th and early 21st centuries, Bayesian methods have rapidly become more popular with the use and availability of computers and new computational methods. Now the question becomes which is better? Or is one truly better than the other? How do we choose?

Frequentists will argue that Bayesian methods are not suitable and cannot be reproduced from one statistician to another as the choice of the prior is purely a guessing game. Bayesians, on the other hand, will argue that Bayesian methods are much more accurate than the frequentist methods since the main focus lies on the posterior distribution and the characteristics that can be drawn from it.

4.1 Frequentist Methods - The Method of Least Squares

The biggest advantage to the Method of Least Squares is its simplicity in the calculations. As well, by definition, the estimates produced give the smallest sum of squared residuals (SSR). "The least-squares estimate of β_j is the best linear unbiased estimate (BLUE). That is, suppose we want to estimate β_j by a linear combination $a_1y_1 + \cdots + a_ny_n$ of the observed response variables and suppose we want the estimate to be unbiased. Among all linear unbiased estimates of β_j, the

one with the smallest variance is the least-squares estimate." [5]

The biggest disadvantage to the Method of Least Squares is that to obtain optimality one must assume the random errors have a normal distribution or attention should be restricted only to linear estimates. "When the distribution of the errors is not normal, least-squares estimates and tests may lose much of their efficiency. A few distant outliers can cause least-squares procedures to perform quite poorly." [5]

4.2 Bayesian Methods

Without any knowledge of prior information, it may seem pointless to use Bayesian techniques. However, with some prior information available it seems pointless not to use it.

> The Bayesian approach can also be regarded as being more satisfactory than the classical approach in that it produces a direct probability statement about a parameter, or hypothesis, as opposed to the somewhat awkward notions of confidence level or p-value, which are frequently misinterpreted by nonprofessional statistical users. It could also be viewed as an advantage that Bayesian analysis allows one to interpret a probability as a measure of degree of belief concerning the actual observed data rather than as a long-run frequency involving hypothetical observations that might have been obtained but were not. Moreover, the Bayesian approach has the appealing feature that it provides a unified, fairly straightforward way to analyze any statistical problem.
>
> Critics of the Bayesian method say that it is too subjective, especially with regard to the choice of a prior distribution for the parameters. In response to this criticism, it can be said that statistical analysis cannot

avoid being subjective, for example, in the choice of the model, and that the data analyst should admit his or her subjectivity and explicitly include it in the analysis. In doing this, however, one runs into the difficulty of quantifying one's prior knowledge and beliefs in the form of a prior distribution. Another difficulty arises if the prior distribution is not restricted to have a mathematically convenient form, because then the computation of the posterior distribution can be unwieldy. [5]

4.3 Conclusion

In the end, which method is better? Ultimately, it is up to the statistician to choose which method he or she prefers to use based on any prior knowledge of the data. While the Method of Least Squares seems ideal because of its simple calculations, the Bayesian approach provides a "direct probability statement about the parameter." Both methods can be equally criticized for the flaws that each possess. Neither method is "better" than the other, it all depends on the prior knowledge of the data and the decision of the statistician as to which method he or she uses.

REFERENCES

[1] *Simple Linear Regression Numerical Example*, http://en.wikipedia.org/wiki/Simple_linear_regression#Numerical_example.

[2] *In praise of Bayes*, http://www.economist.com/node/382968, September 2000.

[3] *Estimating Regression Models Using Least Squares*, http://www.weibull.com/DOEWeb/estimating_regression_models_using_least_squares.htm, 2008.

[4] H. Anton and C. Rorres, *Elementary Linear Algebra with Applications*, John Wiley and Sons, Inc, 1987.

[5] D. Birkes and Y. Dodge, *Alternative Methods of Regression*, John Wiley and Sons, Inc, 1993.

[6] G. Casella and E. George, *Explaining the Gibbs Sampler*, The American Statistician **46** (1992), no. 3, 167–174.

[7] N. Best D.J. Lunn, A. Thomas and D. Spiegelhalter, *WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility*, Statistics and Computing **10** (2000), 325–337.

[8] T. O'Hagan K. Cowles, R. Kass, *What is Bayesian Analysis?*, http://bayesian.org/Bayes-Explained.

[9] M. Parker, *Foundations of Statistics - Frequentist and Bayesian*, `http://www.austincc.edu/mparker/stat/nov04/talk_nov04.pdf`, November 2004.

[10] W.R. Pestman, *Mathematical Statistics*, second ed., Walter de Gruyter GmbH and Co. KG, Berlin, Germany, 2009.

[11] S.J. Press, *Bayesian Statistics: Principles, Models, and Applications*, John Wiley and Sons, Inc, 1989.

[12] P. Sahoo, *Probability and Mathematical Statistics*, Department of Mathematics University of Louisville, 2006.

[13] K. Seefeld, *Statistics Using R with Biological Examples*, University of New Hampshire, Durham, NH, Department of Mathematics and Statistics, 2007.

[14] H.B. Stauffer, *Contemporary Bayesian and Frequentist Statistical Research Methods for Natural Resource Scientists*, John Wiley and Sons, Inc, 2008.

[15] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2011.

CURRICULUM VITAE

SARA E. EVANS
111 Wickfield Dr
Louisville, KY 40245
(502) 439-4332
sara.elizabeth.evans@gmail.com

EDUCATION

University of Louisville, Louisville, KY
Master of Arts, Mathematics; May 2012

Murray State University, Murray, KY
Bachelor of Science, Applied Physics; May 2008

PUBLICATIONS

Arthur Pallone, Melissa Addessi, and Sara Evans. *Indices for Measurement of Seeing Quality by Low-Cost Camcorder Imaging. Journal of the Royal Astronomical Society of Canada*, Vol. 52, Issue 6, p.225.