ABSTRACT

Title of thesis: BAYESIAN BINOMIAL REGRESSION:
PROBABILITY OF SUCCESS IN PRO FOOTBALL

Albert de Hombre, Master of Arts, 2006

Thesis directed by: Professor Benjamin Kedem
Department of Mathematics

Bayesian regression provides a flexible alternative to standard GLM, as it allows for more user control of the data. The Bayesian may use his or her prior knowledge to influence the outcome of an experiment. While GLM is a robust tool that applies to virtually any type of data, there is little room for manipulation on the part of the user, who may have some additional "expert" knowledge beyond the raw data. The assumptions of both practices are outlined and applied to Binomial data acquired from the 2005 NFL season to highlight the advantages and disadvantages of each approach. Under ideal conditions the Bayesian analysis provides more accurate estimators and features relatively simple computations.

BAYESIAN BINOMIAL REGRESSION: PROBABILITY OF SUCCESS IN PRO FOOTBALL

By

Albert de Hombre

Thesis submitted to the Faculty of the Graduate School of the

University of Maryland, College Park, in partial fulfillment

of the requirements for the degree of

Master of Arts

2006

Advisory Committee:

Professor Benjamin Kedem

Professor Paul Smith

Professor Lawrence Washington

Contents

List of Figures

1 Bayesian Binomial Regression

1.1 Assumptions

The proceeding approach for binomial data can be found in Bedrick, Christensen, and Johnson (BCJ 1997). A similar approach is illustrated by Leonard. Suppose that we have data (y_i, x_i') where each y_i is a success proportion from independent binomial (N_i, p_i) random variables and each of the n vectors x_i' is a known vector of covariates. We may assume that the probability, p_i, of success for a given vector of covariates, x_i', is $p_i = r(x_i', \beta)$, where β is an unknown vector of regression coefficients. Here the transformation r may be any cdf, but we will use the logistic transformation exclusively. Namely,

$$p_i = r(x_i', \beta) = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}} \tag{1}$$

Under the above assumptions we may write the Likelihood as a function of β given the vector of success proportions Y and the corresponding vector of known covariates X as follows:

$$L(\beta|Y) = \prod_{i=1}^{n} \frac{N_i!}{(N_i - N_i y_i)! N_i y_i!} [r(x_i'\beta)]^{N_i y_i} [1 - r(x_i'\beta)]^{N_i - N_i y_i} \tag{2}$$

For a given prior, $\pi(\beta)$, the posterior of β is:

$$\pi(\beta|Y) = \frac{L(\beta|Y)\pi(\beta)}{\int L(\beta|Y)\pi(\beta)d\beta} \tag{3}$$

Inference and prediction can be based on a discrete approximation of the distribution of the posterior. An interesting feature of the procedure is that the prior for beta is derived from prior assumptions on the probability, p, of success rather than the obscure regression coefficients. A Beta prior for p is assigned and then change of variables gives the "proper prior (BCJ 1997)" for beta. The reason for the use of the Beta prior for the p_i will be clear shortly as the use of such a prior combined with the above transformation function $r(x'\beta)$ provides a convenient form for the posterior of β. Specifically, one may assign an empirical prior distribution

π_0 (we will use $Beta(a_i, b_i)$ priors for each $\tilde{p_i}$) of P for a given set of covariates $\tilde{X}$ and derive the induced prior $\pi(\beta)$ for β as follows. Let $\pi_0(P) = \pi_0(R(\tilde{X}\beta))$. Then

$$\pi(\beta) = \pi_0(R(\tilde{X}\beta))|\frac{\partial R(\tilde{X}\beta)}{\partial \beta}| \tag{4}$$

where R is the vector transformation that applies r to each $\tilde{x_i}$ of the design matrix $\tilde{X}$, which must be a nonsingular square matrix of the same dimension, say k, as β, the unknown vector of regression coefficients.

It turns out that under the logistic transformation and the assumed $Beta(a_i, b_i)$ prior for the $\tilde{p}_i = r(\tilde{x_i}'\beta)$, where $a_i = \tilde{N}_i\tilde{y}_i, b_i = \tilde{N}_i - \tilde{N}_i\tilde{y}_i$ and $i = 1, 2, ...k$, the induced prior for β has the same form as the likelihood. We may see this fact from the following calculation using change of variables and assuming independence of the $\tilde{p}_i$.

$$\pi(\beta) = \pi_0(R(\tilde{X}\beta))|\tfrac{\partial R(\tilde{X}\beta)}{\partial \beta}|$$

$$= \textstyle\prod_{i=1}^{k} \frac{\Gamma(a_i+b_i)}{\Gamma(a_i)\Gamma(b_i)} r(\tilde{x_i}'\beta)^{a_i-1}[1 - r(\tilde{x_i}'\beta)]^{b_i-1}|\tfrac{\partial R(\tilde{X}\beta)}{\partial \beta}|$$

Note that

$$|\frac{\partial R(\tilde{X}\beta)}{\partial \beta}| = \begin{bmatrix} \frac{\partial r(\tilde{x_1}'\beta)}{\partial \beta_1} & \cdots & \frac{\partial r(\tilde{x_1}'\beta)}{\partial \beta_k} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \frac{\partial r(\tilde{x_k}'\beta)}{\partial \beta_1} & \cdots & \frac{\partial r(\tilde{x_k}'\beta)}{\partial \beta_k} \end{bmatrix}$$

where

$$\frac{\partial r(\tilde{x_i}'\beta)}{\partial \beta_j} = \frac{(1+e^{x_{i1}\beta_1+\ldots+x_{ik}\beta_k})(e^{x_{i1}\beta_1+\ldots+x_{ik}\beta_k})x_{ij} - (e^{x_{i1}\beta_1+\ldots+x_{ik}\beta_k})(e^{x_{i1}\beta_1+\ldots+x_{ik}\beta_k})x_{ij}}{(1+e^{x_{i1}\beta_1+\ldots+x_{ik}\beta_k})^2}$$

$$= \frac{x_{ij}(e^{x_{i1}\beta_1+\ldots+x_{ik}\beta_k})[(1+e^{x_{i1}\beta_1+\ldots+x_{ik}\beta_k}) - (e^{x_{i1}\beta_1+\ldots+x_{ik}\beta_k})]}{(1+e^{x_{i1}\beta_1+\ldots+x_{ik}\beta_k})(1+e^{x_{i1}\beta_1+\ldots+x_{ik}\beta_k})}$$

$$= x_{ij}\frac{(e^{x_{i1}\beta_1+\ldots+x_{ik}\beta_k})}{(1+e^{x_{i1}\beta_1+\ldots+x_{ik}\beta_k})}[1 - \frac{(e^{x_{i1}\beta_1+\ldots+x_{ik}\beta_k})}{(1+e^{x_{i1}\beta_1+\ldots+x_{ik}\beta_k})}]$$

$$= x_{ij} r(\tilde{x_i}\beta)[1 - r(\tilde{x_i}\beta)]$$

Therefore we may write $|\frac{\partial R(\tilde{X}\beta)}{\partial \beta}| =$

$$|det(\begin{bmatrix} r(\tilde{x_1}'\beta)[1 - r(\tilde{x_1}'\beta)] & 0 & ... & 0 \\ . & . & ... & . \\ . & . & r(\tilde{x_i}'\beta)[1 - r(\tilde{x_i}'\beta)] & . \\ . & . & ... & 0 \\ 0 & 0 & ... & r(\tilde{x_k}'\beta)[1 - r(\tilde{x_k}'\beta)] \end{bmatrix} \tilde{X})|$$

$$= |det(\tilde{X})| \prod_{i=1}^{k} r(\tilde{x_i}'\beta)[1 - r(\tilde{x_i}'\beta)]$$

Thus given a particular response vector $\tilde{y}$ of success proportions and corresponding design matrix of covariates $\tilde{X}$, under the logistic transformation, the induced prior for β is:

$$\pi(\beta) = |det(\tilde{X})| \prod_{i=1}^{k} \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} [r(\tilde{x_i'}\beta)]^{a_i} [1 - r(\tilde{x_i'}\beta)]^{b_i} \tag{5}$$

The design matrix $\tilde{X}$ must be nonsingular and have dimension $k \times k$, where k is the dimension of the unknown regression coefficient vector β. Here one may include an additional component to allow for the incorporation of an intercept in the analysis. Typically the design matrix is chosen to represent covariate locations that are rather extreme within the data. These covariates and their corresponding prior distributions can be thought of as some "prior knowledge". Choosing a design matrix will be discussed more in the context of our application to the NFL data set.

1.2 Inference

The Bayesian approach to estimating regression parameters is quite different from that of standard GLM. The philosophy of the Bayesian of course is to estimate (the mean of) a posterior distribution given the data for some prior, while the assumptions for GLM assign the data to some family of distributions in order to estimate such parameters as the mean and variance. Both schools of thought are

based on estimating the mean and variance of some assumed type of distribution. As opposed to attempting some maximization problem (the GLM approach), we will proceed with Importance Sampling to estimate the mean and covariance matrices of β.

It is important to select an importance density function, $g(\beta)$ that is "similar in shape to that of the known kernel of the posterior $L(\beta|Y)\pi(\beta)$ (BCJ 1997)". The tails of $g(\beta)$ should be heavy enough to include all extreme values in the data, so that these locations are not over-weighted in the estimation process. To get an estimator for β consider the following estimate of the conditional mean.

From the density function $g(\beta)$ sample:

$$\beta^1, \beta^2, ..., \beta^t$$

Then we can get, by the Law of Large Numbers, a discrete approximation to the conditional mean as follows:

$$E(\beta|Y) = \int \beta\pi(\beta|Y)d\beta =$$

$$\int \beta \frac{L(\beta|Y)\pi(\beta)/g(\beta)}{\int L(\beta|Y)\pi(\beta)d\beta/g(\beta)}d\beta \approx \sum_{i=1}^{t} \beta^i \frac{L(\beta^i|Y)\pi(\beta^i)/g(\beta^i)}{\sum_{j=1}^{t} L(\beta^j|Y)\pi(\beta^j)/g(\beta^j)} \tag{6}$$

Now let

$$q_i = q(\beta^i) = L(\beta^i|Y)\pi(\beta^i)/g(\beta^i) \tag{7}$$

and

$$\tilde{q_i} = \frac{q_i}{\sum_{i=1}^{t} q_i} \tag{8}$$

Thus we get the estimator for β as

$$\hat{\beta} = \sum_{i=1}^{t} \beta^i \tilde{q_i} \tag{9}$$

and the estimator for the covariance matrix

$$\hat{cov}(\beta|Y) = \sum_{i=1}^{t} \beta^i \beta^{i'} \tilde{q_i} - \hat{\beta}\hat{\beta}' \tag{10}$$

1.3 Prediction

For a new trial y with covariate x the predictive probability of success is

$$p(y = 1|Y, x) = E[r(x'\beta)|Y, x] = \int r(x'\beta)\pi(\beta|Y)d\beta$$

We may use a discrete approximation to the posterior to get

$$\hat{p}(y = 1|Y, x) = \sum_{j=1}^{t} r(x'\beta^j)\tilde{q}_j \tag{11}$$

Here Y represents all previous observations, while y is the unknown outcome, which has the above probability of success.

2 GLM

2.1 Assumptions

The versatility of the Generalized Linear Model allows for application to many assumed distributions of data. Instead of imposing a specific distribution, GLM assumes only that the distribution of the data belongs to the exponential family. The data may even be correlated as described by Kedem and Fokianos (2002). Suppose that we have response data y_i with expected value $E(y_i) = \mu_i$ and density function, $f(y_i, \theta_i)$, from the exponential family. Therefore $f(y_i, \theta_i)$ may be written as:

$$f(y_i, \theta_i) = e^{y_i\theta_i + b(\theta_i) + c(y_i)}$$

The expected value of the y_i is related to the covariate vector x_i and an unobservable regression coefficient vector β by way of a link function g. Particulary, $g(\mu_i) = x_i'\beta$. In this sense a function of the expected value is linear.

2.2 The Binomial Case

The binomial distribution of y_i with parameters n_i and p_i is of the exponential family as it can be written as (Rencher 2000):

$$f(y_i; p_i) = \frac{n_i!}{(n_i - y_i)!y_i!} p_i^{y_i}(1-p_i)^{n_i - y_i} \tag{12}$$

$$= e^{y_i \ln p_i - y_i \ln(1-p_i) + n_i \ln(1-p_i) + \ln(\frac{n_i!}{(n_i-y_i)!y_i!})}$$

$$= e^{y_i \ln(\frac{p_i}{1-p_i}) + n_i \ln(1-p_i) + \ln(\frac{n_i!}{(n_i-y_i)!y_i!})}$$

$$= e^{y_i\theta_i + b(\theta_i) + c(y_i)}$$

where $\theta_i = \ln(\frac{p_i}{1-p_i})$, $b(\theta_i) = n_i \ln(1-p_i) = -n_i \ln(1+e^{\theta_i})$, and $c(y_i) = \ln(\frac{n_i!}{(n_i-y_i)!y_i!})$
The likelihood function is given by:

$$L(\beta) = \prod_{i=1}^{n} e^{y_i\theta_i + b(\theta_i) + c(y_i)} \tag{13}$$

The estimator of β is obtained from maximizing the log-likelihood function:

$$\ln L(\beta) = \sum_{i=1}^{n} y_i\theta_i + \sum_{i=1}^{n} b(\theta_i) + \sum_{i=1}^{n} c(y_i) \tag{14}$$

For the exponential family $E(y_i) = \mu_i = -b'(\theta_i)$ (Venables and Ripley 2002), which gives the relation between θ_i and the link function g.
$g(\mu_i) = x_i'\beta$
The solutions to the corresponding maximization equations (obtained by differentiating the log-likelihood with respect to each β_i) are found iteratively. This highlights a significant difference to the Bayesian procedure, where the estimators are found using a discrete approximation to the posterior of β. Also notice that for the binomial distribution
$-b'(\theta_i) = n\frac{e^{\theta}}{1+e^{\theta}} = np$ since $\theta = \ln(\frac{p}{1-p}) \Rightarrow p = \frac{e^{\theta}}{e^{\theta}+1} = r(\theta)$
therefore,
$g(x) = r^{-1}(x) = \ln(\frac{x}{1-x})$ is the natural link function for binomial data.

3 Application to NFL Data

The preceding outlined Bayesian procedure provides an alternative approach to logistic regression. This method may provide more accuracy than standard GLM, as it allows for a prior manipulation of the data based on an "expert opinion" of the user. As a lifelong football fan, I will put my opinion to work with the Bayesian method and compare the results to GLM with the previously mentioned link function. For simplicity the data are assumed to be independent.

The data set consists of outcomes from the 2005 NFL regular season. For each game i where $i = 1, 2, ..., 256$ one team was randomly selected and the following were recorded; $y_i = 1$ if the selected team won and $y_i = 0$ if not, and the corresponding vector of covariates, x, consists of the selected team's net rushing yards, "NR", net passing yards, "NP", turnover differential, "TO", and a 1 or 0 indicating whether the selected team was the home or away team respectively. The goal was to determine which of these factors have the most effect on the outcome of a pro football game. Of course only one of the above listed covariates (home or away) would actually be known before an unplayed game, so this approach is best suited to find the impact on the outcome of a game of the previously listed measurable performance indicators.

The nature of the data set and the Bayesian method provide plenty of room for manipulation on the part of the user. One important fact to note is that since each covariate vector x_i consists of 5 components, there were no repeated identical observations. Particularly the data are Bernouli, and thus each success proportion y_i is either 0 or 1. This fact does cause some technical difficulties for assigning $Beta$ priors for the $\tilde{p}_i$. Specifically, if the observed covariate $\tilde{x}_i$ occurs only once, then the corresponding prior distribution will be of the form $Beta(0, 1)$ or $Beta(1, 0)$ which are both undefined. Under either circumstance we may still symbolically derive a well defined induced prior for β for inference, but this inhibits our ability for comparing the prior and posterior distributions of the $\tilde{p}_i$. To avoid this problem we employ our knowledge of the data set to find sets of covariate locations for each $\tilde{x}_i$ that have empirically similar odds of success. We can therefore regard each observation of these sets as a repeated observation of the selected design covariate $\tilde{x}_i$. In that sense we will specify values $\tilde{N}_i > 0$ and $0 < \tilde{y}_i < 1$ in order to get well

defined priors for the $\tilde{p}_i$.

3.1 Experimental Design

Before we can derive a prior for β, we must specify a design matrix $\tilde{X}$. The authors BCJ (1997) suggest selecting "values of the variables that were relatively extreme within the data, but still had substantial probabilities (of success)." To do this we employ our "expertise" of the subject at hand and consider different choices for the design. As mentioned above, for each of the selected design covariate locations $\tilde{x_i}$ we selected from the data other covariate vectors x_j which satisfied similar conditions as the design location $\tilde{x_i}$. This process of course requires intimate knowledge of the data in order to determine what specific conditions must be met to achieve such probabilistic similarity.

There is a widespread belief throughout all levels of organized football that three of the biggest factors in determining the outcome of a game are turnovers, success rushing the ball, and home field advantage (drbobsports.com). With this in mind we specify our design:

$$\tilde{X} = \begin{bmatrix} 1 & -123 & -34 & 1 & 0 \\ 1 & -100 & 112 & 1 & 1 \\ 1 & -12 & 18 & -1 & 1 \\ 1 & 100 & -12 & 1 & 1 \\ 1 & 139 & -44 & 0 & 0 \end{bmatrix} \tag{15}$$

$\tilde{x_i}'$=(intercept,NR,NP,TO,H/A)

A 1 in the turnover column indicates that that team won the turnover battle by 1, i.e. that particular team had one less turnover than their counterpart that day. Similarly a -123 in the rushing column indicates that that team was out rushed by 123 yards in that game. The philosophy of this design was to select observations that had a low turnover differential and then contrast high, medium, and low net rushing totals with home and away. The belief here is that while keeping the turnovers relatively constant, we may be able to determine the relative importance of rushing versus being the home or away team. This design could be interpreted as being relatively centered in the turnover dimension and extreme within the rushing and

home or away dimension.

As previously mentioned the data are Bernouli, and some analysis must be done to determine an appropriate prior for each $\tilde{p}_i$. For each design location $\tilde{x}_i$ the empirical odds of success were determined by selecting from the data other covariate vectors that had similar components. For example, consider $\tilde{x_2}' = (1, -100, 112, 1, 1)$. This team's performance can be characterized as "getting out rushed by a lot", "gaining a lot more passing", and having a low turnover differential while playing at home. We then made suitable numeric conditions that reflect this characterization and found 10 games that satisfied the criterion. Of these 10 games only 2 teams won with this type of performance, and the associated prior for $\tilde{p_2} = r(\tilde{x_2}'\beta)$ is a $Beta(2, 8)$. This is an admittedly crude process, but it reflects our belief that such a performance results in a low probability of success, while maintaining much uncertainty about the actual distribution of the probability of success. For each of the covariate locations $\tilde{x_i}$ listed above, the following priors were assigned.

$$\pi_0(\tilde{p_1}) = Beta(1, 5)$$

$$\pi_0(\tilde{p_2}) = Beta(2, 8)$$

$$\pi_0(\tilde{p_3}) = Beta(2, 2)$$

$$\pi_0(\tilde{p_4}) = Beta(19, 1)$$

$$\pi_0(\tilde{p_5}) = Beta(7, 3)$$

Choosing an appropriate importance density function is of great importance for the accuracy of prediction. The authors BCJ (1997) suggest a multivariate t distribution centered at the MLE for β with dispersion "proportional to the asymptotic covariance matrix evaluated at the mode." The goal is to get a density that is "similar in shape to the known kernel $L(\beta|Y)\pi(\beta)$" with "tails that decay less rapidly than the likelihood (BCJ 1997)." If the importance density function is of a poor match, then some of the resulting probabilities $\tilde{q_i}$ can be too heavily weighted causing the posterior distributions to be inaccurate. After several trial runs and using a numerical maximization method, we found that a multivariate t density with 6 degrees of freedom and

$$mean = (-0.29713023, 0.01005308, 0.00337546, 0.52726051, -0.02104360) \quad (16)$$

$$\Sigma = \begin{bmatrix} .263^2 & 0 & 0 & 0 & 0 \\ 0 & .003^2 & 0 & 0 & 0 \\ 0 & 0 & .002^2 & 0 & 0 \\ 0 & 0 & 0 & .131^2 & 0 \\ 0 & 0 & 0 & 0 & .366^2 \end{bmatrix} \tag{17}$$

works well for the data.

If a poorly fitting importance density function is chosen, it will be clear from the resulting discrete approximation of the posterior as some of the sampled β^t will have an extremely high relative weight $\tilde{q}_t$. In trial runs with t densities of different means we sampled 8000 β^t and found that in each of these trials there were sampled β^t with associated probabilities, $\tilde{q}_t$, of nearly .98. This situation is very problematic for acquiring an appropriate posterior for the $\tilde{p}_i$. While these particular β^t with a high associated sampled probability were pretty good estimates to the posterior mean, this was not evident in the resulting posterior for $\tilde{p}_i$ since all of the other sampled β^t had an associated probability of less than 10^{-3}. The result is a posterior density that is virtually indistinguishable among all of the prior covariate locations $\tilde{x}_i$.

For example, in one trial with a poor fitting t density we found that the largest associated probability for any of the sampled β^t was approximately .97 and the corresponding $\tilde{p_1} = r(\tilde{x_1}'\beta^t) \approx 0.15$, a seemingly reasonable probability of success for that particular covariate vector. However the mean of the entire augmented sample $r(\tilde{x_1}'\beta^j)$ for $j = 1, 2, ...8000$ was nearly .5, and the resulting discrete approximation to the posterior for $\tilde{p_1}$ had the form of a $Beta$ with a similar mean. In fact the mean of all of the $r(\tilde{x_i}'\beta^j)$ for $i = 1, 2, ...5$ and $j = 1, 2, ...8000$ were virtually the same. Clearly this is not a desirable situation as it dilutes any of the information regarding the accuracy of the posterior prediction. We can make the best of this situation by centering the importance t density at the vector β^t with the disproportionately high associated probability $\tilde{q}_t$. Choosing the mean vector for the importance t density in this way and adjusting the degrees of freedom to allow for "slower" decaying tails provides a good strategy for finding an appropriate importance density function.

3.2 Results

We sampled vectors β^t for $t = 1, 2, ...8000$ from a multivariate t density with the above named parameters for inference and prediction. The estimated vector of regression coefficients was:

$$\hat{\beta} = \begin{bmatrix} -0.47015 \\ 0.01490 \\ 0.00286 \\ 0.58997 \\ 0.33056 \end{bmatrix} \tag{18}$$

compared to the estimated coefficients using GLM:

$$\hat{\beta}_{glm} = \begin{bmatrix} -0.71893 \\ 0.01999 \\ 0.00607 \\ 0.79196 \\ 0.77375 \end{bmatrix} \tag{19}$$

and the covariance matrix with the Bayesian procedure:

$\hat{cov}(\beta|Y) =$

$$\begin{bmatrix} 2.389e-02 & 3.605e-05 & -4.037e-06 & 3.174e-04 & -1.276e-02 \\ 3.605e-05 & 2.698e-06 & -2.800e-08 & 7.388e-06 & 2.267e-05 \\ -4.037e-06 & -2.800e-08 & 1.885e-06 & 4.120e-06 & -1.133e-05 \\ 3.174e-04 & 7.388e-06 & 4.120e-06 & 7.896e-03 & -1.511e-03 \\ -1.276e-02 & 2.267e-05 & -1.133e-05 & -1.511e-03 & 4.741e-02 \end{bmatrix} \tag{20}$$

Therefore the variance of the estimator $\hat{\beta}$ is:

$$\hat{var}(\beta|Y) = \begin{bmatrix} 2.3897e-02 \\ 2.6982e-06 \\ 1.8858e-06 \\ 7.8967e-03 \\ 4.7410e-02 \end{bmatrix} \tag{21}$$

compared to the variance of the estimator found using GLM:

$$var(\hat{\beta}_{glm}) = \begin{bmatrix} 5.1166e-02 \\ 9.7969e-06 \\ 4.0763e-06 \\ 1.7192e-02 \\ 1.3417e-02 \end{bmatrix} \tag{22}$$

It is encouraging that the estimated regression coefficients and the estimated variances of the two respective methods are similar. The fact that the variance of the estimators under the Bayesian method are slightly smaller than those found using GLM suggests that we may have more accurate estimators with the Bayesian method, which could translate to better prediction.

To compare the prediction accuracy of the two respective methods we find the predicted probability of success with the estimated parameters for each one of the 256 observations (games). For a given game with observed covariate vector x_i the predicted probability of success under the GLM is:

$$p_i = \frac{e^{x_i'\hat{\beta}_{glm}}}{1+e^{x_i'\hat{\beta}_{glm}}}$$

If $p_i > .5$ then we predict the outcome to be a win, and if not then a loss. The predicted outcomes were then compared to the actual outcome with the following results. The Bayesian procedure with the above given importance t density and design matrix $\tilde{X}$ correctly predicted 84.675 (217 out of 256) percent of the games, while GLM with the logit link function and the above estimated parameters correctly predicted 83.9 (215 out of 256) percent.

An approximation to the posterior density of each of the $\tilde{p}_i$ for $i = 1, 2, ..., 5$ was found by smoothing a random sample from the discrte approximation that takes the value $r(\tilde{x}_i\beta^j)$ with probability $\tilde{q}_j$. Histograms of the smoothed samples as well as the prior densities for each of the prior locations $\tilde{x}_i$ are given below(figures 1 and 2). It is apparent that the means of the posterior densities are close to what we anticipated as the probability of success for each of the design locations. For example, the prior mean of design location 1 was $1/6 \approx .17$, and from the histogram the posterior mean appears to be close to that. The actual mean of the smoothed sample for location 1 was .24. So in this sense our prior beliefs are enforced, but we

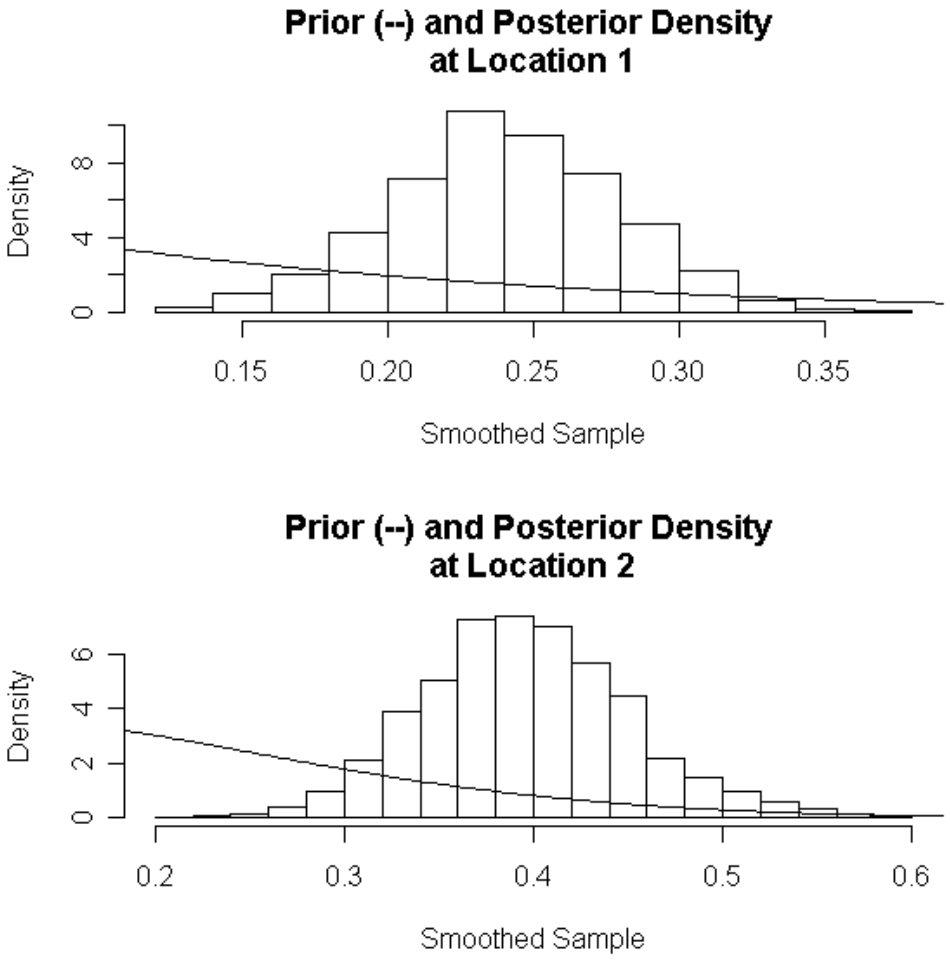

Figure 1: Prior and Posterior Densities at Locations 1 and 2

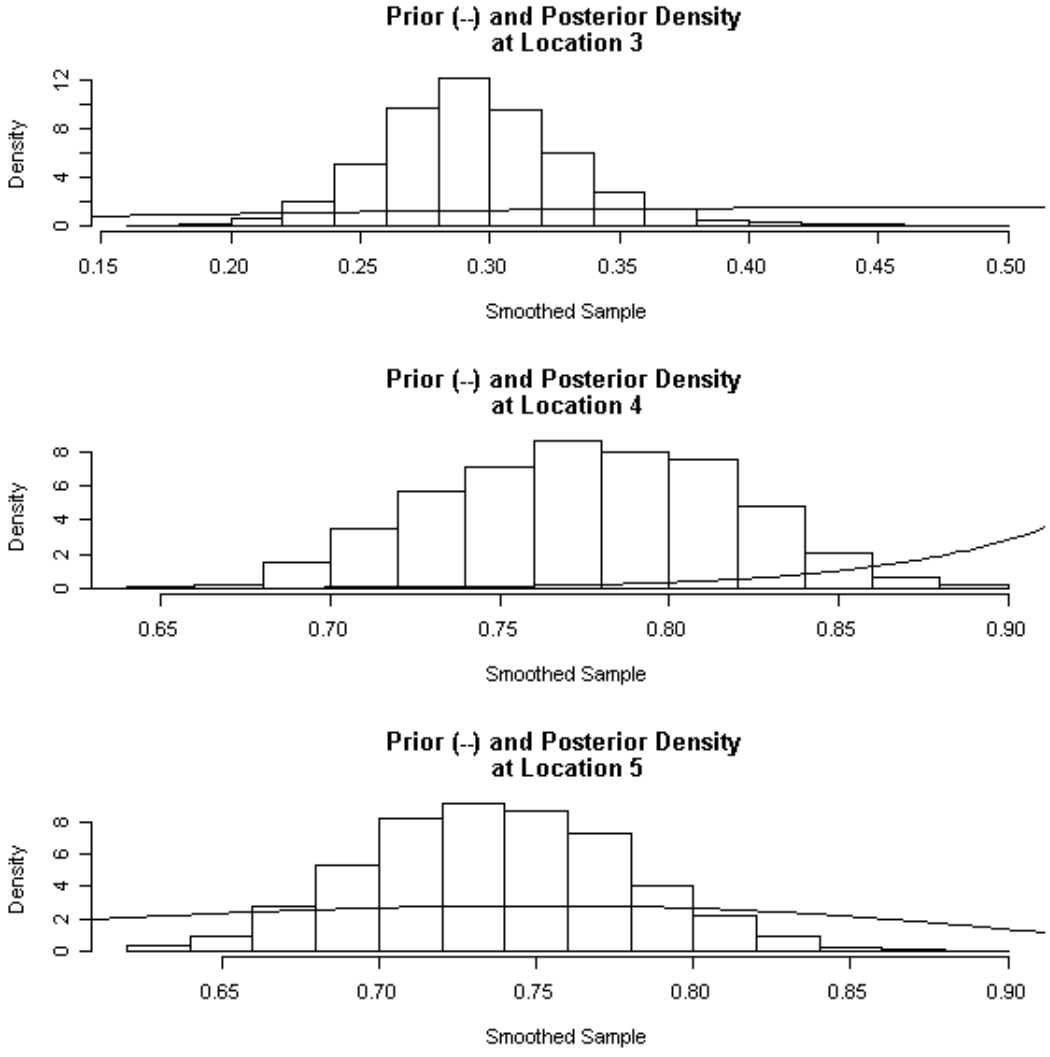

Figure 2: Prior and Posterior Densities at Locations 3, 4, and 5

get a much better understanding as to the actual distribution of the probability of success for such covariates. There is a significant difference in the prior and posterior densities as far as shape and allowance of extreme events.

It is exciting that the Bayesian procedure provides better prediction for the given data, but it may be more interesting to investigate the particular covariate locations at which the two models differ in predicted probability and consider altering the assigned priors.

To compare the significance of rushing yardage and turnover differential to the widely accepted home field advantage, we consider predictive probabilities of winning as a function of rushing yardage with different combinations of passing and turnover totals for home and away teams. The first diagram (figures 3 and 4) contrasts predictive probabilities as a function of rushing yardage for the fixed values for net passing yardage, $NP = -100$ and turnover differential, $TO = -2$ for home vs. away teams. The second is identical except that $TO = 2$. Then we consider the same values for TO and adjust the net passing to $NP = 100$. The respective plots give insight into the effect of turnover differential for fixed passing yardage as a function of rushing yardage for home and away teams. A comparison of the respective distributions shows the relative importance of passing yardage to rushing yardage for fixed turnover differentials, $TO = -2$ and $TO = 2$, that can be characterized as "relatively low."

We can draw some very interesting conclusions from the above distributions. First notice that the shape, particularly the rate of growth, is very similar among the 4 distributions with common turnover differentials (counting home vs. away). For those with $TO = -2$ the growth is relatively slow compared to those with $TO = 2$. This suggests that it requires a certain amount of success in the running game to overcome turnovers. The plots also strongly indicate that net passing yards is relatively insignificant regardless of turnovers. For example, both distributions with a fixed $TO = -2$ indicate an approximately equal need of between $75 - 85$ rushing yards by the home team to have $pr(win) > .5$ regardless of having a 100 yard deficit in the passing game. The plots with $TO = 2$ are similar with respect to the little effect of passing yardage on $pr(win)$, but obviously show a clear difference in that the necessary rushing yardage is significantly less to have $pr(win) > .5$.

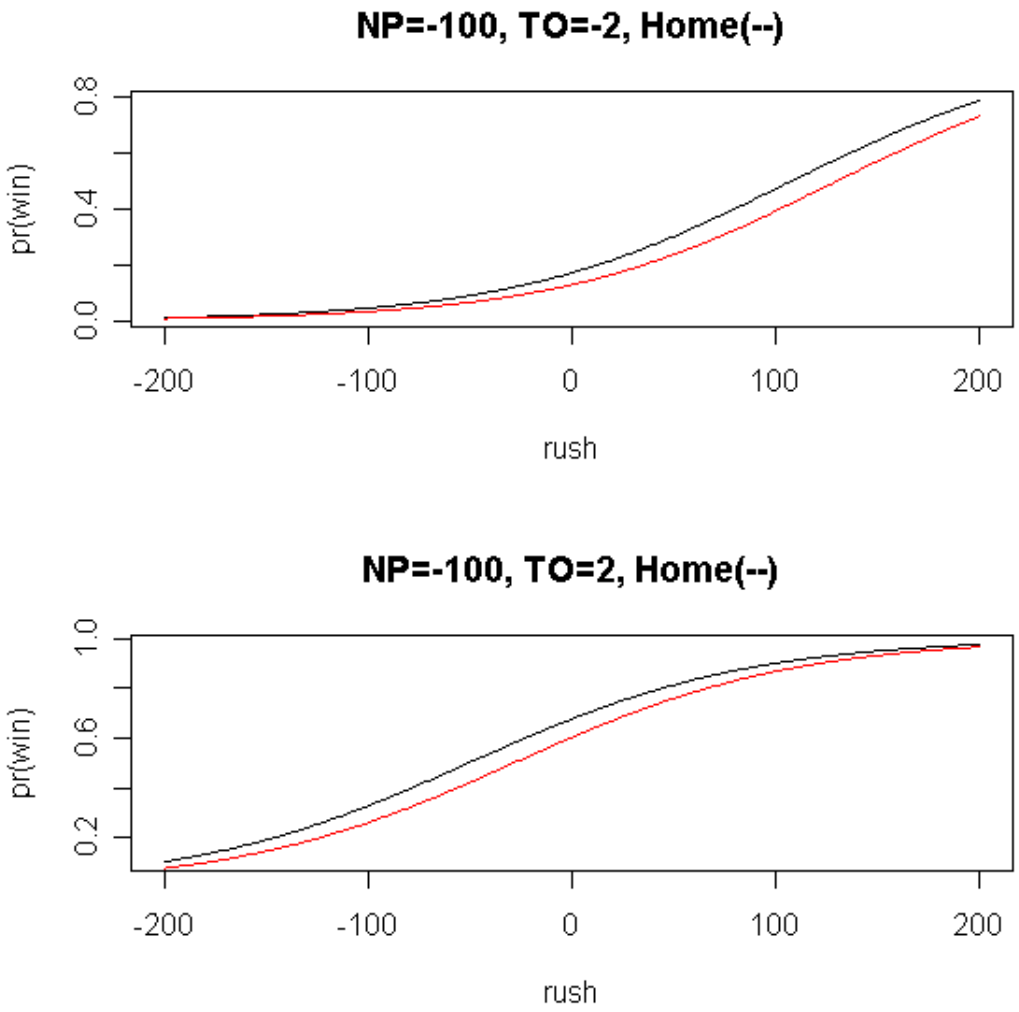

Figure 3: Predictive Probabilities of Winning as a Function of Rushing Yardage for Fixed NP and TO

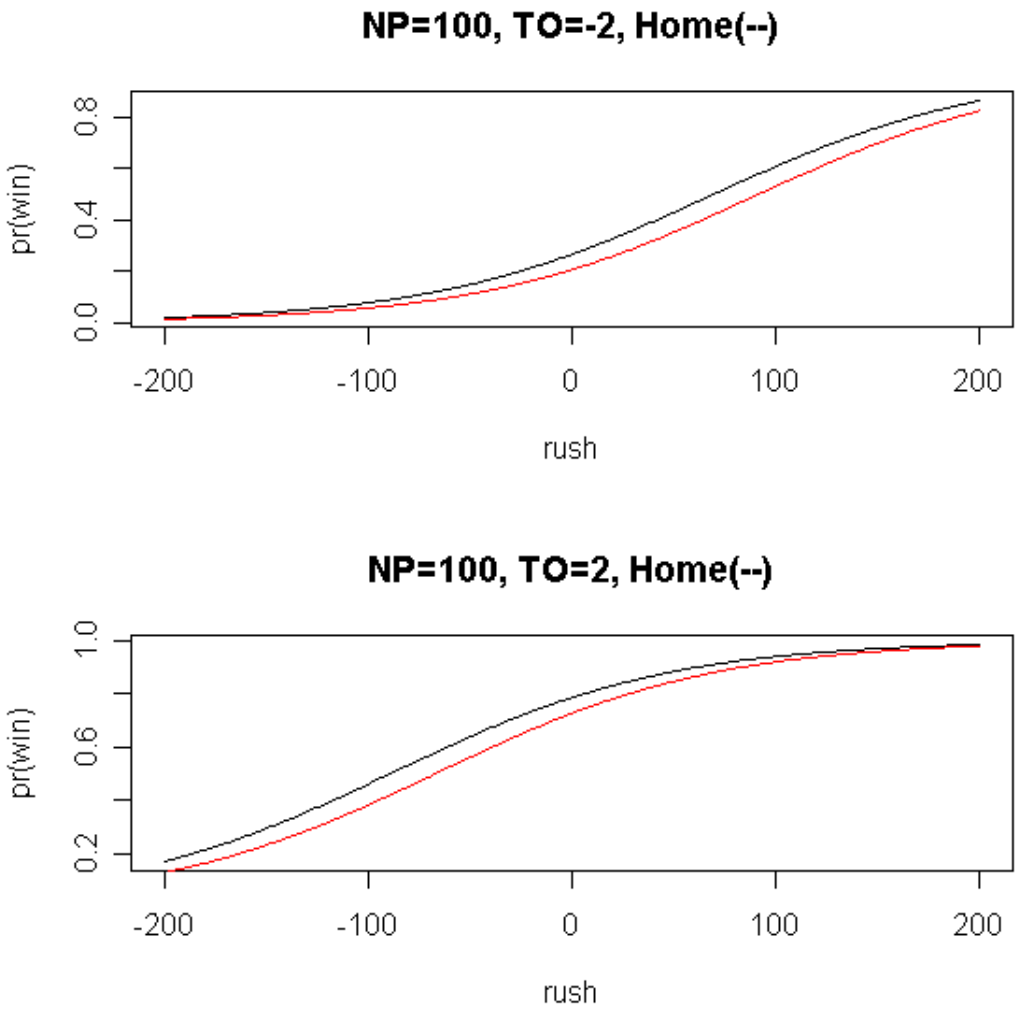

Figure 4: Predictive Probabilities of Winning as a Function of Rushing Yardage for Fixed NP and TO

As expected the $pr(win|home) >= pr(win|away)$ in all distributions, but is interesting to note where the significant separation in $pr(win)$ occurs between the home and away team. It appears that different NR and TO combinations directly affect home field advantage. Apparently home field advantage becomes negligible for extreme NR and TO totals as indicated in both figure 3 and figure 4. Particularly, there is little difference between $pr(win)$ for the home and away team for negative values of TO and large (in the negative sense) net rushing totals. Also there is little difference between $pr(win)$ for the home and away team for positive values of TO and large net rushing totals. This observation is not surprising as it simply means that teams that run the ball well and win the turnover battle, which usually means that the particular team plays good defense, effectively eliminate any disadvantage of playing on the road, but it also indicates that these teams would have an even greater advantage when playing at home.

In light of the above observations we will consider distributions for $pr(win|home)$ vs. $pr(win|away)$ as functions of net passing yardage (figures 5 and 6), NP, for fixed values of NR and TO. The values $TO = 2$, $TO = -2$, $NR = 50$, and $NR = -50$ are chosen to give a clear contrast while maintaining significant win and loss probabilities.

It is clear from figures 5 and 6 that NP has a relatively small effect on $pr(win)$ compared to NR and TO. For example, in figure 5 we have the fixed values $NR = 50$ and $TO = 2$ and the range of $pr(win|home)$ is approximately between $.75 - .9$ for domain values of NP between -200 to 200. This means that a 400 yard swing in the passing game has a minimal effect on $pr(win)$. Also in figure 6 for the fixed values $NR = -50$ and $TO = -2$ the $pr(win)$ changes very little and barely exceeds .2 for the home team and .15 for the away team. Again the indication here is that winning the turnover battle and out rushing the other team has the most effect on $pr(win)$. We also see that there is a significant "gap" between the home and away team in both figure 5 and 6. Since the rushing totals chosen here were not extremely large or small, we see the home field advantage at work.

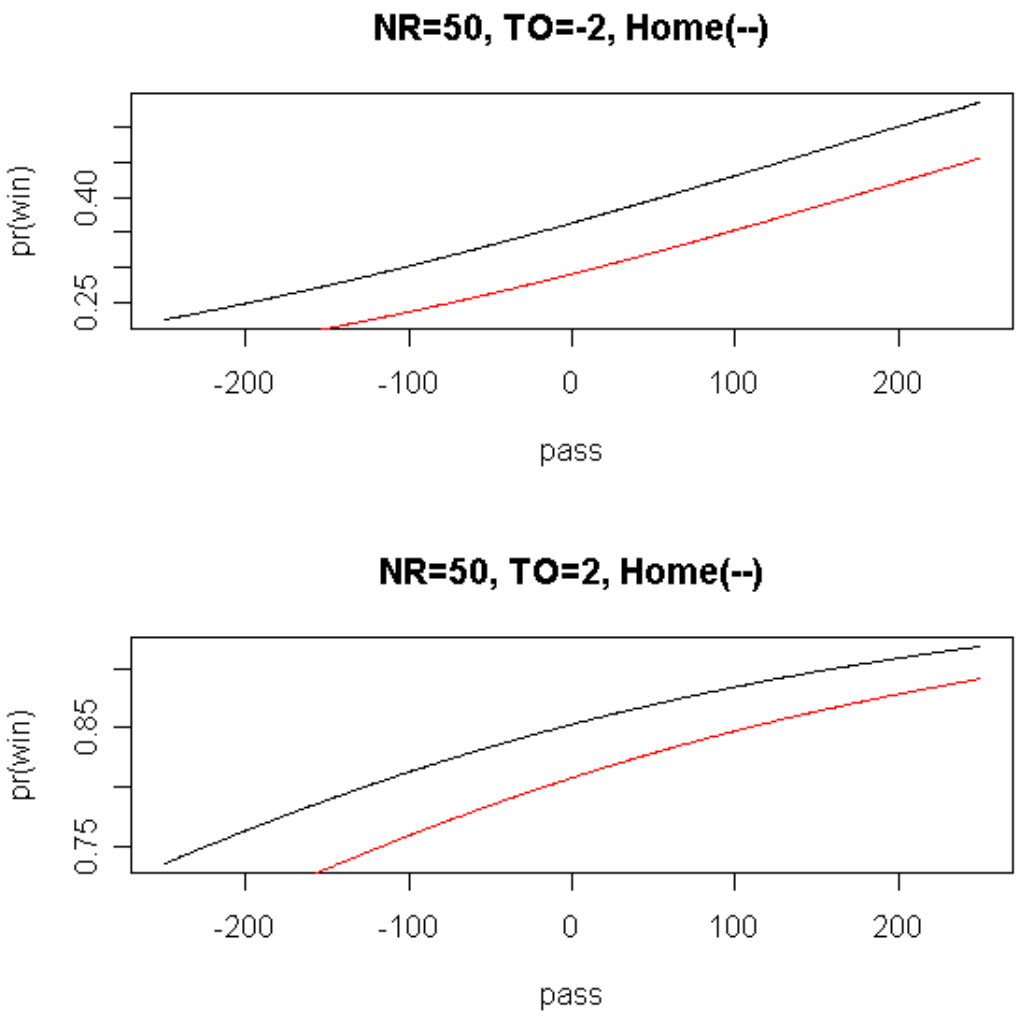

Figure 5: Predictive Probabilities of Winning as a Function of Passing Yardage for Fixed NR and TO

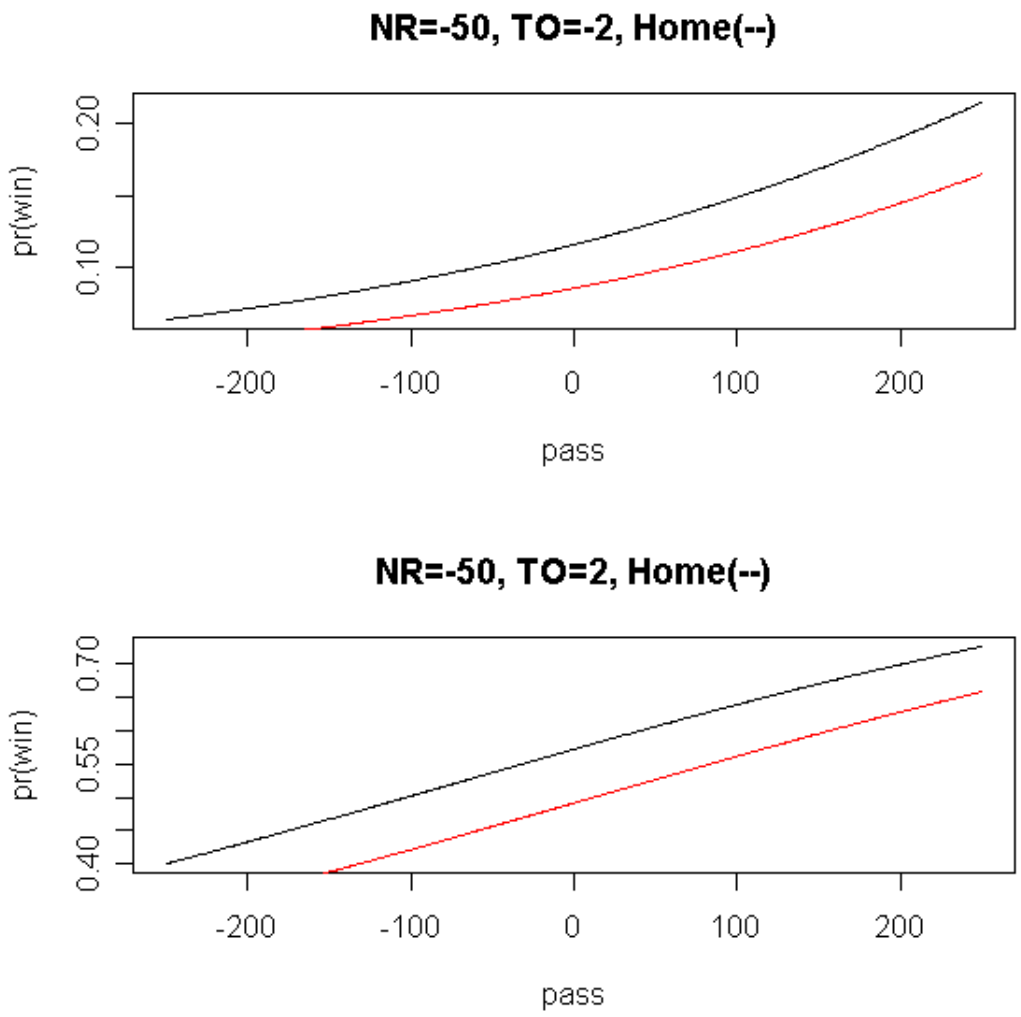

Figure 6: Predictive Probabilities of Winning as a Function of Passing Yardage for Fixed NR and TO

3.3 Test Data

For an additional measure of the relative accuracy of the two methods, all computations were repeated using only the first 226 observations. Then the same prediction decision as before was employed to predict the outcome of the final 30 observations. The results of the two methods were identical. Both predicted $26/30 \approx 0.8667$ of the remaining games, which is slightly better than the outcome using all 256 observations.

3.4 An Altered Prior

To consider the effect of the chosen prior the procedure was repeated with both an empirically poorly fitting prior and the diffuse prior, i.e. $\pi(\beta) = 1$. For a poorly fitting prior simply switch each a_i for b_i and b_i for a_i.

$$\pi_0(\tilde{p_1}) = Beta(5, 1)$$

$$\pi_0(\tilde{p_2}) = Beta(8, 2)$$

$$\pi_0(\tilde{p_3}) = Beta(2, 2)$$

$$\pi_0(\tilde{p_4}) = Beta(1, 19)$$

$$\pi_0(\tilde{p_5}) = Beta(3, 7)$$

Each π_0 is the opposite from the original prior, except for $\pi_0(\tilde{p_3})$ which is identical. Computation with the above prior yielded the following results:

$$\hat{\beta} = \begin{bmatrix} -0.42812 \\ 0.00658 \\ 0.00367 \\ 0.597518 \\ 0.086264 \end{bmatrix} \tag{23}$$

$\hat{cov}(\beta|Y) =$

$$\begin{bmatrix} 2.3399e-02 & 6.9333e-08 & 1.3754e-05 & -1.9367e-03 & -1.7714e-02 \\ 6.9333e-08 & 2.2376e-06 & 2.3135e-07 & -1.9549e-05 & -5.1948e-05 \\ 1.3753e-05 & 2.3134e-07 & 1.4561e-06 & 8.8129e-06 & -3.0065e-05 \\ -1.9367e-03 & -1.9549e-05 & 8.8129e-06 & 6.1055e-03 & 8.1005e-05 \\ -1.7714e-02 & -5.1948e-05 & -3.0065e-05 & 8.1005e-05 & 4.0605e-02 \end{bmatrix} \quad (24)$$

As before, each x_i was determined to be a success if $\hat{p}_i > .5$ and a failure otherwise. This time the results were not as favorable but still respectable. Prediction in this manner with the "poor" prior correctly predicted 0.8047 of the games, compared to 0.84765 before. Clearly the new prior has an undesirable effect on prediction. Considering the chosen priors, it seems that extreme valued $\hat{p}_i$ would be impacted the most since the design locations with an empirically low probability of success were assigned $Beta$ distributions with a large mean and vice versa. With this in mind consider the posterior densities for each $\tilde{p}_i$(figures 7 and 8).

There are some interesting, apparent differences among the respective posterior densities under the two priors. The posterior means of the first two locations are noticeably larger, while the third location is relatively unchanged. The densities of the last two locations, particularly location 5, are extremely skewed compared to the original results. The skewing of the densities is to be expected, as the new prior has provided empirically false information. For example in the case of location 5, the left hand tail (figure) decays slower than with the original prior allowing for a greater probability of failure.

We now perform all calculations using the diffuse prior, $\pi(\beta) = 1$. Results:

$$\hat{\beta} = \begin{bmatrix} -0.4324 \\ 0.0139 \\ 0.0031 \\ 0.6057 \\ 0.2856 \end{bmatrix} \quad (25)$$

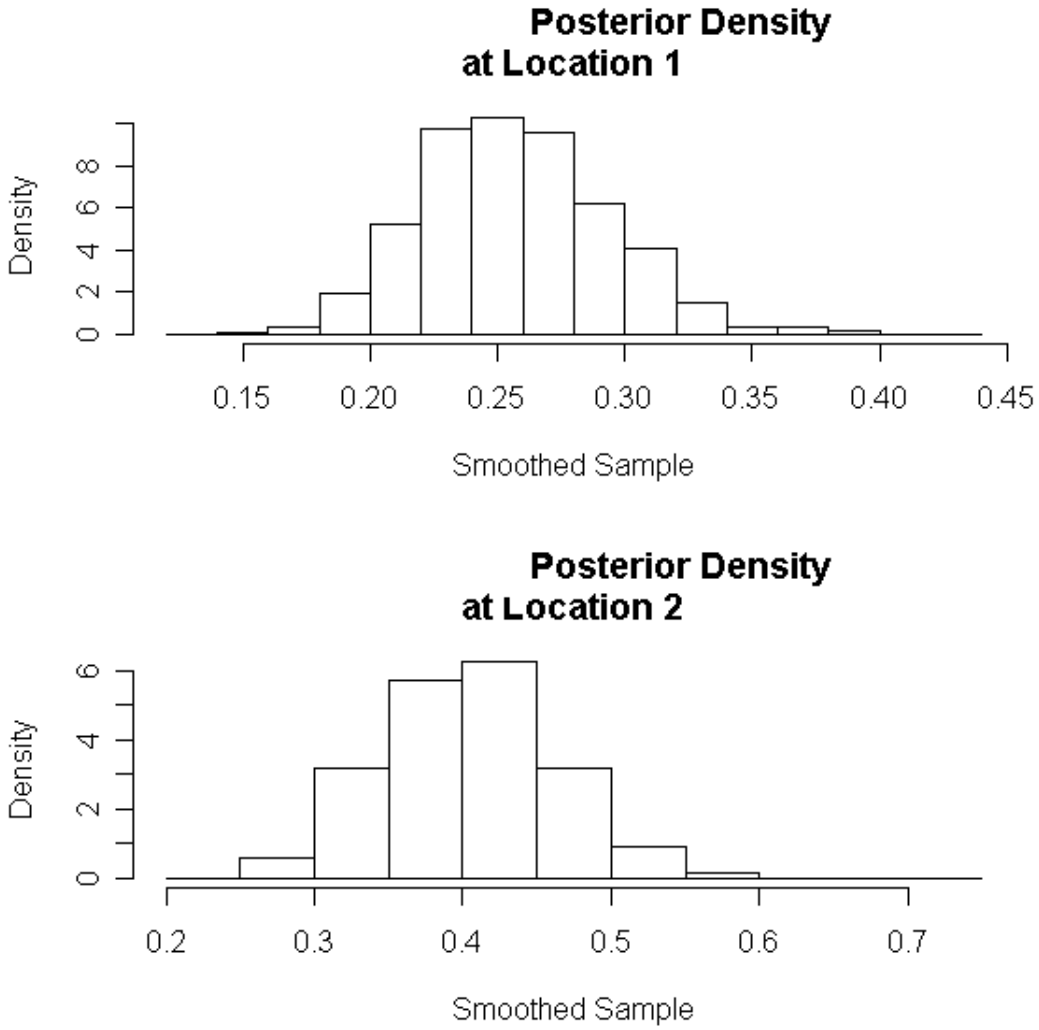

Figure 7: Posterior Densities at Locations 1 and 2 with Altered Prior

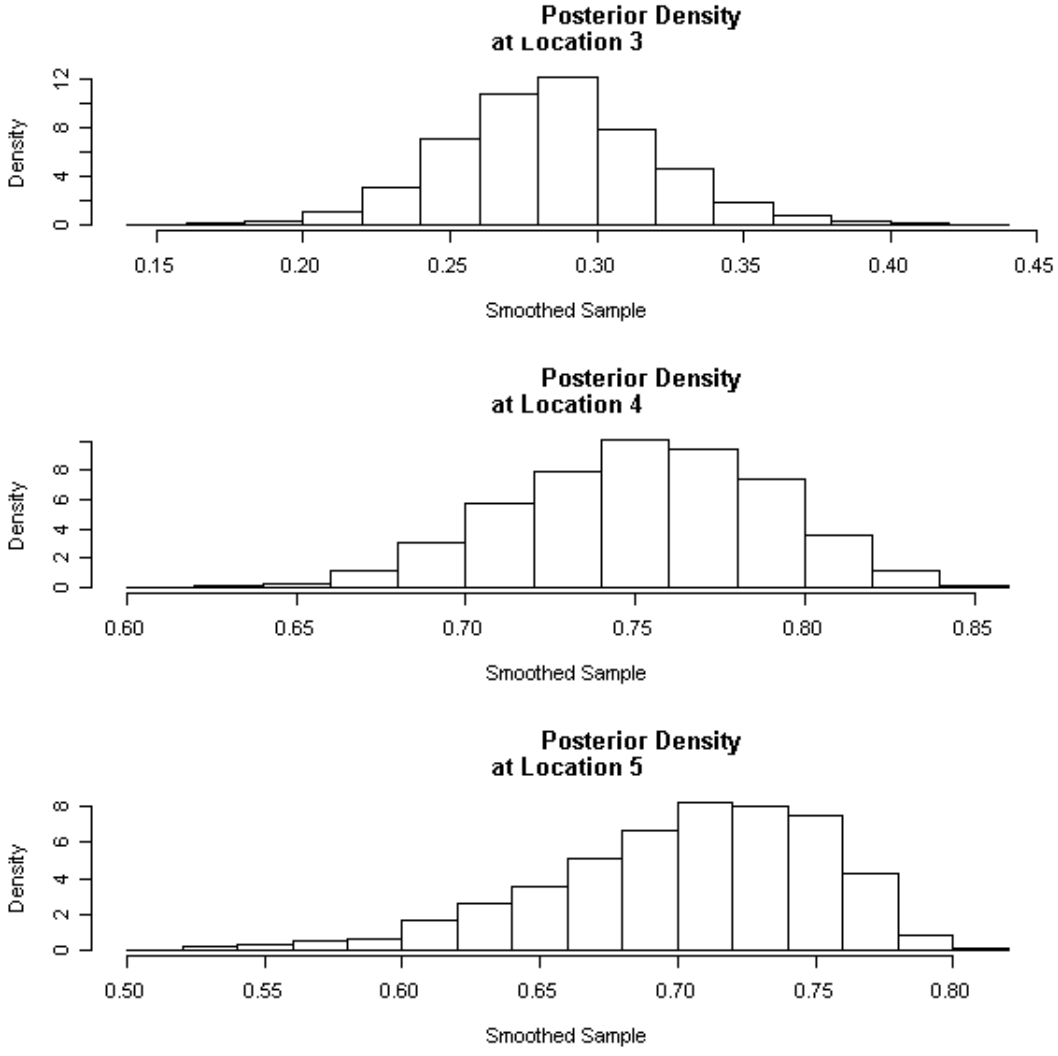

Figure 8: Posterior Densities at Locations 3, 4, and 5 with Altered Prior

$\hat{cov}(\beta|Y) =$

$$\begin{bmatrix} 2.7576e-02 & 1.4572e-05 & -3.7450e-06 & -4.8933e-04 & -1.6528e-02 \\ 1.4572e-05 & 3.8754e-06 & -3.7560e-08 & 1.9311e-05 & 9.5400e-06 \\ -3.7450e-06 & -3.7560e-08 & 2.1300e-06 & 1.1266e-05 & 1.0243e-05 \\ -4.8933e-04 & 1.9311e-05 & 1.1266e-05 & 7.5078e-03 & 4.2294e-04 \\ -1.6528e-02 & 9.5400e-06 & 1.0244e-05 & 4.2294e-04 & 5.5042e-02 \end{bmatrix} \quad (26)$$

The mean of β differs slightly from the estimates found using both the original prior and GLM. For example, the estimate found using GLM puts more emphasis on home field advantage, and it is interesting to note that the "poor" prior provides estimates for β_2 and β_5, the coefficients for NR and home field advantage, that differ significantly from the others. This is of particular interest since the design, $\tilde{X}$ was chosen to contrast these two parameters. Posterior densities for the $\tilde{p_i}$ (figures 9 and 10) are not noticeably different from those found with the original prior aside from a slight difference in the rate of decay of the tails at some locations.

3.5 Case Deletion Diagnostics

We examine the effect on the predictive probabilities of removing "prior" observations. Specifically we would like to know if there are any data that will significantly affect the predictive probabilities upon deletion. Following the notation of BCJ (1997), if $Y_{(i)}$ denotes the data without y_i, then the likelihood for β for all data except y_i is:

$$L(\beta|Y_{(i)}) = \frac{L(\beta|Y)}{L(\beta|y_i)} \quad (27)$$

Therefore the posterior of β from the reduced data is:

$$\pi(\beta|Y_{(i)}) = \frac{L(\beta|Y_{(i)})\pi(\beta)}{\int L(\beta|Y_{(i)})\pi(\beta)d\beta}$$

$$= \frac{\pi(\beta|Y)/L(\beta|y_i)}{\int \pi(\beta|Y)/L(\beta|y_i)d\beta} \quad (28)$$

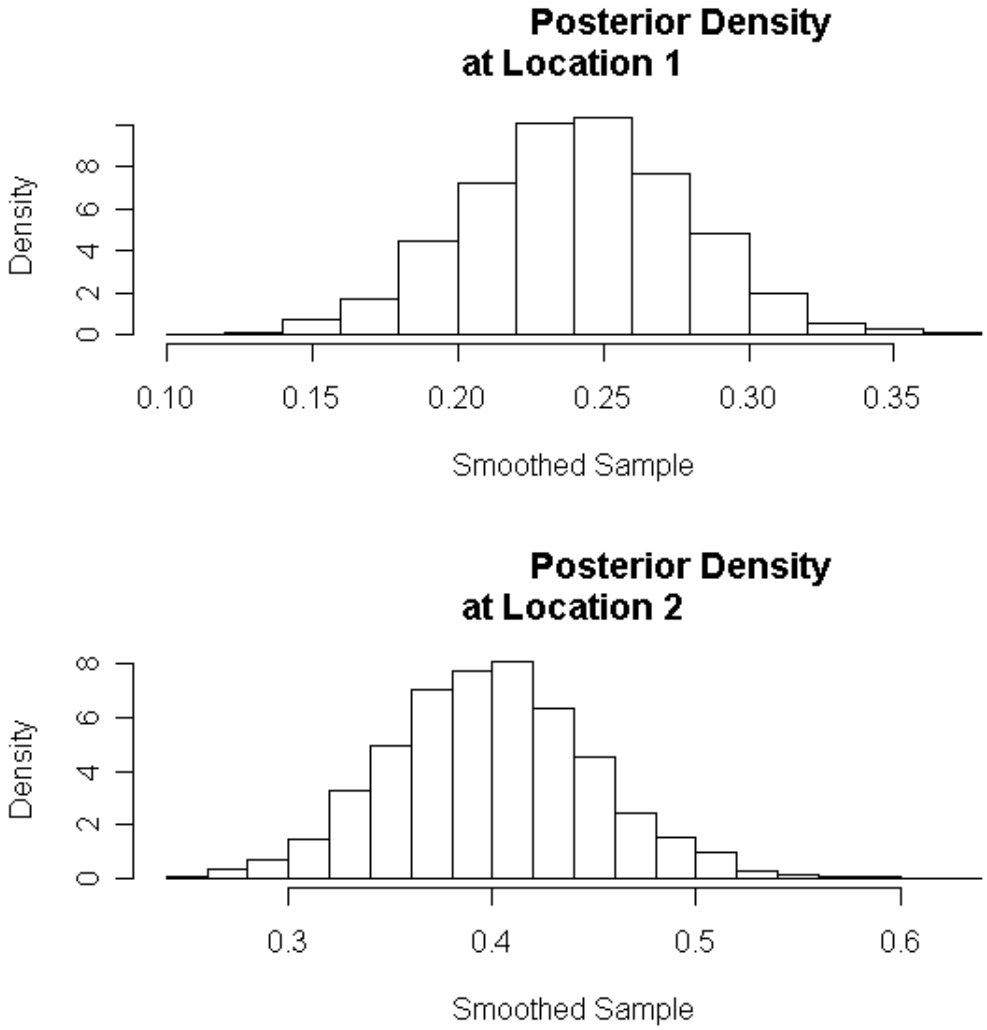

Figure 9: Posterior Densities at Locations 1 and 2 with Diffuse Prior

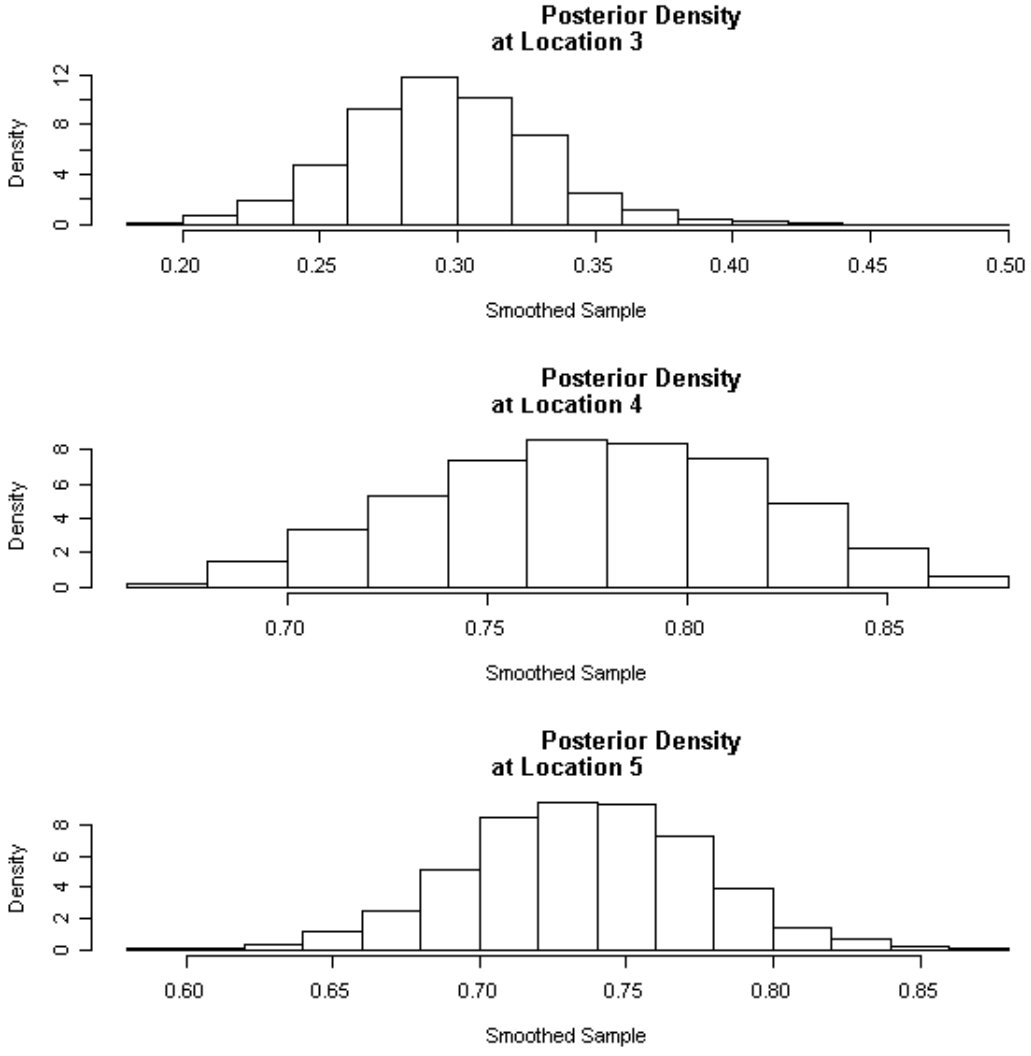

Figure 10: Posterior Densities at Locations 3, 4, and 5 with Diffuse Prior

As before we use the discrete approximation to the posterior, which takes values β_j with probability $\tilde{q}_{j(i)}$ where

$$\tilde{q}_{j(i)} = \frac{\tilde{q}_j / L(\beta^j | y_i)}{\sum_{k=1}^{t} \tilde{q}_k / L(\beta^k | y_i)} \tag{29}$$

Thus for each observation y_i we may compute the new predictive probability of success for any covariate vector x_j having removed the observation y_i from the data and compare the results to the full data set. "The symmetric Kullback-Leibler (KL) divergence is used to measure the discrepency between full and reduced data predictive distributions (BCJ 1997)." The predictive distribution of a single trial is Bernoulli, and the KL divergence between two Bernoulli distributions with probabilities p and q is

$$J(p, q) = (p - q) log[\frac{p(1-q)}{q(1-p)}] \tag{30}$$

We would like to have a measure of the effect on all predictive probabilities of removing say y_i from the data. The authors, BCJ (1997), define a "symmetric predictive divergence diagnostic" as

$$D_i^p = \sum_{j=1}^{n} J(p(y = 1 | Y, x_j), p(y = 1 | Y_{(i)}, x_j)) \tag{31}$$

Where $p(y = 1 | Y_{(i)}, x_j)$ is the probability of success having removed y_i from the data, and $p(y = 1 | Y, x_j)$is the predictive probability based on all the data. We may estimate the probability of success based on the reduced data having removed case i by using the weights $\tilde{q}_{j(i)}$, and then

$$p(y = 1 | Y_{(i)}, x) = \int r(x'\beta) \pi(\beta | Y_{(i)}) d\beta \approx \sum_{j=1}^{t} r(x'\beta^j) \tilde{q}_{j(i)} \tag{32}$$

Clearly a larger D_i^p value indicates that the observation y_i is more extreme within the data, and perhaps contributes more to the variability of the estimates, which is of course undesirable. In the application to the NFL data set the maximum D_i^p values were .034, .033, and .026 which occurred twice. For simplicity the D_i^p were computed in two groups; teams that won and teams that lost. It is clear from the plots of the respective D_i^p (figure 11) that the above values are outliers and have a relatively large effect on the predictive probabilities. The D_i^p values above correspond to the following observations:

$$.034 \leftrightarrow (49, -67, -4, 0)$$

$$.033 \leftrightarrow (-112, -156, 4, 0)$$

$$.026 \leftrightarrow (22, -149, 3, 0) and (53, 51, 3, 1)$$

Upon further review of the data it was discovered that the largest D_i^p value of .034 corresponded to a data vector that was incorrect. The data vector $x_{122} = (49, -67, -4, 0)$ corresponds to a team that won the game but has a very low predicted probability of winning $p = .0974$. The correct data vector should include a TO total of 4 instead if -4. This change of course drastically alters the predicted probability of winning for the corrected vector $x_{122} = (49, -67, 4, 0)$ to $p = .914$. We would therefore like to know the effect on all predicted probabilities of removing this observation from the data set (figure 12). Note that in the following calculations the data vector x_{122} is still in its original incorrect form, but we can consider the effect of removing it from the data set by looking at the difference $pr(win|alldata) - pr(win|alldataexceptcase122)$.

Most of the differences are close to zero, but there are some outliers. For example, it is interesting that of the winning teams the case with the largest decrease in predicted probability was case $x_{29} = (87, 20, -2, 1)$, which is similar to the removed x_{122}. It is reasonable that similar data vectors to x_{122} would have a decreased predicted probability of success after removing this case since x_{122} has a very low predicted probability of success but was characterized as a win. The vector $x_{65} = (-112, -156, 4, 0)$ experienced the largest increase in predicted probability of success among winning teams, and it is virtually the opposite of the original x_{122}. The most obvious conclusion that can be drawn from these two observations is that the incorrectly recorded x_{122} skewed the weighting of the effect of turnovers on predicted probabilities. The original data vector describes a team that still won the game having given up 4 more turnovers than the opposition. Thus the detrimental effect of turning the ball over was slightly diminished by this extreme case.

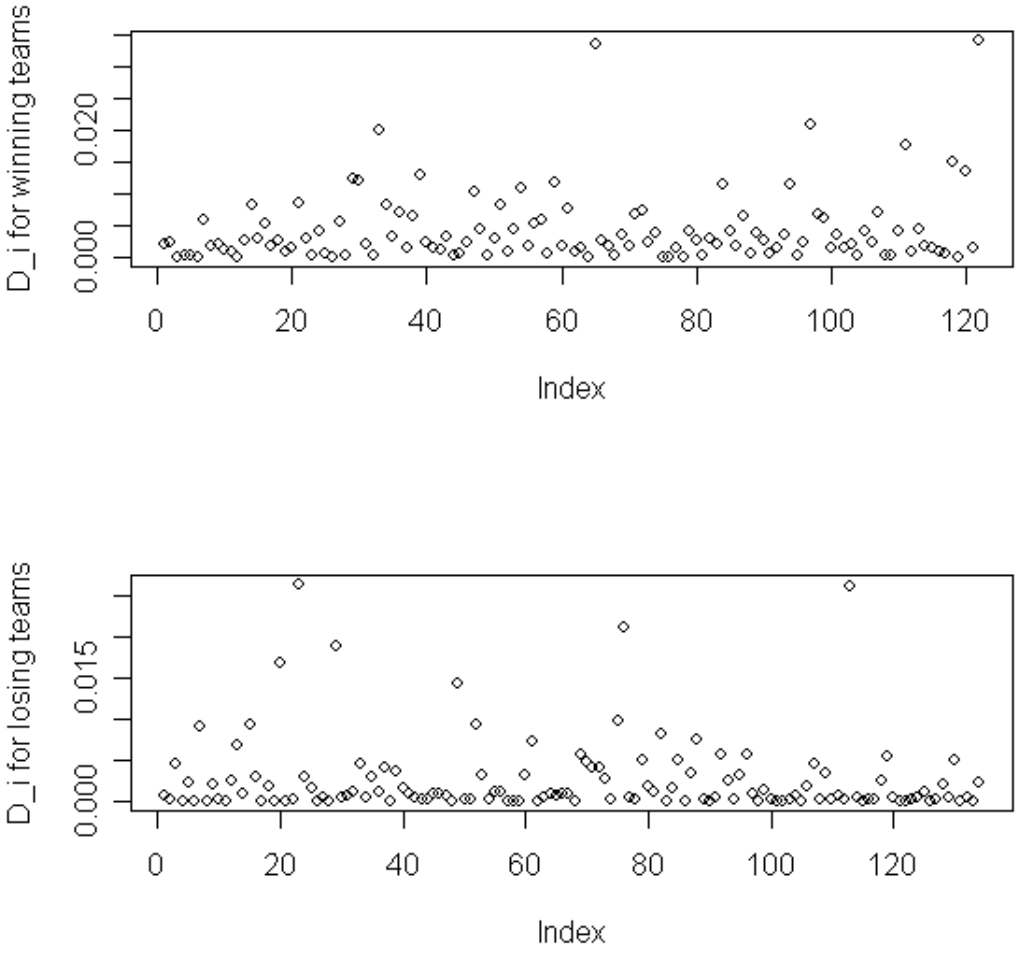

Figure 11: D_i^p Values for Winning and Loosing Teams

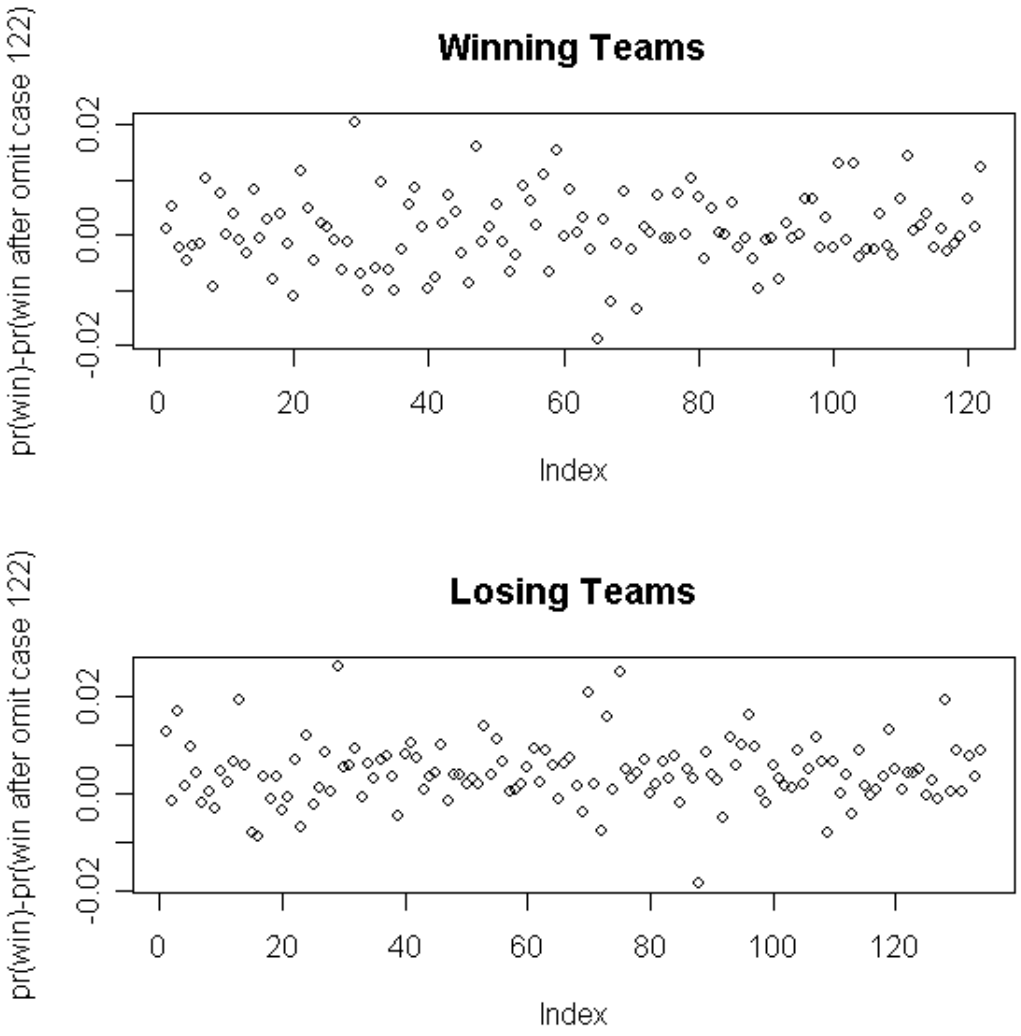

Figure 12: Differences in Predicted Success Probabilities After Removing Case 122

4 Simulation

The results of inference and prediction with the NFL data set were favorable towards the Bayesian procedure. To get an idea of which procedure may provide better results in general, both are applied to simulated data. Consider the stationary (Shumway and Stoffer 2000) $AR(2)$ time series:
$z_t = \frac{1}{6}z_{t-1} - \frac{5}{6}z_{t-2} + \epsilon_t$ and let $x' = (1, z_{199}, z_{198})$ with response $y = 1$ if $z_{200} > 0$ and $y = 0$ otherwise. Repeat 200 times to get a binomial data set with 200 observations. Then $\hat{\beta}$ and predicted probabilities are computed using both methods. This process was repeated 100 times giving 100 $v\hat{a}r\beta_i$ and 20000 $\hat{p}_i$ collectively acquired from the simulation with the following results.
Average variance of β from Bayesian procedure :

$$(1/100) * \sum_{i=1}^{100} v\hat{a}r(\beta_i) = \begin{bmatrix} 0.0156 \\ 0.0055 \\ 0.0144 \end{bmatrix}$$

Average variance of β from GLM:

$$(1/100) * \sum_{i=1}^{100} v\hat{a}r(\beta_i) = \begin{bmatrix} 0.0436 \\ 0.0157 \\ 0.0495 \end{bmatrix}$$

The mean squared error from Bayesian procedure:

$$(1/20000) * \sum_{i=1}^{20000} (\hat{p}_i - y_i)^2 = 0.14018$$

The mean squared error from GLM:

$$(1/20000) * \sum_{i=1}^{20000} (\hat{p}_i - y_i)^2 = 0.12954$$

The results of the two procedures are very similar. Bayesian inference produces $\beta's$ with smaller variance, while on average GLM predicted probabilities have a

slightly smaller error. It should be noted that to automate the simulation process, a general algorithm was developed to select a prior and design matrix for each simulated data set. This of course has an effect on inference and prediction. In practice the user has the ability to more carefully select the prior and design matrix and thus achieve better results as was the case with the NFL data.

5 Comments

There is a rather large body of literature related to modeling NFL data, but most research focuses on predicting the point spread of a future contest. Harville (1977, 1980) provides an extensive treatment of prediction for high school, college, and pro football using a mixed linear model, which assumes a Normal distribution for the point spreads. Thompson (1975) gives an MLE approach with similar Normal assumptions, while Glickman and Stern (1998) describe a MCMC analysis with Bayesian features. Most of the aforementioned procedures use only past point spreads and do not account for such factors as rushing and passing yards as described here. Of course, with effective predictors for the performance indicators used here one could use these results to predict point spreads with perhaps more accuracy than the previously mentioned methods since these performance indicators can give a more complete picture of a team's strengths and weaknesses. Football handicappers, such as drbobsports.com, have noticed some of the trends relating football success to such factors as rushing yardage and turn overs, yet it is unclear whether these conclusions were made numerically or intuitively.

The computations used in the Bayesian procedure outlined herein are relatively simple compared to the numerical maximization methods used for GLM, although the use of any statistical computation software renders this point mute. The software package used here was R, and there is no supplemental package that will perform the computations with the logit link function. There was therefore considerable effort put into writing the necessary code from scratch.

6 References

Ashburn, J. R. and Colvert, P. M. (2006), "A Bayesian Mean-Value Approach with a Self-Consistently Determined Prior Distribution for the Ranking of College Football Teams"

Bedrick, E. J., Christensen, R., and Johnson, W. (1997), "Bayesian Binomial Regression: Predicting Survival at a Trauma Center", The American Statistician, Vol. 51, No. 3, 211-218

Glickman, M. E. and Stern, H. S. (1998), "A State-space Model for National Football League Scores", *Journal of the American Statistical Association*, Vol.93, No. 441, 25-35

Harville, D. (1977), "The Use of Linear-Model Methodology to Rate High School or College Football Teams", *Journal of the American Statistical Association*, Vol. 72, No. 358, 278-289

Harville, D. (1980), "Predictions for National Football League Games Via Linear-Model Methodology", *Journal of the American Statistical Association*, Vol. 75, No. 371, 516-524

Kedem, B. and Fokianos, K. (2002), *Regression Models for Time Series Analysis*, John Wiley and Sons, Inc.

Leonard, T. (1972), "Bayesian Methods for Binomial Data", *Biometrika*, Vol. 59, No.3, 581-589

Pollard, R. (1973), "Collegiate Football Scores and the Negative Binomial Distribution", *Journal of the American Statistical Association*, Vol. 68, No. 342, 351-352

Rencher, A. C., (2000), *Linear Models in Statistics*, John Wiley and Sons, Inc.

Shumway, R. H. and Stoffer, D. S. (2000), *Time Series Analysis and Its Applications*, Springer-Verlag

Stern, H. (1991), "On the Probability of Winning a Football Game", *The American Statistician*, Vol. 45, No. 3, 179-183

Thompson, M. (1975), "On Any Given Sunday:Fair Competitor Orderings with Maximum Likelihood Methods", *Journal of the American Statistical*

Association, Vol. 70, No. 351, 536-541

Venables, W.N. and Ripley, B.D., (2002), *Modern Applied Statistics with S*, Springer-Verlag

www.drbobsports.com