Chase Joyner

882 Election Prediction Project

October 27, 2016

Model formulation

Here we document the model formulation process. To begin, we will consider the response variables

$$Y_{ij} = \begin{cases} 1, & i\text{th state votes Democrat in election year } j \\ 0, & i\text{th state votes Republican in election year } j. \end{cases}$$

The standard regression framework is no longer applicable here because it is not reasonable that the Y's are normally distributed, i.e. the errors are not normally distributed. Also, a model of the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

does not guarantee that the predicted value will be between 0 or 1. Therefore, we will consider a probit link model,

$$P(Y_{ij} = 1 \mid \mathbf{X}_{ij}) = \Phi(\mathbf{X}_{ij}\boldsymbol{\beta}),$$

where $\mathbf{X}_{ij}$ is a vector of covariates for the ith state during the jth election year. Additionally, we do not have the assumption that the Y's are independent because surrounding states have influence over each other. Specifically, if South Carolina votes Republican, then Georgia is likely to vote Republican. Therefore, we consider a spatial random effects model of the form

$$P(Y_{ij} = 1 \mid \mathbf{X}_{ij}) = \Phi(\mathbf{X}_{ij}\boldsymbol{\beta} + b_i)$$

where b_i represents the spatial random effect for state i in any election year, succinctly written as $\mathbf{b} = (b_1, ..., b_n)' \sim \text{CAR}(\tau^2, \rho)$, i.e. $\mathbf{b} \sim N\big(\mathbf{0}, \tau^2(\mathbf{D} - \rho\mathbf{W})^{-1}\big)$, where $\mathbf{D}$ is a diagonal matrix with entries $D_{ii} = \sum_{k=1}^{n} W_{ik}$, $W_{ik} = 1$ if the ith spatial unit is a neighbor of the kth spatial unit, and $W_{ik} = 0$ otherwise with the convention that $W_{ii} = 0$. Under this formulation, the likelihood is

$$L(\boldsymbol{\beta}, \mathbf{b} \mid \mathbf{Y}) = \prod_{i=1}^{51} \prod_{j=1}^{J} \big(\Phi(\mathbf{X}_{ij}\boldsymbol{\beta} + b_i)\big)^{Y_{ij}} \big(1 - \Phi(\mathbf{X}_{ij}\boldsymbol{\beta} + b_i)\big)^{1 - Y_{ij}},$$

where J is the number of election years being considered. However, we will have no hope of implementing a Gibbs sampler. To remedy this, we will perform a data augmentation step. Namely, consider

$$Z_{ij} \sim N(\mathbf{X}_{ij}\boldsymbol{\beta} + b_i, 1) \quad \text{where } z_{ij} > 0 \text{ if } Y_{ij} = 1, \text{ and } z_{ij} < 0 \text{ if } Y_{ij} = 0.$$

Then, we can reformulate our likelihood to the equivalent form of

$$L(\boldsymbol{\beta}, \mathbf{b}, \mathbf{Z} \mid \mathbf{Y}) = \prod_{i=1}^{51} \prod_{j=1}^{J} \phi(z_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta} - b_i) \Big[I(z_{ij} > 0, Y_{ij} = 1) + I(z_{ij} < 0, Y_{ij} = 0) \Big].$$

To capture data sets over different election years, we will implement a matrix $\mathbf{G}$ whose rows identify which state the observation belongs to. Therefore, we can control which random effect is being placed on observation i, j. As a result, our likelihood function becomes can now be written as

$$\begin{aligned} L(\boldsymbol{\beta}, \mathbf{b}, \mathbf{Z} \mid \mathbf{Y}) \propto & \exp\left\{ -\frac{1}{2} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{G}\mathbf{b})'(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{G}\mathbf{b}) \right\} \\ & \prod_{i=1}^{51} \prod_{j=1}^{J} \Big[I(z_{ij} > 0) I(Y_{ij} = 1) + I(z_{ij} < 0) I(Y_{ij} = 0) \Big]. \end{aligned}$$

To complete the model formulation, we impose the following prior specifications:

$$\begin{aligned} \boldsymbol{\beta} &\sim N(\mathbf{a}, \mathbf{R}) \\ \tau^2 &\sim IG(a_\tau, b_\tau). \end{aligned}$$

Denote $\boldsymbol{\Sigma} = \big(\mathbf{D} - \rho\mathbf{W}\big)^{-1}$. Combining all of this, the joint posterior distribution is

$$\begin{aligned} \pi(\boldsymbol{\beta}, \mathbf{b}, \tau^2, \mathbf{Z} \mid \mathbf{Y}) \propto & \exp\left\{ -\frac{1}{2} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{G}\mathbf{b})'(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{G}\mathbf{b}) \right\} \cdot \\ & \prod_{i=1}^{51} \prod_{j=1}^{J} \Big[I(z_{ij} > 0) I(Y_{ij} = 1) + I(z_{ij} < 0) I(Y_{ij} = 0) \Big] \\ & \times \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - \mathbf{a})' \mathbf{R}^{-1} (\boldsymbol{\beta} - \mathbf{a}) \right\} \\ & \times \big(\tau^2\big)^{-\frac{51}{2}} \exp\left\{ -\frac{1}{2\tau^2} \mathbf{b}' \boldsymbol{\Sigma} \mathbf{b} \right\} \times \big(\tau^2\big)^{-a_\tau - 1} \exp\left\{ -\frac{b_\tau}{\tau^2} \right\}. \end{aligned}$$

By inspection, we see that the latent variables have the marginal posterior distribution

$$Z_{ij} \mid \boldsymbol{\beta}, b_i, \mathbf{Y} \sim \begin{cases} TN\big(\mathbf{X}_{ij}\boldsymbol{\beta} + b_i, 1, (0, \infty)\big), & \text{if } Y_{ij} = 1 \\ TN\big(\mathbf{X}_{ij}\boldsymbol{\beta} + b_i, 1, (-\infty, 0)\big), & \text{if } Y_{ij} = 0. \end{cases}$$

The posterior distribution for the regressors is

$$\pi(\boldsymbol{\beta} \mid \text{else}) \propto \exp\left\{ -\frac{1}{2} \Big[(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{G}\mathbf{b})'(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{G}\mathbf{b}) + (\boldsymbol{\beta} - \mathbf{a})' \mathbf{R}^{-1} (\boldsymbol{\beta} - \mathbf{a}) \Big] \right\}.$$

After some algebra, we find that $\boldsymbol{\beta} \mid \mathbf{Z}, \mathbf{b}, \mathbf{Y} \sim N(\boldsymbol{\mu}, \mathbf{C})$, where

$$\begin{aligned} \boldsymbol{\mu} &= (\mathbf{X}'\mathbf{X} + \mathbf{R}^{-1})^{-1} \big(\mathbf{R}^{-1}\mathbf{a} + \mathbf{X}'(\mathbf{Z} - \mathbf{G}\mathbf{b}) \big) \\ \mathbf{C} &= (\mathbf{X}'\mathbf{X} + \mathbf{R}^{-1})^{-1}. \end{aligned}$$

The posterior distribution of the spatial random effects is

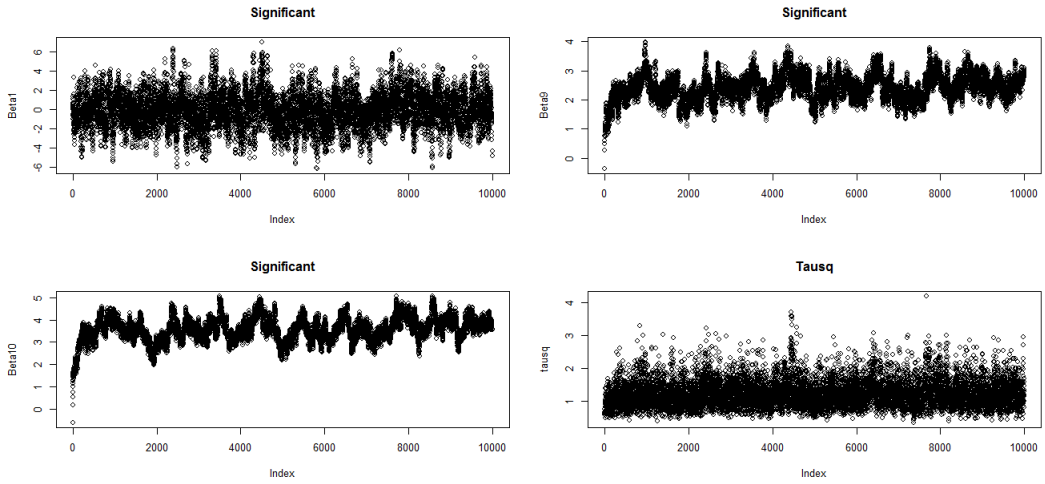
$$\pi(\mathbf{b} \mid \text{else}) \propto \exp\left\{-\frac{1}{2\tau^2}\left[\tau^2(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\mathbf{G}\mathbf{b})'(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta}-\mathbf{G}\mathbf{b})+\mathbf{b}'\boldsymbol{\Sigma}\mathbf{b}\right]\right\}.$$

Therefore, $\mathbf{b} \mid \boldsymbol{\beta}, \mathbf{Z}, \tau^2 \sim N\big((\tau^2\mathbf{G}'\mathbf{G}+\boldsymbol{\Sigma})^{-1}\big(\tau^2\mathbf{G}'(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta})\big), (\tau^2\mathbf{G}'\mathbf{G}+\boldsymbol{\Sigma})^{-1}\big)$. Lastly, the posterior of τ^2 is easily seen as

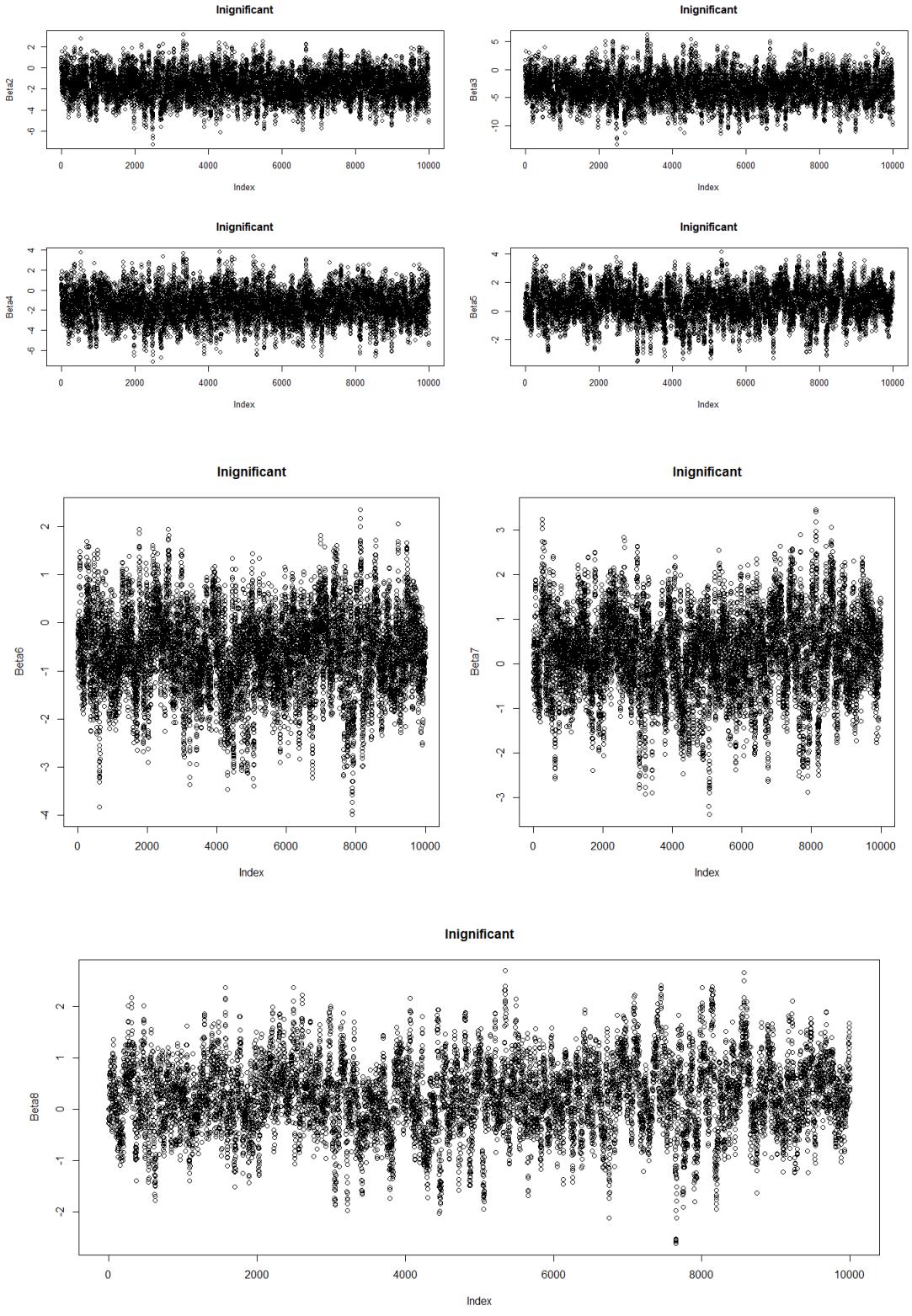
$$\tau^2 \sim IG\left(\frac{51}{2}+a_\tau, \frac{1}{2}\mathbf{b}'\boldsymbol{\Sigma}\mathbf{b}+b_\tau\right).$$

Simulation study

To assess our model's ability of distinguishing significant covariates and insignificant covariates, a simulation study was performed. The true covariates from the subsequent data application section were used. The probability of success, i.e. $P(Y_{ij}=1 \mid X_{ij})$, was generated for all i, j and a set of observations were generated from this. We performed the Gibbs sampler algorithm developed in the model formulation section for each data set and repeated this process for 250 total data sets. To assess the mixing of a single chain, below are the MCMC plots $\boldsymbol{\beta}$ and τ^2. The first four plots are the significant covariate coefficients with a true value of $1, 2$, and 3, and τ^2 set to 1.



The next seven plots are for the insignificant coefficients.



Overall, it appears that the chain has converged for all parameters and is mixing very well. The significant parameters appear to be picked up and the insignificant parameters stay close to 0 and do not deviate away. Therefore, the model seems to be doing a good job of flagging significant covariates and not flagging insignificant covariates. To back this claim, for each data set, we reported the arithmetic mean of the last 5000 estimates from the Gibbs sampler. After doing this for all 250 data sets, we averaged these arithmetic means to obtain:

Parameter	True	Estimate
β_1	1	1.770
β_9	2	2.815
β_{10}	3	4.209
τ^2	1	1.201

These estimates give rise to the fact that the model is doing a decent job at identifying the significant covariates and estimating their true values on the correct magnitude. It is worth noting that the largest estimate for the insignificant covariates was 0.465 and the smallest estimate was 0.131.

Data application

The data consists of the past 5 elections, i.e. 255 observations. The covariates that were available are education, race, median household income, and unemployment rates for all states in the 5 past elections. More specifically, for education we considered for each state the (x_1) percentage of an advanced degree, the (x_2) percentage of only Bachelor's degree, the (x_3) percentage of only High School diploma, and the (x_4) percentage of no High School diploma. For race, we considered the (x_5) percentage of those that are white, (x_6) percentage of black, (x_7) percent hispanic, and (x_8) percent asian. These covariates were kept constant across the 5 election years as a result of lacking data and would probably not change drastically across years. For (x_9) median household income and (x_{10}) unemployment rate, we used available data to include the income and unemployment rate for each state during each election year. This constitutes our design matrix. Our goal is to predict the outcomes for this election year. Running the Gibbs sampler, we get the mean estimates

$$\begin{array}{ll} \beta_1 = 1.212 & \beta_2 = -0.517 \\ \beta_3 = -0.223 & \beta_4 = -0.754 \\ \beta_5 = 0.655 & \beta_6 = 0.319 \\ \beta_7 = 0.719 & \beta_8 = 1.462 \\ \beta_9 = -0.320 & \beta_{10} = 0.511. \end{array}$$

We report the 95% HPD intervals of these parameters, which are

Parameter	HPD interval	Parameter	HPD interval
β_1	$[-2.059, 4.609]$	β_2	$[-2.724, 1.764]$
β_3	$[-4.617, 4.473]$	β_4	$[-3.505, 1.939]$
β_5	$[-1.187, 2.488]$	β_6	$[-0.991, 1.636]$
β_7	$[-0.619, 2.076]$	β_8	$[0.159, 2.882]$
β_9	$[-0.703, 0.057]$	β_{10}	$[0.194, 0.845]$.

Based on the HPD intervals, the only significant covariates are the percent of asian people in the state and the unemployment rate. This goes slightly against intuition. Of course, unemployment

is expected to be related to voting, and also race. However, the other races should be significant as well, and probably would have different voting habits. However, we did not consider model selection and therefore use these parameter estimates to generate the predictions. We used the covariates available to us for this years presidential election to generate

$$P(Y_i = 1 \mid \mathbf{X}_i) = \Phi(\mathbf{X}_i\boldsymbol{\beta} + b_i)$$

for $i = 1, ..., 51$ using the last 50,000 iterates of $\boldsymbol{\beta}$ and b_i. Therefore, for each state we obtain $50,000$ samples of $P(Y_i = 1 \mid \mathbf{X}_i)$ and then we take the average of these. The outcomes are as follows:

$$\begin{array}{lll}
P(Y_1 = 1 \mid \mathbf{X}_1) = 0.138 & P(Y_2 = 1 \mid \mathbf{X}_2) = 0.120 & P(Y_3 = 1 \mid \mathbf{X}_3) = 0.605 \\
P(Y_4 = 1 \mid \mathbf{X}_4) = 0.105 & P(Y_5 = 1 \mid \mathbf{X}_5) = 0.980 & P(Y_6 = 1 \mid \mathbf{X}_6) = 0.525 \\
P(Y_7 = 1 \mid \mathbf{X}_7) = 0.998 & P(Y_8 = 1 \mid \mathbf{X}_8) = 0.961 & P(Y_9 = 1 \mid \mathbf{X}_9) = 0.999 \\
P(Y_{10} = 1 \mid \mathbf{X}_{10}) = 0.704 & P(Y_{11} = 1 \mid \mathbf{X}_{11}) = 0.328 & P(Y_{12} = 1 \mid \mathbf{X}_{12}) = 0.994 \\
P(Y_{13} = 1 \mid \mathbf{X}_{13}) = 0.137 & P(Y_{14} = 1 \mid \mathbf{X}_{14}) = 0.979 & P(Y_{15} = 1 \mid \mathbf{X}_{15}) = 0.492 \\
P(Y_{16} = 1 \mid \mathbf{X}_{16}) = 0.607 & P(Y_{17} = 1 \mid \mathbf{X}_{17}) = 0.199 & P(Y_{18} = 1 \mid \mathbf{X}_{18}) = 0.472 \\
P(Y_{19} = 1 \mid \mathbf{X}_{19}) = 0.199 & P(Y_{20} = 1 \mid \mathbf{X}_{20}) = 0.980 & P(Y_{21} = 1 \mid \mathbf{X}_{21}) = 0.975 \\
P(Y_{22} = 1 \mid \mathbf{X}_{22}) = 0.998 & P(Y_{23} = 1 \mid \mathbf{X}_{23}) = 0.978 & P(Y_{24} = 1 \mid \mathbf{X}_{24}) = 0.855 \\
P(Y_{25} = 1 \mid \mathbf{X}_{25}) = 0.104 & P(Y_{26} = 1 \mid \mathbf{X}_{26}) = 0.491 & P(Y_{27} = 1 \mid \mathbf{X}_{27}) = 0.069 \\
P(Y_{28} = 1 \mid \mathbf{X}_{28}) = 0.088 & P(Y_{29} = 1 \mid \mathbf{X}_{29}) = 0.928 & P(Y_{30} = 1 \mid \mathbf{X}_{30}) = 0.849 \\
P(Y_{31} = 1 \mid \mathbf{X}_{31}) = 0.992 & P(Y_{32} = 1 \mid \mathbf{X}_{32}) = 0.755 & P(Y_{33} = 1 \mid \mathbf{X}_{33}) = 0.998 \\
P(Y_{34} = 1 \mid \mathbf{X}_{34}) = 0.267 & P(Y_{35} = 1 \mid \mathbf{X}_{35}) = 0.016 & P(Y_{36} = 1 \mid \mathbf{X}_{36}) = 0.690 \\
P(Y_{37} = 1 \mid \mathbf{X}_{37}) = 0.043 & P(Y_{38} = 1 \mid \mathbf{X}_{38}) = 0.960 & P(Y_{39} = 1 \mid \mathbf{X}_{39}) = 0.944 \\
P(Y_{40} = 1 \mid \mathbf{X}_{40}) = 0.993 & P(Y_{41} = 1 \mid \mathbf{X}_{41}) = 0.076 & P(Y_{42} = 1 \mid \mathbf{X}_{42}) = 0.025 \\
P(Y_{43} = 1 \mid \mathbf{X}_{43}) = 0.243 & P(Y_{44} = 1 \mid \mathbf{X}_{44}) = 0.155 & P(Y_{45} = 1 \mid \mathbf{X}_{45}) = 0.101 \\
P(Y_{46} = 1 \mid \mathbf{X}_{46}) = 0.964 & P(Y_{47} = 1 \mid \mathbf{X}_{47}) = 0.780 & P(Y_{48} = 1 \mid \mathbf{X}_{48}) = 0.958 \\
P(Y_{49} = 1 \mid \mathbf{X}_{49}) = 0.315 & P(Y_{50} = 1 \mid \mathbf{X}_{50}) = 0.932 & P(Y_{51} = 1 \mid \mathbf{X}_{51}) = 0.089.
\end{array}$$

We will use the idea that if $P(Y_i = 1 \mid \mathbf{X}_i) > 0.5$, then state i will vote Democrat this year.