*Permanent Visa Application Certification in the United States: Bayesian vs. Frequentist Approaches*

Jackson Crum | Group 6

**Introduction**

Immigration and the place of immigrants in our society is the focus of a significant amount of the political rhetoric and debate in our current government. With the current administration’s “America First” agenda being highly anti-immigrant, it may become more difficult for foreign workers to attain a permanent labor certification and work in this country legally. For many immigrants, this labor certification is the only means to a better, safer, more prosperous life. As the process of attaining US citizenship is paved with bureaucracy and long waiting periods, labor certification and permanent residency (green card) provide the best opportunities for many hopeful immigrants. This project will attempt to determine the variables that increase or decrease the probability of a permanent labor certification application being denied.

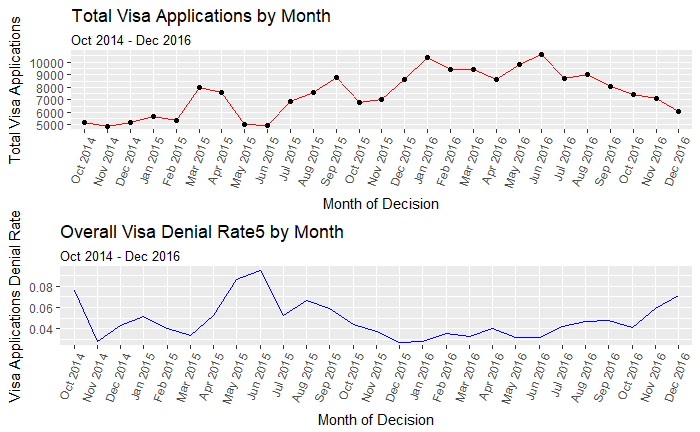
A permanent labor certification issued by the Department of Labor (DOL) allows an employer to hire a foreign worker to work permanently in the United States. The DOL must certify to the U.S. Citizenship and Immigration Services that there are not sufficient U.S. workers able, willing, qualified and available to accept the job opportunity in the area of intended employment and that employment of the foreign worker will not adversely affect the wages and working conditions of similarly employed U.S. workers.[[1]](#footnote-1)

The data covers 2012-2017 and includes 374,362 applications with 161 variables, including information on employer, company size, position, offered wage, job posting history, employee education and past visa history, associated lawyers, and final decision. Final decision, certified or denied, is the variable the project is attempting to predict.

**Data Cleaning and Exploratory Data Analysis**

The dataset was very messy with regard to the number of missing values, excessive and redundant variables, and mislabeled data. Overall, there were 33,669,345 missing values, with over half of the variables missing from the majority of observations. After removing the unnecessary or majority NA variables and removing rows with any NA values for the remaining variables, the workable dataset consisted of 202,203 observations with 14 variables. Discrepancies between state labels (abbreviations versus full names) were resolved in Excel. After initial EDA, some observations had obviously mislabeled pay units as determined by the excessive pay amount. Those observations with pay units that could be identified were fixed, the others were removed. Any observation with state labels outside the United States or in US territories were removed. Applications labelled “certified-expired” were converted to “certified” and 8,000 withdrawn applications were also removed.

All variables were explored to understand the dataset and determine possible trends. In total, 95.2% (185,017 out of 194,307 applications) were certified. Of the total applications, 58.5% (113,633 applications) came from India, with the next most being 7.9% (15,324 applications) coming from China, and only 19 countries having more than 1,000 applications. Of all of the countries, those with the highest denial rates mostly came from Central and South America. California had the most applicants, with 24.3% (47,218) of the total applications, followed by Texas at 14% (27,286). Mississippi, New Mexico, and Louisiana had the highest denial rate, at 16.4%, 13.5%, and 12.6%, respectively. When, as seen in *Figure 1,* decision date was plotted both against total applications and overall denial rate by month, it was noticed that there appears to be an inverse relationship between total applications and overall denial rate. As such, total visas in the month of decision was created and added as a variable. After exporting this dataset for later training, all categorical variables were set to dummy variables (323 in total) and all numeric data was normalized to put them all on the same scale. Case status was set to 0 for denial and 1 for certification to create a logistic equation with a Bernoulli distribution.



(Upper) Total Visas by Month (Lower) Denial Rate by Month

*Figure 1*

**Frequentist Models**

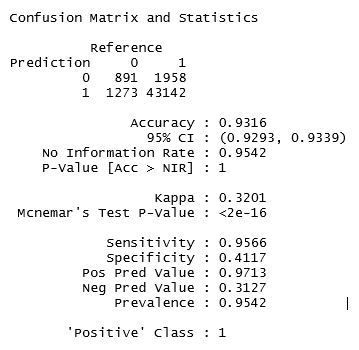
The first models chosen for training, testing, and evaluation were a series of increasing simplistic logistic regression models. The logistic model takes a linear combination of the variables and assigned weight coefficients to each variable to adjust their influence on the outcome, a value between 0 and 1. Logistic regression is ideal as the predicted variable in this case is binary, meaning it only takes two values, a 0 for denied and a 1 for certified, and there are multiple variables with influence. The first logistic model used was a kitchen sink containing all the 322 predictor variables (Full Logistic). After this model was run, a function was written to extract all important variables with a p-value < 0.01 and a second logistic regression was run containing only these 30 variables (Subset Logistic).

Finally, a stepwise logistic regression was run on the subset model. The stepwise regression runs a logistic regression and then provides the AIC (Akaike Information Criterion) score the model would have if each variable or no variable were removed. The variable that creates the worst scored model is removed the logistic regression is run again. This process continues until removing no variable would result in the best model. As can be seen by the metric scores, all logistic models perform well in terms of accuracy ((TP + TN) / (TP + TN + FP +FN)), sensitivity (TP / TP + FN), and positive predictive value (TP / (TP + FP)), but perform very poorly in terms of specificity (TN / (TN + FP)) and negative predictive value (TN / (TN + FN)).

In these model, variance inflation factors (VIFs) are used to measure collinearity in the logistic models. As a rule-of-thumb, economists usually drop variables with a VIF of greater than five to remove multicollinearity between variables. When run on the full model, the vif() function returns an error and states that there are alias coefficients in the model, implying that the model contains perfect collinearity. When tested on the significant variables model, all multicollinearity has been removed from the model with the exception of the collinearity between pay amount and unit of pay, which is expected. Further analysis on both variables was desired, so both were left in future models. As this dataset was used on all further models, it can be confidently stated that there is little multicollinearity in the models.

Based on the coefficients shown in *Figure 4*, the model claims that the three most important variables in determining certification are whether or not the applicant is receiving an hourly wage, how much pay they receive, and company size, with all increasing the log odds as they grow. Surprisingly, American citizens have the highest negative coefficient of statistically significant variables, which means the model claims that being an American decreases your log odds the most for receiving certification.

Following the logistic models, two XGBoost (Extreme Gradient Boosting) models were developed to see if they would have improved predictive power. XGBoost is an optimized distributed gradient boosting library that is ideal for this sort of dataset. It grows classification decisions trees one after other and attempts to reduce misclassification rate in subsequent iterations. The next tree is built by giving a higher weight to misclassified points by the previous tree, so the model slowly learns over many iterations while utilizing parallel computing to provide fast training and regularization to avoid overfitting. The model uses gradient descent to optimize the loss function by tuning different values of coefficients to minimize the error. XGBoost models are ideal for sparse matrices. As this dataset is very large and mostly dummy variables, this is an ideal model for the problem.



*Figure 2*

The first XGBoost model used the xgboost function, a simple interface for XGBoost models, while the second model used the xgb.train function, a more advanced interface. The classification threshold was set at 0.8. While both models maintained the high accuracy, sensitivity, and positive predictive value of the logistic models, there is a noticeable improvement in both specificity and negative predictive value, meaning these models are able to better predict application denials than the logistic models, with XGBoost1 being the best performing model. XGBoost1 can plot variable importance (*Figure 4)* and has somewhat different results than the logistic model results on the matter, with, among others, being Indian, being well educated, and applying in California increasing your log odds while applying to Texas employer decreasing your log odds.

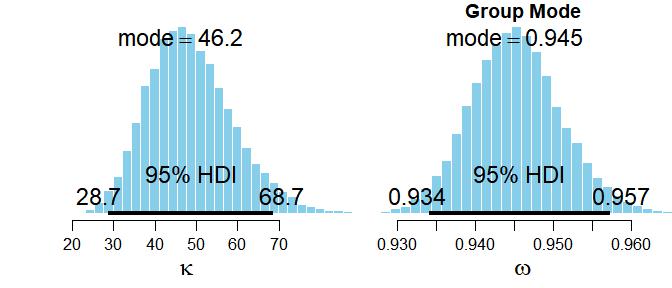
As shown by the metric scores below and the confusion matrix (*Figure 2*), there is still a significant amount of uncertainty in frequentist models. While the models are able to predict certification with high accuracy, with the best specificity being only 41.17% and the best negative prediction value being only 47.48%. These scores were both recorded with a 0.80 threshold, meaning that when told to score any new observation with a probability of 0.80 or less as “denied”, the best models were able to correctly predict less than half the time.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Sensitivity** | **Specificity** | **Pos Pred Value** | **Neg Pred Value** |
| Full Logistic | 0.9315 | 0.9655 | 0.2246 | 0.9629 | 0.2378 |
| Subset Logistic | 0.8987 | 0.9280 | 0.2860 | 0.9644 | 0.1602 |
| Stepwise Logistic | 0.9362 | 0.9735 | 0.1594 | 0.9602 | 0.2240 |
| XGBoost2 | 0.9316 | 0.9566 | 0.4117 | 0.9713 | 0.3127 |
| XGBoost1 | 0.9527 | 0.9831 | 0.3175 | 0.9678 | 0.4748 |

**Bayesian Models**

Given the amount of prior uncertainty there is in the dataset regarding the importance of variables and the parameter values, hierarchical models were developed. The hierarchical models are ideal for this sort of problem as they allow for parameters to account for uncertainty by being assigned prior distributions that are influenced by prior knowledge and create long MCMC chains to adequately explore the sample space.

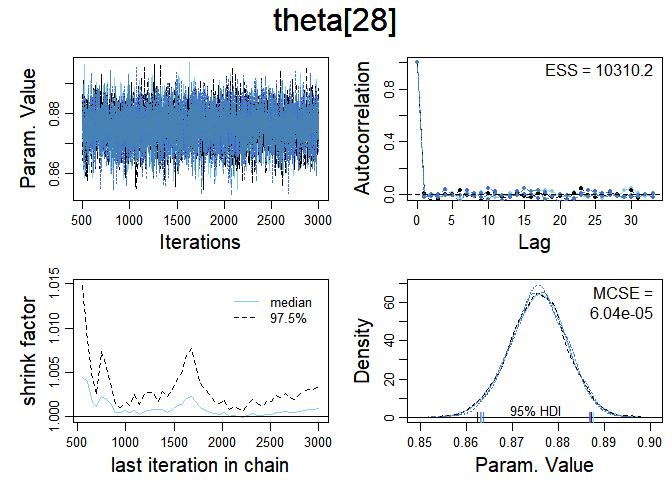
Three Bayesian models were developed with JAGS to examine both nominal and metric variables in the dataset. The first Bayesian model (Jags-Model-2.R) is a basic hierarchical model to determine the outcome from a single categorical variable. The likelihood function is a Bernoulli distribution and the prior distribution is a beta density distribution. The shape parameters of the beta distribution are re-expressed with the mode ω and the concentration κ of the beta density. The mode ω is dependent upon a prior beta distribution beta( 8 , 2 ) that represents the prior belief that all visa applications have a high probability (heavily right-skewed beta distribution) of certification, while the concentration is given a gamma distribution. The value of κ was updated to a more reasonable gamma distribution after initial runs. The results of this model give posterior distributions for individual factors in the chosen variable along with diagnostic information, including plotting the MCMC chain, effective sample size, and posterior distribution with the 95% HDI. The results of the running the model are very promising, with all runs having a very ESS (Effective Sample Size). For example, in *Figure 3*, results show that the model adequately explored the sample space for Mexico (ESS = 10310.2) and gave a posterior distribution with and HDI between 0.862 and 0.888, well below the group mode distribution with the HDI between 0.934 and 0.957. This model also allows for comparisons between multiple levels of a variable, such as between countries, states, or education levels. Some comparisons can be seen below in *Figure 5*. As this model shows, there is statistically significant probability distributions between some factors in country of citizenship, state of employment, and unit of pay, with the highest variation (difference between individual 95% HDIs and group HDI) coming from country of citizenship, especially from Central and South American nations.



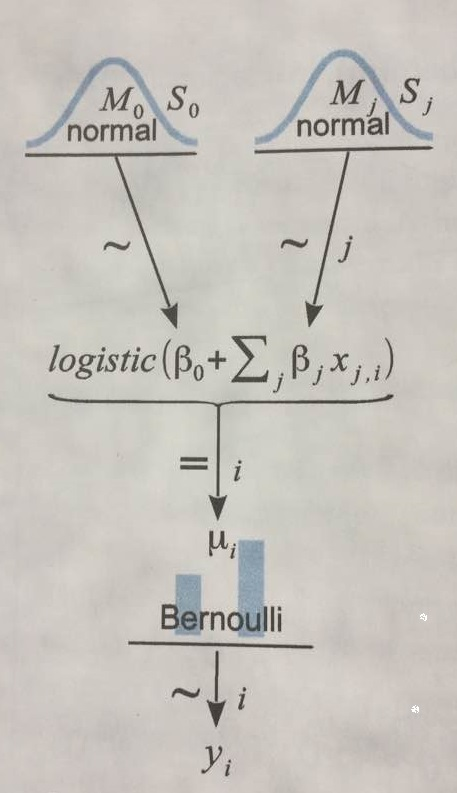
(Upper) Diagnostics of Mexico

(Lower) Country of citizenship parameter distributions

*Figure 3*



The second model (Jags-Model-1.R) is a hierarchical logistic ANOVA-like model that shows that the modal certification rate of the categorical variable *j* (ω**j**) is a logistic function of a baseline certification rate β0 plus a deflection for the variable β[j]. All variables are given the same parameter distributions, 10,000 MCMC chain steps, and a 1,000-step burn-in period. Data from each variable was manipulated in their own tables in Excel with three columns, one for variable name, one for total certified applications, and one for total applications, for use in this model. Some comparison results from this model can be seen below in *Figure 5* in the plots of the posterior beta distributions across category. For country of citizenship, most have very concentrated posteriors with modes near the group mode HDI interval. However, countries such as Ecuador, Jamaica, Mexico, Poland, and the United States have wider dispersion towards lower probabilities, implying they might have statistically significant lower certification rates and less certainty about their posterior distributions. The same can be claimed for state of employment, with most state posteriors being tightly concentrated around the group mode with a few, including Mississippi, Alaska, Montana, and Hawaii, concentrating more towards lower certification rate, though much less dispersion than in country of citizenship. The posteriors for level of education, however, have very low concentration and thus more uncertainty in their results. While bachelor’s and doctorate have modes that concentrated more towards the overall mode, as expected from the frequentist models, master’s is more dispersed while degrees skew towards lower rates of certification.



The third model (Jags-Model-4.R) is a hierarchical model designed to test the metric values in the dataset, specifically pay amount and number of employees, two variables the frequentist models deemed highly influential. The model follows the structure shown on the right. This model took sample of 20,000 observations and only 5,000 MCMC steps to make it runnable. A linear combination of the metric predictors is mapped to a probability value via the logistic function. The intercept and slope parameters (β0, β1, and β2) are each given normal distributions with mean = 0 and sd = ¼. This model is good for this situation as it results is a binary outcome from a Bernoulli distribution and will account for the combined influence of the variables on the outcome as well as accounting for uncertainty by giving the beta values vague prior distributions. The initial run of the model shows some correlation between the two variables. One major drawback of this model is that the majority of the predicted values are 1, which, as seen in the diagnostics in *Figure 7* by the plot and wide coefficient HDI’s, leads to a wide decision boundary distribution and ambiguous parameter estimates. As this model is robust, it allows for randomness by including a guess parameter that could correctly estimate the visa application by chance.

**Conclusions and Comparison of Methods**

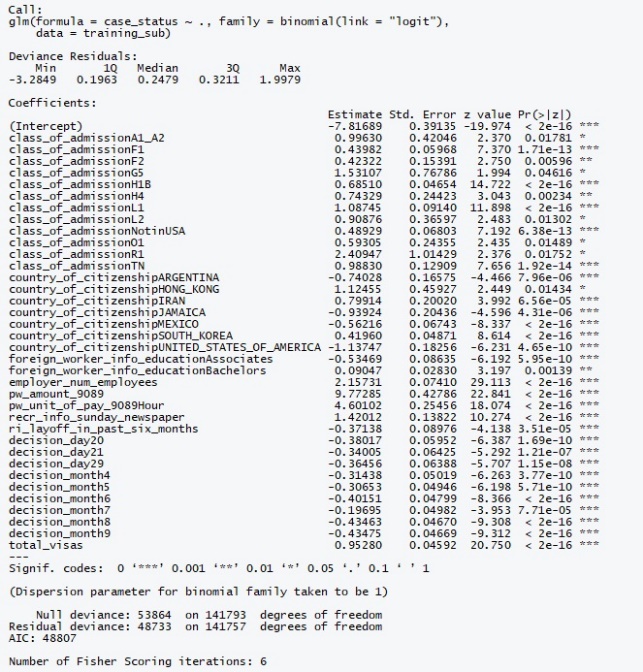
This type of problem is ideal for a Bayesian approach to statistical analysis. Frequentist models do not account for previous knowledge or prior distributions, rather focusing entirely on the likelihood function. As shown in this project, the p-value is difficult to interpret and over thirty variables had some significance, leading to possible overfitting. The lack of predictability the frequentist models have is a clear sign that the uncertainty in the data is too much for the frequentist models to account for. The Bayesian models allow for including prior knowledge and giving probability distributions to parameters to account for the inherit uncertainty of the data. The hierarchical models allow for the accounting of the further uncertainty by developing more and more layers of probability distributions of prior knowledge of parameter distributions.

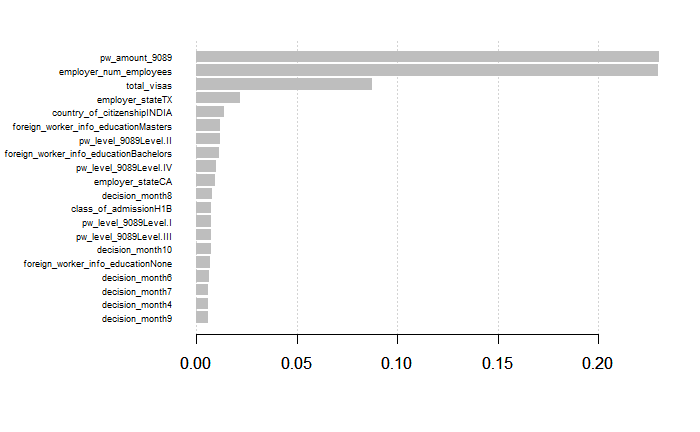
From the coefficients from the logistic regressions, some basic conclusions can be drawn. Larger salaries and larger companies correlated with a higher rate of certification. State and country of citizenship have some effect on the probability of certification, with employers in Texas having a lower rate and employers in California and applicants from India having a higher rate of certification. Having a higher education results in a higher certification rate, though having a doctorate does not improve one’s chances from masters or bachelors in a statistically significant way.

The Bayesian models show some similar results with other posterior probabilities that show statistically significant results. Ecuador, Mexico, Jamaica, Thailand, and the United States citizens have a statistically probable lower rate of certification than the group. Mississippi, Hawaii, and Louisiana have a statistically probable lower rate of certification than the group, though state posterior distributions are tighter than country posterior distributions. Higher education is less predictive in the Bayesian models than the frequentist models calculate. Company size has a higher correlation with certification probability than pay amount, which also has a significant positive correlation with certification probability as shown by the coefficient distributions. Number of applications in the month of decision was shown to have no impact on certification probability.

While the Bayesian models captured more of the uncertainty in the data, a more robust model would be needed to include a linear combination of all variables. A more complete dataset may lead to more certainty in the Bayesian models.

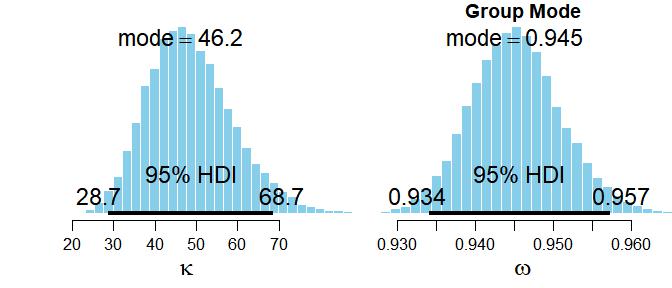
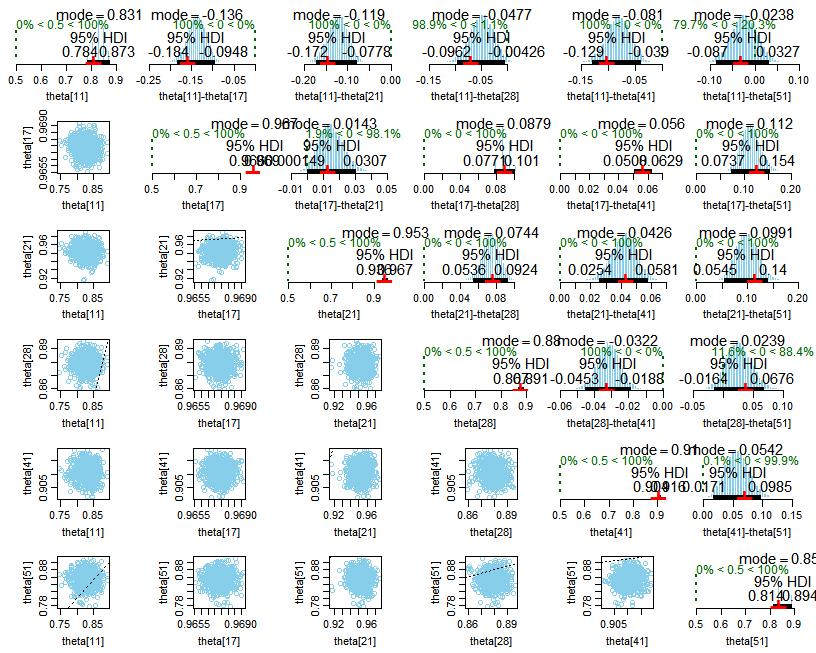
**Figures**





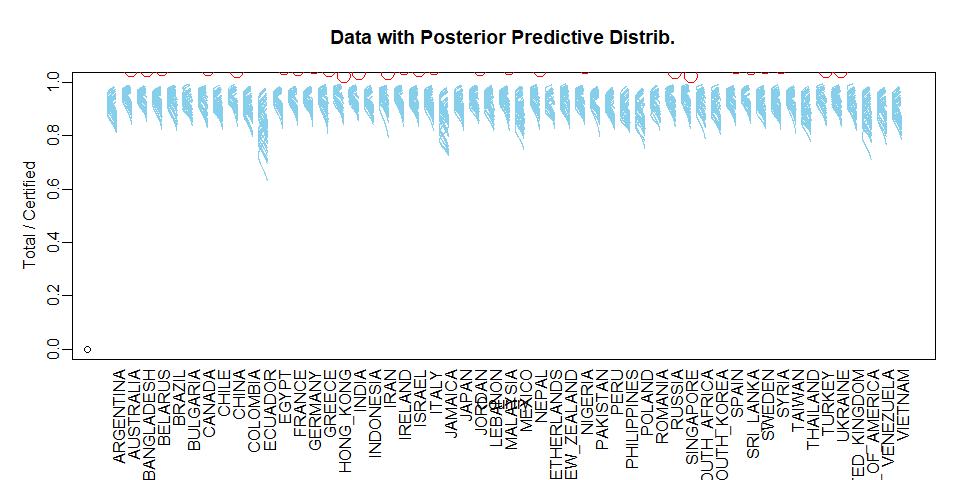
(Upper) Variable metrics of subset logistic model (Lower) Variable importance as derived by XGBoost model

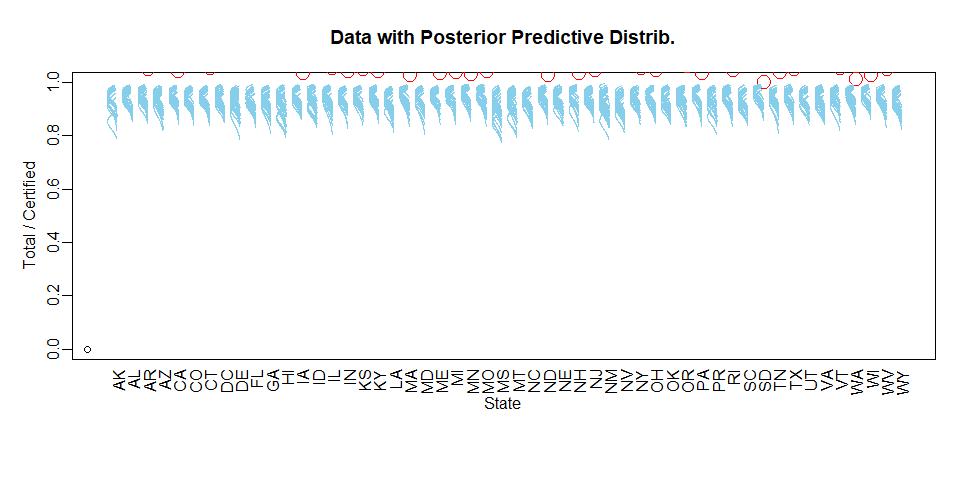
*Figure 4*

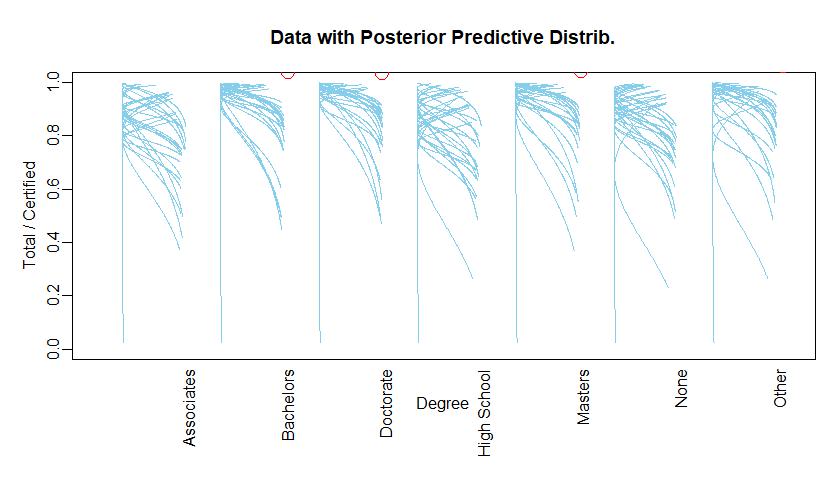


Comparison for Country of Citizenship: Ecuador (11), India (17), Israel (21), Mexico (28), South Korea (41), and USA (51)

*Figure 5*



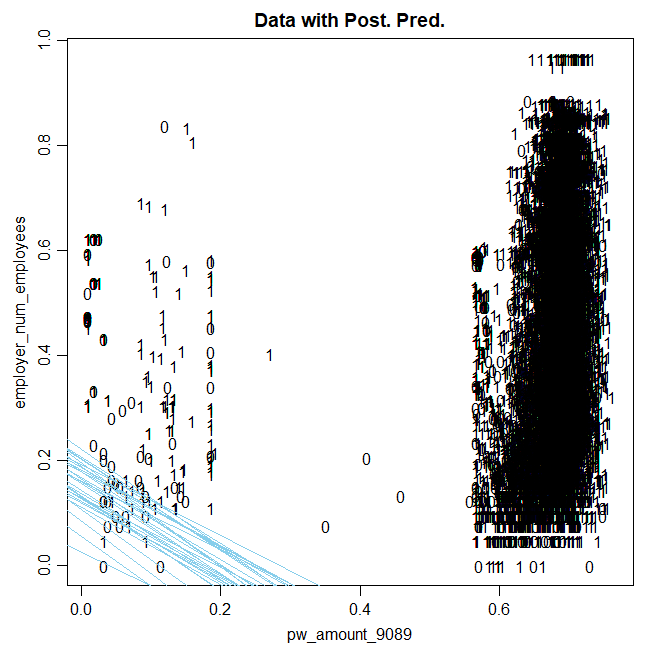
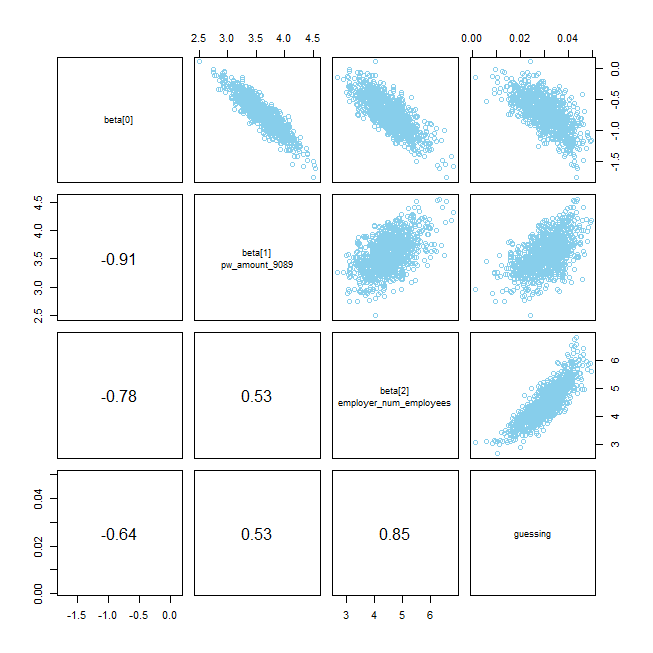


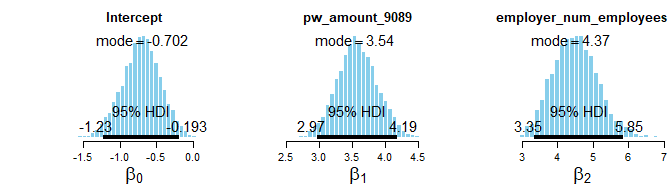


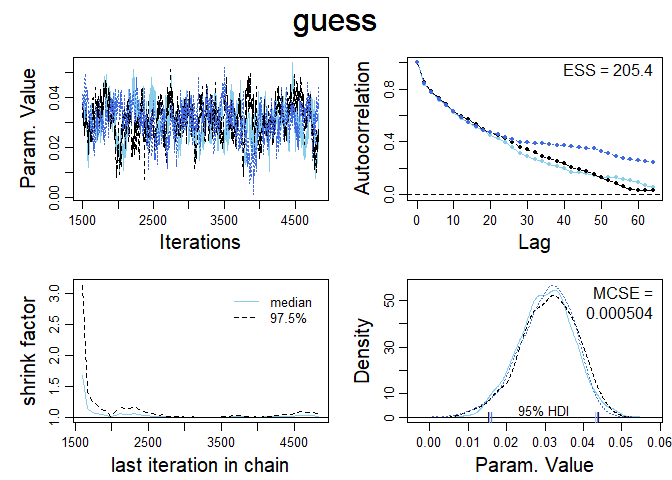
(Upper) Posterior distributions of countries of citizenship (Middle) Posterior distributions of employer states

(Lower) Posterior distributions of levels of education

*Figure 6.*





Diagnostics of linear combination of pay amount and company size

*Figure 7*

**References**

1. Permanent Labor Certification. Accessed March 07, 2018. <https://www.foreignlaborcert.doleta.gov/perm.cfm>
2. Kruschke, John K. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. 2nd ed. London: Elsevier, 2015.

1. Permanent Labor Certification. Accessed March 07, 2018. https://www.foreignlaborcert.doleta.gov/perm.cfm. [↑](#footnote-ref-1)