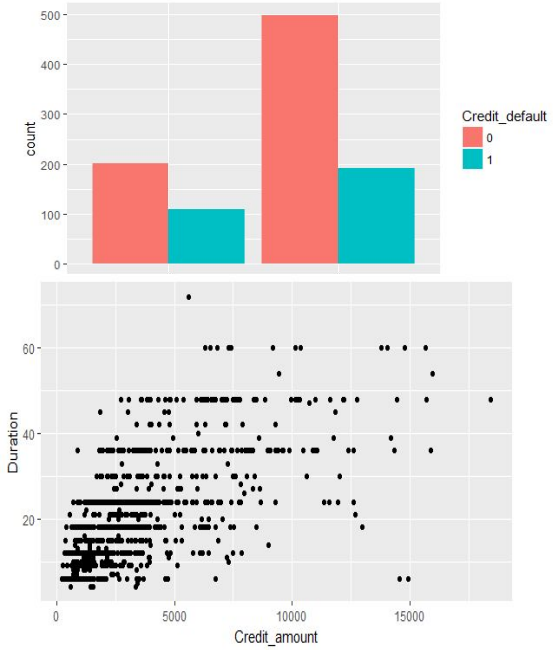
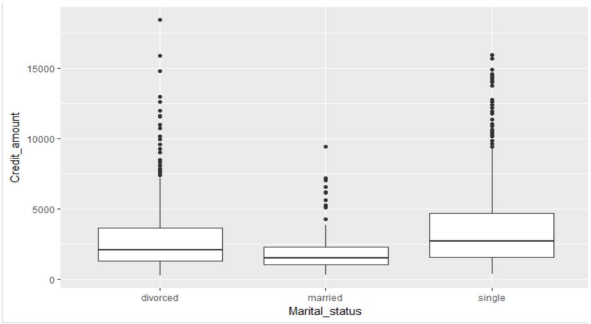
Bayesian Method for Data Science
Instructor: Yuxiao Huang
Group 2: Krystin Sinclair, Ye-in Jeon, Christian Cleber Masdeval Braz

UCI German Dataset

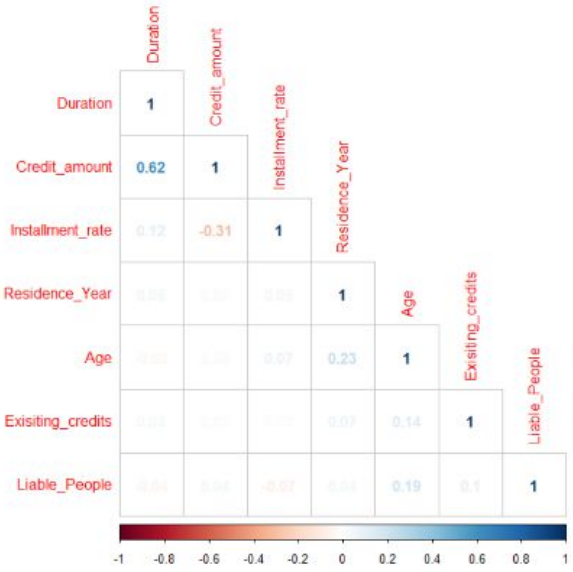
The data that we have chosen to work with is regarding German credit data. This data has almost 1,000 observations and 21 attributes. The main variable of interest is whether or not the client would be classified as good credit or bad credit. A secondary response variable is credit amount. The attributes are the status of a checking account, how long the account has been open for, credit history, purpose of the loan, if employed, how long they have held a job, savings amount, marital status, residence, and age, among others. Banks would be interested in what type of applicant will default and applicants would be interested in how much they can ask for.

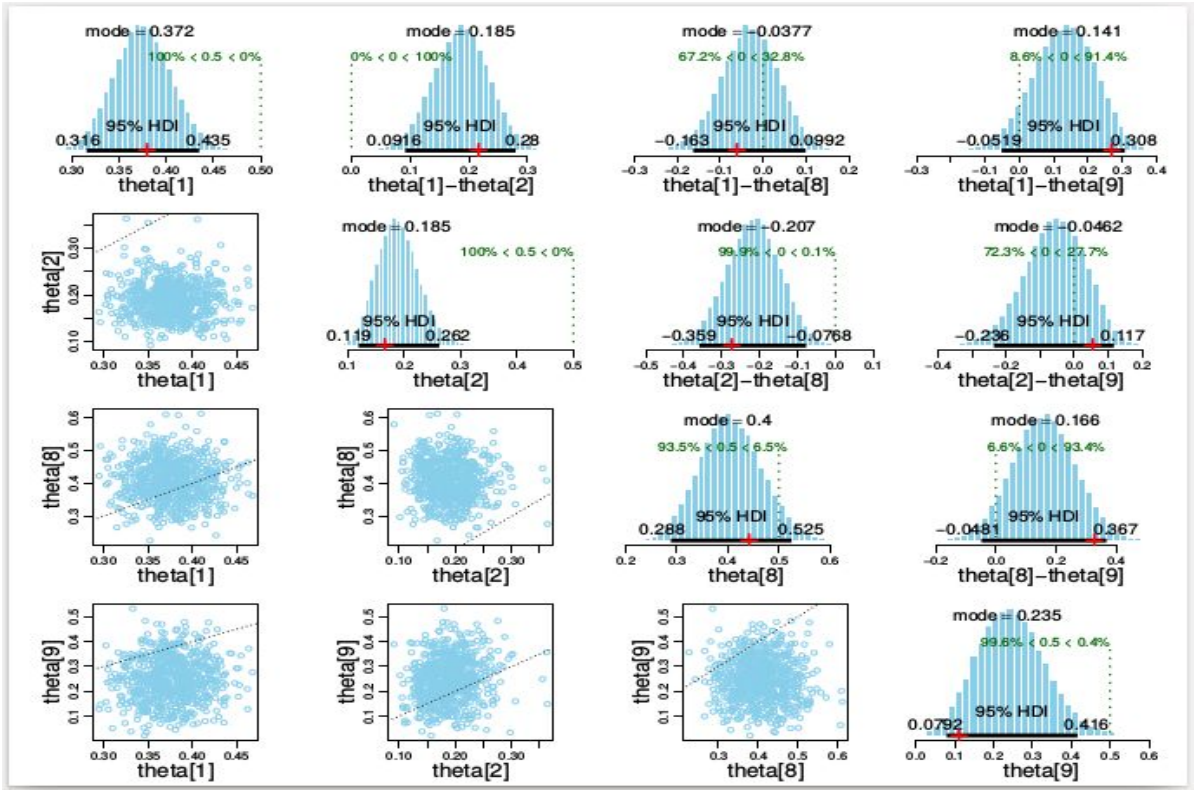
We began by doing initial exploratory data analysis to understand the data. We ran both structure and summary functions in R. Then we used bar plots, correlation matrix, scatter plots and boxplots to understand the different variable types that we had. Below we have a bar plot showing the gender based on credit default. For credit default 1 means default and 0 means no default. We can see that there are more male applicants and that the majority of the applicants did not default. Next, there is a duration scatter plot by credit amount. This shows that they have a positive correlation. Next, there is a marital status and credit amount box plot. The credit amount is fairly similar between single, divorced and married. However, there does appear to be less variance amongst the married crowd. The correlation matrix confines this theory, showing a correlation of **.62**. This is the highest correlation of all of the numerical attributes.



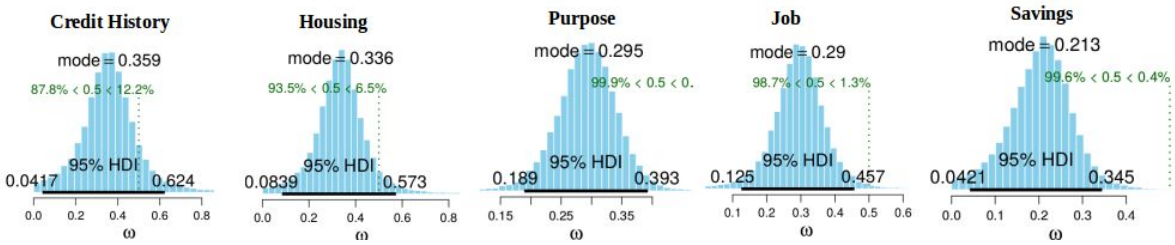


We use MCMC to assess the likelihood of default based on the different categories within categorical variables. Doing that, we can get a sense of how likely is a subject to default and also the tendency of default of a whole group. The figure shows the result for the category Purpose and the mode of four possible loan purposes: 1 - Car (new), 2 - Car (used), 8 - Education, and 9 - Retraining. As one would expect, as the classes are unbalanced and 70% of the examples are non-defaulters and only 30% are defaulters, all the modes are close to the default rate of the dataset. We can see from Case 1, that the mode of





the probability of someone of this subgroup default is 0.37, and also that no shrinkage is occuring, as it matches the proportion in the data (as indicated by the red cross). On the other hand, Case 9 presents a noticeable shrinkage away from the raw proportion towards the mode of the group (which is 0.3). Notice also that the widths of their 95% HDIs is broad. This occurs because the little amount of data for this specific subgroup, leading it to be strongly influenced by the metrics of the group and in a wider uncertainty. Finally, Case 8 is a subject that the financial institution should pay attention. Despite its wide HDI, it has a high proportion of defaulters and is the only one with a credible 0.5 estimative. Moreover, the posterior estimate of the differences between it and Case 2 and (almost) Case 9, excluded zero even with the shrinkage pulling their individual estimates toward the mode, what reinforces the precaution.

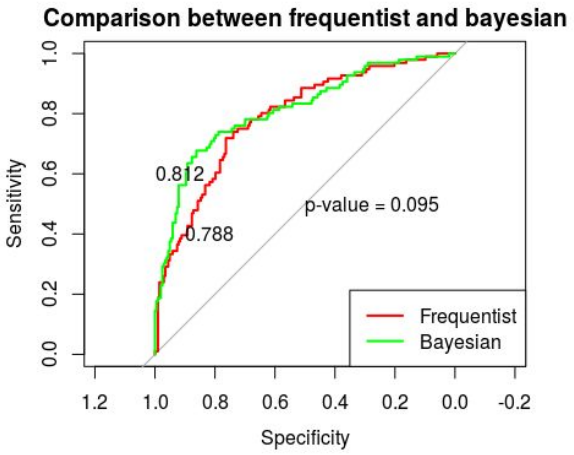


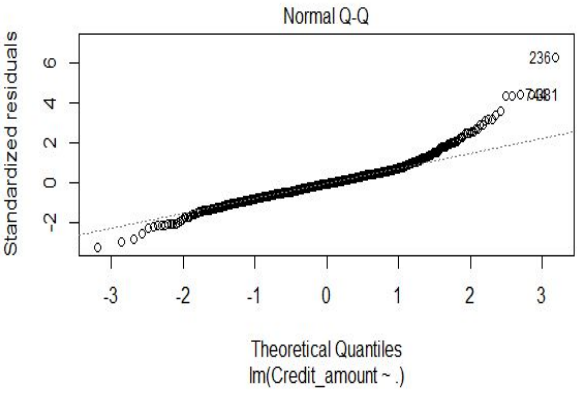
Organized from higher modes to smaller ones, five categorical features with qualitative information about bank account holders. We can see that they variate a lot and that some categories seems to concentrate more information about past defaulters. All these analysis would help to enlighten which client would be more propense to default than other.

In this section we develop a model to predict the probability of whether a person will default or not based on the data. To accomplish this, we first fit a **logistic regression** model using the *glm* function available in R. The overall accuracy for the frequentist model is about **73%** and tells how often this classifier is correct. The precision is **0.63** and measures how often the model is correct in predicting a positive outcome. The recall is **0.44** and tells us when it is actually positive, how often it predicts as positive. This simple model has modest statistics and as our main objective is solely using it as a frequentist estimation to be able to compare it with a bayesian approach, we do not going to try further improvements. Anyway, it is worth noting that in this problem we are more concerned about the false negative rates (recall) instead of the true positives rates (precision). This is because a bank (or any other financial institution lending money to a untrustworthy party) is interested in to minimize the chance of not being paid back the borrowed amount. So, in this setting, it is more problematic to state that a person is reliable when it is not (false negative), instead of saying that a person is not reliable and, in fact, it is (false positive). A way to handle this problem using logistic, is modifying the threshold for the decision of classifying as positive. In other words, one can increase the rate of false positives to decrease the rate of false negatives. This is possible lowering the decision threshold from 0.5, for example, to 0.3. Hence, any probability outcome from the logistic model greater than 0.3 would be classified as a "default risk". In our following attempt, we evaluate the same problem, predicting the chance of a person default to pay a loan or not, using the **bayesian logistic regression** approach. To be able to estimate the parameters of the linear function of the logistic model, i.e., the posterior of our bayesian inference, we rely on the MCMC strategy. A snippet of the model written using the software JAGS is

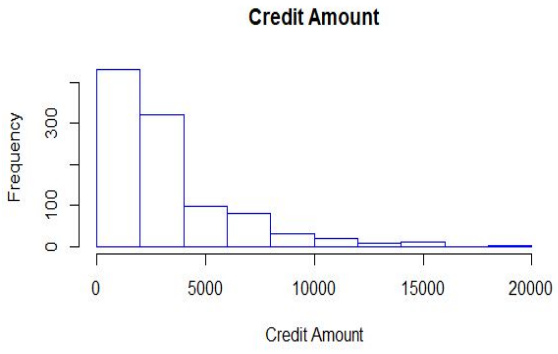
```
model {
      for ( i in 1:Ntotal ) {
      # In JAGS, ilogit is logistic:
      y[i] ~ dbern( mu[i] )
      mu[i] <- ( guess*(1/2) + (1.0-guess)*ilogit(zbeta0+
                          sum(zbeta[1:Nx]*zx[i,1:Nx])) )
    }
    # Priors vague on standardized scale:
    zbeta0 ~ dnorm( 0 , 1/2^2 )
    for ( j in 1:Nx ) {
      zbeta[j] ~ dnorm( 0 , 1/2^2 )
    }
    guess ~ dbeta(1,9)
  }
```

also presented. You can observe that we tried to build a robust model against outliers, once we are employing the guessing variable. This model generates the estimates for all the coefficients, which we then use to make predictions using the *sigmoid* function. To compare the result, we use the same train/test set created before. The bayesian approach shows a better performance than the frequentist, with an overall accuracy of **80%**, a recall of **0.55** and a precision of **.77**. Finally, we plot the ROC curve of the two models. This is a commonly used graph that summarizes the performance of a classifier over all possible thresholds. It is generated by plotting the True Positive Rate (y-axis) against the False Positive Rate (x-axis) as you vary the threshold for assigning observations to a given class. We print the AUC (Area Under the Curve) for both models. An AUC higher than 0.5 means that the classifier is better than chance. Here, both have a good AUC with the bayesian classifier slightly better. The p-value informs that the difference between them is not significant.



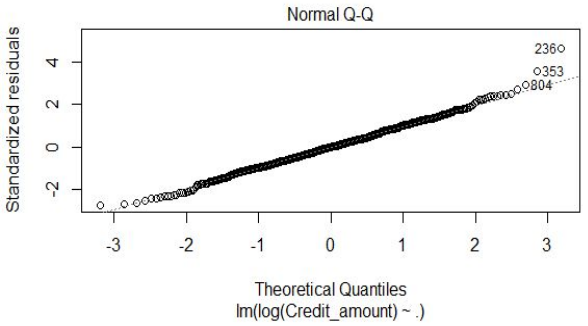


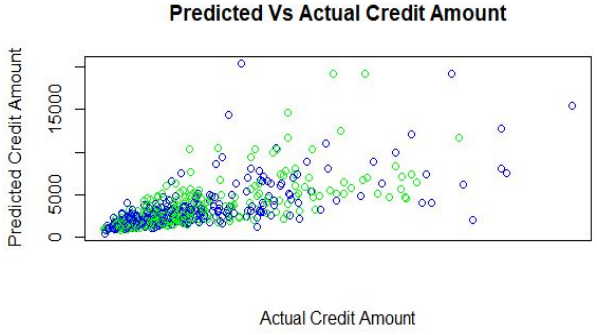
The regression problem explores how the attributes relate to credit amount. This analysis began by using a linear model and all of the attributes. The results showed an adjusted R-squared value of **.6**. This is an okay result but we would like a model that explains more of the variance. The Q-Q plot of the residuals of this model is shown to the left. An ANOVA test was performed to see which attributes were the most important when it comes to credit amount. A p-value of .001 was used to determine which attributes would be deemed important and kept in the model. This left the model with six attributes: status checking, duration, purpose, installment rate, property, and job. A test of normality was performed on credit amount and it was found to be heavily skewed, as showed in the histogram. This led the next model to utilize the log function to normalize the data. The final frequentist regression model only used the six main attributes and the log of credit amount. This resulted in a model with an adjusted R-squared of **.63**, which is slightly better than the initial model.

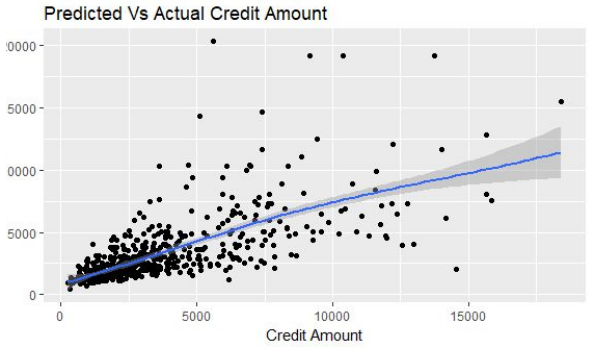


The Q-Q plot of this new model is presented and we can verify that, after applying the log transformation on the response variable, the distribution of the residuals is much more normal. After

making feature selection and adjusting the data, we performed a Bayesian Analysis using the same attributes and with credit amount being the response variable. BPM, BMA, MPM and HPM were all utilized. It was found that BPM was the best model based on the RMSE values. The next step of the analysis is to compare the Bayesian and Frequentist method. The best model for the frequentist method had an RMSE of **2080.457** while the best model of the Bayesian Analysis had a RMSE of **2019.237**. The scatterplots below are of the best model available which is the Bayesian BPM.







In conclusion, there is much advice that can be given to both banks and loan applicants. For banks, they should be wary of giving loans to the unemployed. Interestingly, renters are less likely to default than homeowners. As for loan applicants, the applicants that receive the higher credit amount were by those that have management positions or jobs with high qualifications and when the purpose of the loan is to purchase a car. The analysis and use of a variety of methods showed that Bayesian Analysis is a very useful tool. It provided better models than frequentist method and for future research Bayesian Analysis should be utilized.

References

Dataset

http://home.cse.ust.hk/~qyang/221/Assignments/German/

Articles

https://loans.usnews.com/beyond-credit-scores-factors-that-affect-a-loan-application

https://studentloanhero.com/featured/personal-loan-purpose-happens-change/

https://www.cbsnews.com/news/5-things-that-can-torpedo-your-mortgage-application/

http://www.ijbf.uum.edu.my/images/pdf/5no1ijbf/6ijbf51.pdf

https://www.sciencedirect.com/science/article/pii/0883902688900183

https://dl.acm.org/citation.cfm?id=131259

https://www.emeraldinsight.com/doi/pdfplus/10.1108/eb013696

http://www.rcmloan.com/credit-building-solana-beach-ca/