Urine Data Analysis in Frequentist and Bayesian Framework

Zhishan Li

Introduction

The urine data frame contains 79 urine specimens which are analyzed to study which of the physical characteristics of the urine might be related to the formation of calcium oxalate crystals. In this dataset, there are 6 physical characteristics of the urine and they are gravity, pH, osmolarity, conductivity, urea concentration, and calcium concentration. In addition, including the indicator of the presence of calcium oxalate crystals, r, 7 total columns are listed in this dataset. Below, is a summary table for explaining each variable in the urine dataset.

Variable Name	Interpretation
r	Indicator of the presence of calcium oxalate crystals; r=0 indicates no calcium oxalate crystals are present, and r=1 indicates the presence of calcium oxalate crystals
gravity	Specific Gravity of the urine sample
ph	pH value of the urine sample
osmo	Osmolarity of the urine. It is proportional to the concentration of molecules in the solution
cond	Conductivity of the urine. It is proportional to the concentration of charged ions in the solution
urea	The urea concentration in millimoles per liter
Calc	The calcium concentration in millimoles per liter

The project aims to analyze and predict the presence of calcium oxalate crystals (r) in patients using the physical characteristics of their corresponding urine samples in both frequentist and Bayesian framework.

Exploratory Data Analysis

Scanning through the dataset through R, two NA values (missing values) are found in the dataset; after cleaning all the NA values, the data frame’s dimension reduces to 77 rows and maintains all 7 columns. By creating the pairwise scatterplot between variables (Figure 1), some information regarding the nature of the dataset are provided.

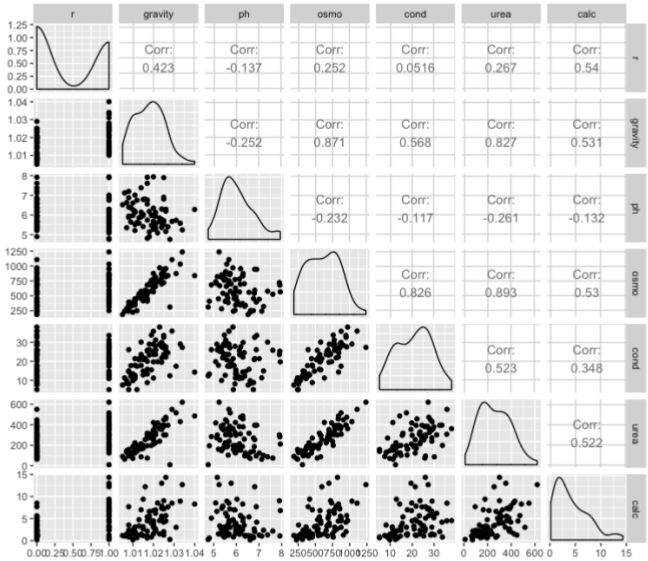


Figure 1: pairwise scatterplot between variables

By the scatterplot, it is easy to find out that there are some strong correlations between variables. For example, the positive correlation between gravity and osmolarity(about 0.871), between gravity and urea(about 0.827), and more; also, there are some negative correlations between pH and some other variables, such as gravity(-0.252), conductivity(-0.117), osmolarity(-0.232), urea(-0.261), calcium(-0.132). More important, for the response (r), the indicator of the presence of calcium oxalate crystals, it can be seen that the result is a binary response with a value of either 0 or 1; the response r (whether there is a presence in calcium oxalate crystal) is positively correlated with all the other variables, except for pH.

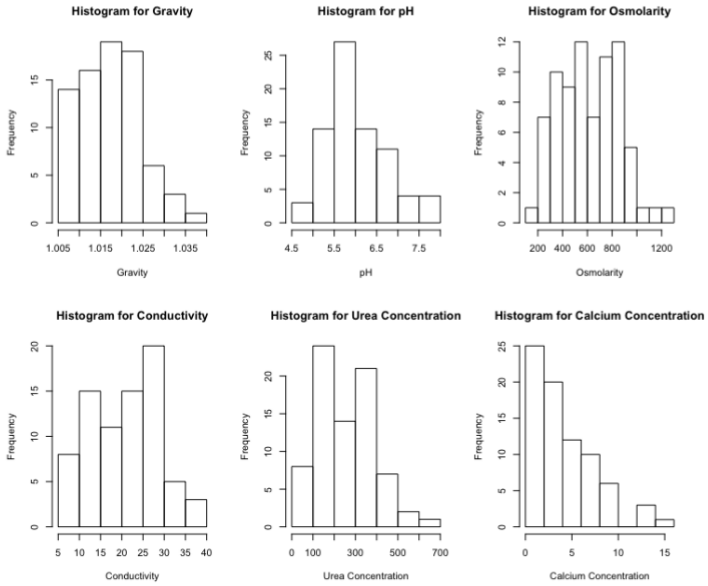
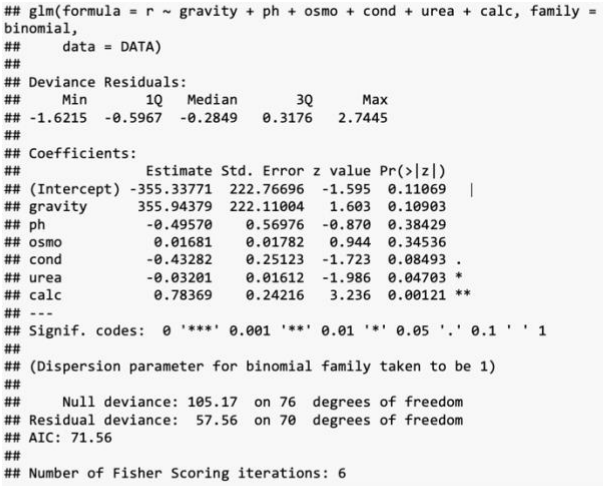


Figure 2: Histograms for each of the six features in urine

On the other side, when looking through the histograms(Figure 2) for each physical characteristics of urine (the 6 features), the highest frequencies in each variables' range is from 1.015 to 0.020 for gravity, from 5.5 to 6.0 for pH, from 550 to 600 and 800 to 900 for osmolarity, 25 to 30 for conductivity, 100 to 200 for urea concentration, and 0 to 2.5 for calcium. Overall, both scatterplots and histograms provided some intuitions about what is going on between the response, r, and the variables.

Methods

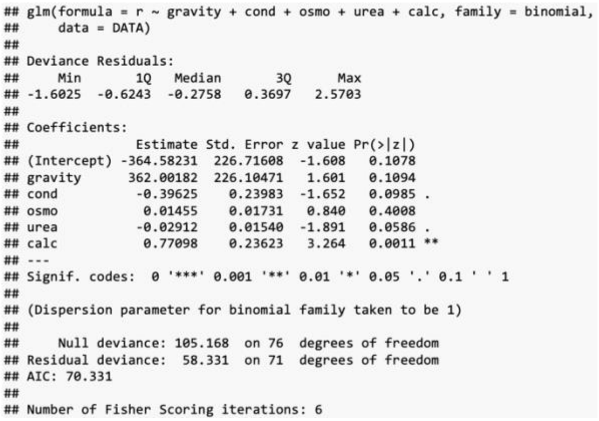
In this project, there are mainly two models used: Frequentist and Bayesian models. Under the frequentist regression model, we can utilize the generalized linear model (glm) to yield the formula (formula = r~ gravity + ph + osmo + cond + urea + calc). Since the response, r, is a binary response, we set the family to equal to binomial (family = binomial). After getting the summary function (Figure 3) from glm, we get to see the variables' coefficient estimates, standard error, z-values, and p-values; and now we can check which factors/variables have a significant effect on r.



```
## glm(formula = r ~ gravity + ph + osmo + cond + urea + calc, family =
binomial,
##     data = DATA)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6215  -0.5967  -0.2849   0.3176   2.7445
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -355.33771  222.76696  -1.595  0.11069   |
## gravity      355.94379  222.11004   1.603  0.10903
## ph            -0.49570    0.56976  -0.870  0.38429
## osmo           0.01681    0.01782   0.944  0.34536
## cond          -0.43282    0.25123  -1.723  0.08493 .
## urea          -0.03201    0.01612  -1.986  0.04703 *
## calc           0.78369    0.24216   3.236  0.00121 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 105.17  on 76  degrees of freedom
## Residual deviance:  57.56  on 70  degrees of freedom
## AIC: 71.56
##
## Number of Fisher Scoring iterations: 6
```

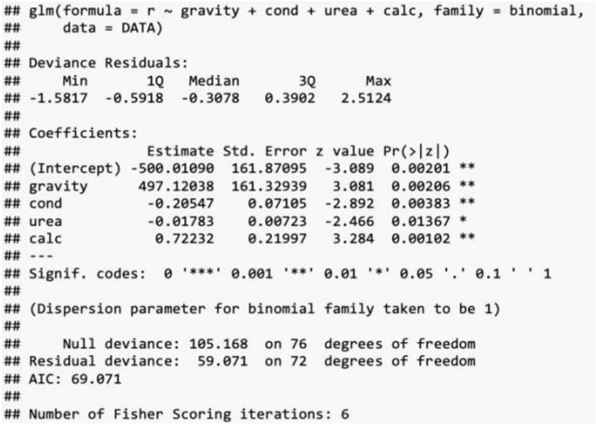
Figure 3: statistics summary information in glm with all predictors

Differences in AIC values as we remove variables from our model provide evidence for choosing one set of variables over another. The lower the AIC value the better the model is. Thus, by removing one predictor which has the highest p-value at a time, and then making another glm model until we found the lowest AIC will result in the best possible glm model from the dataset. In figure 3, the highest p-value is 0.38429 which means we dropped the "ph" predictor and made another glm and its summary function (Figure 4).



```
## glm(formula = r ~ gravity + cond + osmo + urea + calc, family = binomial,
##     data = DATA)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6025  -0.6243  -0.2758   0.3697   2.5703
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -364.58231  226.71608  -1.608   0.1078
## gravity      362.00182  226.10471   1.601   0.1094
## cond          -0.39625    0.23983  -1.652   0.0985 .
## osmo           0.01455    0.01731   0.840   0.4008
## urea          -0.02912    0.01540  -1.891   0.0586 .
## calc           0.77098    0.23623   3.264   0.0011 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 105.168  on 76  degrees of freedom
## Residual deviance:  58.331  on 71  degrees of freedom
## AIC: 70.331
##
## Number of Fisher Scoring iterations: 6
```

Figure 4: statistics summary information in glm with all predictors except "ph"



```
## glm(formula = r ~ gravity + cond + urea + calc, family = binomial,
##     data = DATA)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5817  -0.5918  -0.3078   0.3902   2.5124
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -500.01090  161.87095  -3.089  0.00201 **
## gravity      497.12038  161.32939   3.081  0.00206 **
## cond          -0.20547    0.07105  -2.892  0.00383 **
## urea          -0.01783    0.00723  -2.466  0.01367 *
## calc           0.72232    0.21997   3.284  0.00102 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 105.168  on 76  degrees of freedom
## Residual deviance:  59.071  on 72  degrees of freedom
## AIC: 69.071
##
## Number of Fisher Scoring iterations: 6
```

Figure 5: statistics summary information in glm with all predictors except "ph" and "osmo"

In figure 4, the summary function for the glm model which contains all the predictors except the “ph” predictor. Repeating the same method, we can finally end up having the lowest AIC, 69.071, by using only gravity, conductivity, urea, and calcium concentration in the glm model (see figure 5). Therefore, the final glm for the response r is:

$$Log(\frac{\theta}{1-\theta})=-500.01090+497.12038*gravity-0.20547*cond-0.01783*urea+0.72232*calc$$

Now, let’s look at figure 7. After checking the residuals and normal Q-Q plot, we can see the glm model is not fitting our data well enough; as we see that there is a heavy tail in the upper tail for the normal Q-Q plot, so it might not be the best try for our dataset. However, since we would like to use the logisticRegressionBayes function that is provided in class, we will continue analyzing the dataset in Bayes inference under the assumption of a normal prior.

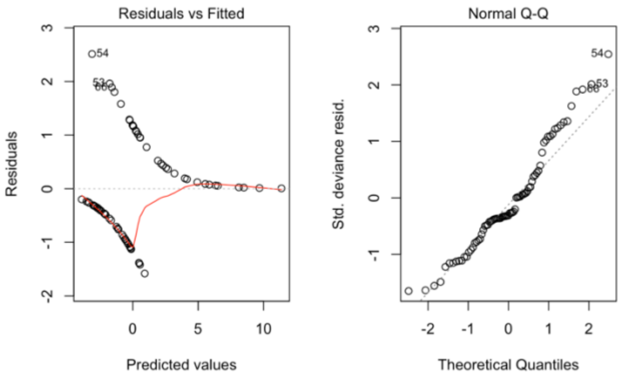


Figure 7: Residual plots and normal Q-Q plots for model glm

Switching to Bayesian framework, we now generalize a logistics regression model in Bayes. By generating 50,000 samples from the posterior distribution of a logistic regression using a Metropolis algorithm for all 6 predictors (the logisticRegressionBayes function), then we plot the trace plots for all six predictors, it turns out that there are lots of fluctuations (not stable at all) for four out of six predictors no matter it has 1 billion or 500 thousand sample sizes(there is an issue on convergence). Thus, at this time, we will try to drop out some variables after using bayesglm function in R and comparing their estimation and p-values. After several trials, it ends up the 4 predictors (ph, conductivity, urea, and calcium concentration) generate some stationary trace plots (see figure 8 for reference) by discarding the first 5,000 burn-in.

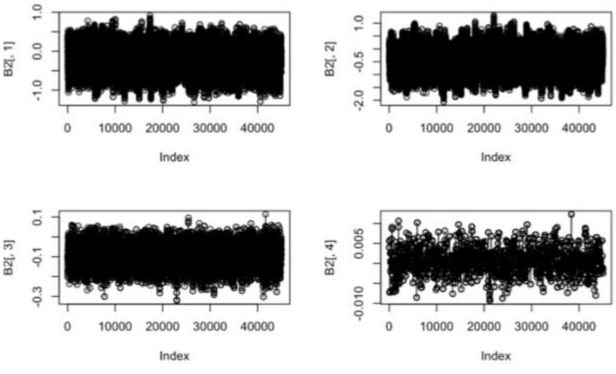


Figure 8: trace plots for pH, conductivity, urea, calcium concentration

Model diagnosis and Variable selection

In frequentist regression model, the appropriateness of the generalized linear model above can be determined through the residual plot, normal Q-Q plot, sentinel-residual plot, and sentinel-Normal Q-Q plot (Figure 6).

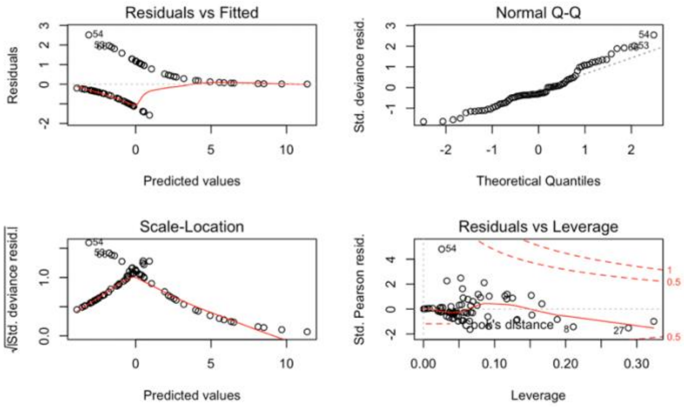
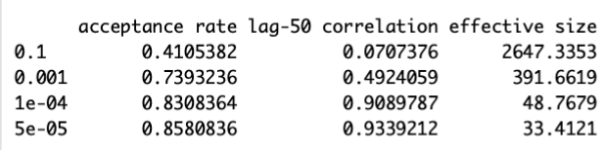


Figure 6: Residual plots and normal Q-Q plots for model glm

Based on the regression equation, the residuals plot (left upper corner in figure 6) shows the predicted value on the x axis and the residuals on the y axis. The residuals are essentially the difference between the predicted value and the actual value. Moreover, we do not want any clear patterns where the residuals either increase or decrease in line with our predicted value, or any pattern where the residuals appear non-linear (U or upside-down U shape). Coming to the normal Q-Q plot (right upper side in figure 6), this is used to assess whether our residuals are normally distributed; from this plot, we can see that the residuals deviated from the diagonal line in the upper tail; in this case we see that the tails are observed to be “heavier” (larger values) than what we would expect. This can be implied that the points form a steeper line than the diagonal. In the residuals vs leverage plot (right lower side in figure 6), we see that there are several points that have high residual and high leverage. The plots that lie close to or outside of the dashed red curves are worth investigating further. In the scale-location plot (left lower side of figure 6), we can see that our residuals are not spread equally along the ranges of predictors; this also implies that they are not equal variance. Based on the normal Q-Q plot, we can conclude that the explanatory variables gravity, conductivity, urea, and calcium concentration are best for predicting the response, r, under the frequentist framework. However, based on the two plots in the left side of figure 6, they show us that the glm model does not work well enough with our data. Thus, next, we switch to the Bayes framework.



```
      acceptance rate lag-50 correlation effective size
0.1         0.4105382          0.0707376      2647.3353
0.001       0.7393236          0.4924059       391.6619
1e-04       0.8308364          0.9089787        48.7679
5e-05       0.8580836          0.9339212        33.4121
```

Figure 9: Report for average acceptance rate, lag-50 correlation, and effective number of samples.

In Bayes framework, the function logisticRegressionBayes implements a Metropolis algorithm. With candidates generated from normal distribution with mean equal to the current sample and variance V; and we set v value for 0.1, 0.001, 0.0001, 0.00005 (4 predictors, thus 4 values). See the result in figure 9 which reports the acceptance rate, the lag-50 correlation and effective number of samples. V=0.1 has the largest effective size among samples and the smallest lag-50 correlation which implies that it provides more information than the others. Thus, in our prediction, small values of lambda lead to high rates of acceptance but high correlation between samples.

Conclusion and Discussion

In this project, we determine the relatively better result on predicting our dataset based on the scoring rule we set: if the prediction is 0.5 or greater that there are calcium oxalate crystals present and response is 1, we add 1 point for the corresponding model's scores (variables that are named "fm0" and "bay0" in the R file). Similarly, we add one point if the prediction is less than 0.50 and the response is 0. Note that we use bayesglm function for predicting the model here, because R's built in functions for doing this do not work with the logisticRegressionBayes function we were provided in class. After running through the total 77 samples in our dataset, we have the scores for frequentist framework of 65, and Bayesian framework of 44. These yield the approximated accuracy of 84% in frequentist, and 57% in Bayes. Note here that the comparison of accuracies for both inferences is calculated on the identical 4 predictors (gravity, conductivity, urea, and calcium concentration). Surprisingly, we would say the glm model might perform better under this scoring rule. However, there is lots of room to adjust Bayes to better fit the data. For example, the Bayesian model might perform better by considering a different prior distribution and increasing the sample sizes. Even though the frequentist inference performs better under our scoring rule, there are still some strong reasons to work with Bayes based on its advantages and characteristics. Overall, the advantage that Frequentist framework has is its rejection on using any prior knowledge: In throwing away all a prior information, they do worse when the prior information was useful, and they do better when the prior information was systematically biased. In Bayes, we make a decision based on the posterior probability and the parameters of interest is considered a random variable while in the frequentist inference they are considered fixed.