Analysis of Calcium Oxalate Crystals in Urine Using Logistic Models

Yuma Mizushima

I. Introduction

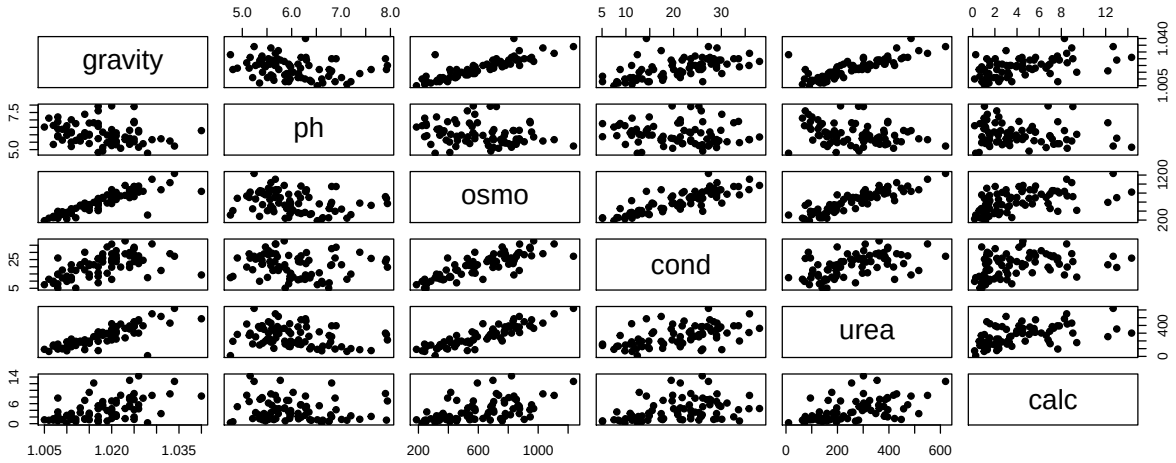
Calcium oxalate crystals are the most common type of kidney stones. When there are high levels of calcium, oxalate, cystine, or phosphate and a lack of liquids, solid masses form in the kidney. Dehydration due to lack of drinking enough liquids and a diet high in protein, oxalate, salt, and sugar are some of the risk factors that may lead to development of these solid masses.

79 samples of urine were analyzed in order to determine whether certain physical attributes of urine may be related to the formation of calcium oxalate crystals. The "urine" dataset in the "boot" package contains 7 variables (r, gravity, ph, osmo, cond, urea, and calc). The "r" variable is our response, with binary outcome - 1 denoting presence of calcium oxalate crystals and 0 denoting no presence. The data has been cleaned first from observations with missing values, resulting in 77 samples of usable data. All of the explanatory variables are continuous variables and explained in detail below.

gravity - The specific gravity of the urine.
ph - The pH reading of the urine.
osmo - The osmolarity of the urine. Osmolarity is proportional to the concentration of molecules in solution.
cond - The conductivity of the urine. Conductivity is proportional to the concentration of charged ions in solution.
urea - The urea concentration in millimoles per litre.
calc - The calcium concentration in millimoles per litre.

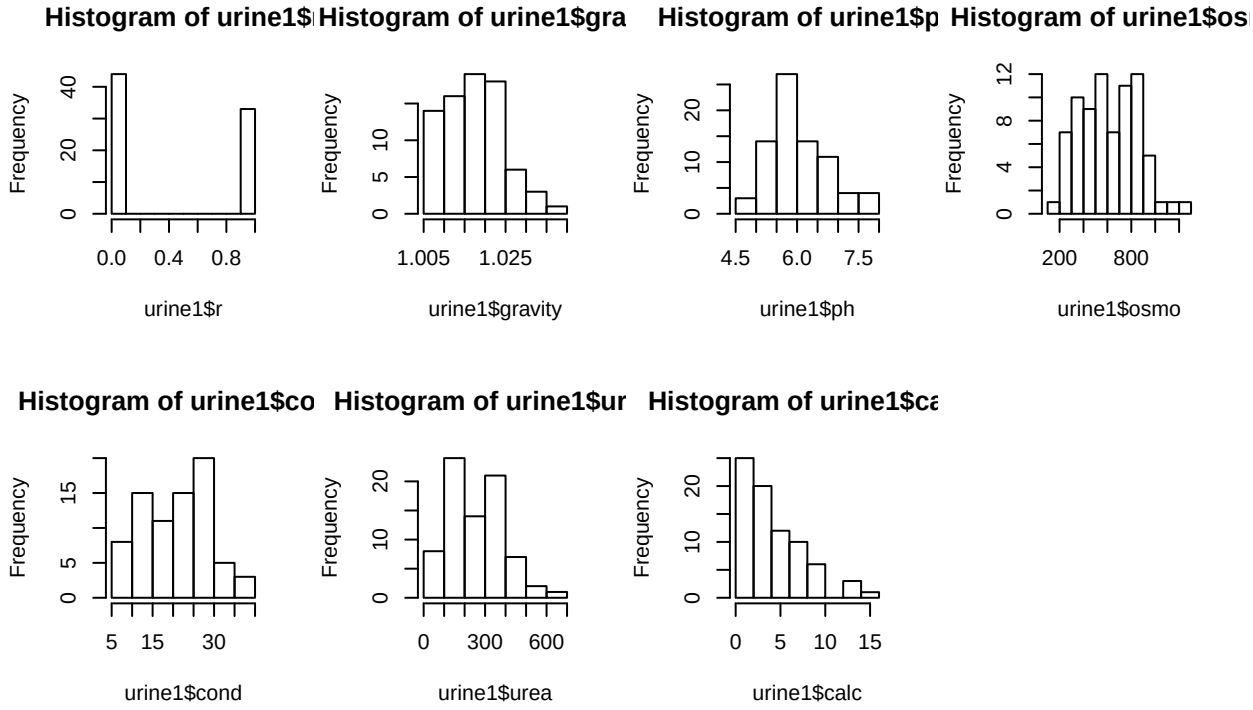
We will analyze this dataset to fit appropriate models for an indicator of the presence of calcium oxalate crystals.

II. Exploratory Data Analysis



We see from the scatterplot matrix of our predictors that several pairs have linear, positive correlation.

One example is that gravity seems to be highly correlated with osmo and urea. This is an indication of multicollinearity, which may pose problems in acquiring accurate estimates of the coefficients of each variable. The 4 pairs with the most correlation are: "osmo" vs "urea", "gravity" vs "osmo", "gravity" vs "urea", and "osmo" vs "cond", in decreasing order. "Osmo" vs "urea" has a Pearson correlation coefficient of 0.89, which is extremely high and will be handled later.
The distribution of all variables are shown below:



III. Methods

Frequentist Regression Model

Our logistics regression model is as follows:

$$Y_i = \begin{cases} 1, & \text{presence of calcium oxalate crystals} \\ 0, & \text{no presence of calcium oxalate crystals} \end{cases}$$

$$\text{logit}(\theta_i) = \log(\frac{\theta_i}{1-\theta_i}) = \tau + \beta_1 X_1 + ... + \beta_p X_p$$

$$E(Y_i) = \theta_i = \frac{e^{\tau+\beta_1 X_1+...+\beta_p X_p}}{1+e^{\tau+\beta_1 X_1+...+\beta_p X_p}}$$

The glm() function allows us to fit generalized linear models, which includes logistic regression as one of the models. Using the 'family = binomial' argument, we perform a logistic regression rather than a different generalized linear model.

```
##                  Estimate   Std. Error    z value    Pr(>|z|)
## (Intercept) -355.33770709 222.76696455 -1.5951095 0.110687746
## gravity      355.94379390 222.11004140  1.6025561 0.109032700
```

```
## ph           -0.49570188   0.56975565 -0.8700254 0.384286539
## osmo          0.01681128   0.01781576  0.9436182 0.345364804
## cond         -0.43281891   0.25123314 -1.7227779 0.084928692
## urea         -0.03201315   0.01611884 -1.9860707 0.047025469
## calc          0.78369128   0.24216382  3.2362030 0.001211312
```

"Urea" and "calc" have a significant effect, at the 0.05 level, on "r" - the presence of calcium oxalate crystals. "Calc" is by far our most significant variable, per the summary results. This is expected, as we know high levels of calcium result in higher probability of formation of calcium oxalate crystals. Both "ph" and "osmo" have high p-values, indicating minimal significance on our response variable "r".

Fitting a Reduced Model

Because too many covariates in a multivariate model may cause problems of overfitting, we will look to reduce the number of covariates. As previously mentioned, the "osmo" vs "urea" pair had the highest correlation, so we will first remove the least significant variable (i.e "osmo" predictor) out of the 2.

```
##                  Estimate   Std. Error    z value    Pr(>|z|)
## (Intercept) -504.94023332 1.613051e+02 -3.1303423 0.001746027
## gravity      504.70610351 1.611253e+02  3.1323832 0.001733934
## ph            -0.41584195 5.555369e-01 -0.7485406 0.454134130
## cond          -0.20922197 7.243907e-02 -2.8882476 0.003873948
## urea          -0.01869147 7.296142e-03 -2.5618292 0.010412252
## calc           0.72611893 2.229300e-01  3.2571618 0.001125323
```

The "ph" predictor has a high p-value, so we will remove it and regress "r" on the remaining predictors.

```
##                  Estimate   Std. Error   z value    Pr(>|z|)
## (Intercept) -500.01089750 1.618710e+02 -3.088948 0.002008668
## gravity      497.12038416 1.613294e+02  3.081400 0.002060297
## cond          -0.20546936 7.104578e-02 -2.892070 0.003827129
## urea          -0.01782852 7.230404e-03 -2.465771 0.013671887
## calc           0.72231675 2.199721e-01  3.283675 0.001024630
```

Upon removal of both "ph" and "osmo", all other predictors - gravity, cond, urea, and calc - have a significant effect on "r" at the 0.05 level.

The resulting regression model is as follows:

$$\text{logit}(\theta_i) = -500.01 + 497.12 * X_1 - 0.205 * X_2 - 0.0178 * X_3 + 0.722 * X4,$$

where X_1 indicates gravity, X_2 indicates cond, X_3 indicates urea, and X_4 indicates calc

Generalized Linear Model in the Bayesian Framework

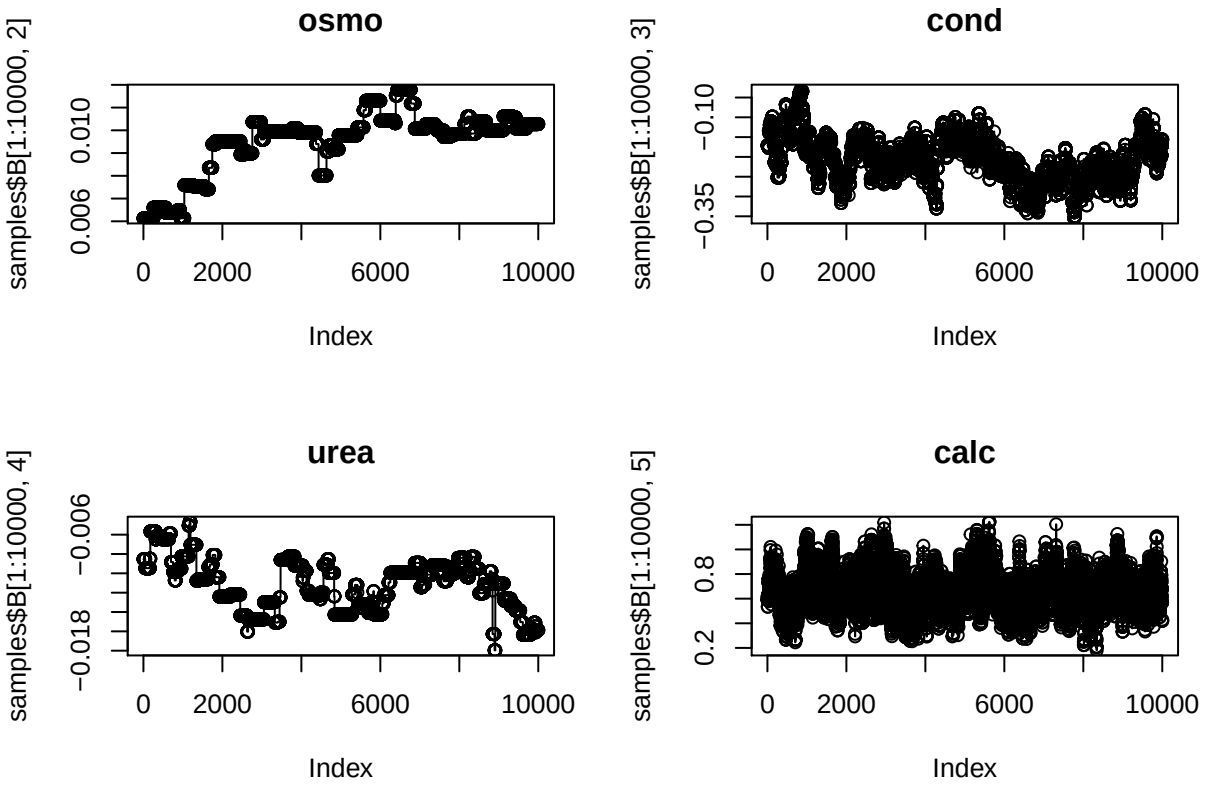
As the frequentist method may be outperformed by Bayesian methods, we will attempt to fit a generalized linear model in the Bayesian framework. Because the closed form for a posterior distribution is unknown, we will use the Metropolis sampler logisticRegressionBayes() function to fit the logistic regression. This function generates samples from the posterior distribution, which will be derived with a flat normal prior with variance 0.02 and binomial likelihood function, using a Metropolis algorithm. Initially, we used the same predictors from the frequentist model, but later removed "gravity" as its trace plot did not converge. Instead, we replaced "gravity" with "osmo", which had the highest correlation with "gravity". We will first

collect 20,000 samples and discard the first 10,000 iterations of the MCMC chain for burn-in so that it converges.

Each regression coefficients' posterior mean, posterior standard deviation, and 95% posterior credibility regions are below:

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                  Mean       SD  Naive SE Time-series SE
## (Intercept) -1.426124 0.864658 8.647e-03      0.1380617
## osmo         0.009502 0.001383 1.383e-05      0.0006962
## cond        -0.208431 0.055532 5.553e-04      0.0088627
## urea        -0.011172 0.002603 2.603e-05      0.0007844
## calc         0.653466 0.161145 1.611e-03      0.0123635
##
## 2. Quantiles for each variable:
##
##                  2.5%       25%       50%       75%     97.5%
## (Intercept) -3.255993 -2.036240 -1.391868 -0.824878  0.191032
## osmo         0.006132  0.009334  0.009912  0.010274  0.011781
## cond        -0.311696 -0.248020 -0.210357 -0.171488 -0.094900
## urea        -0.016283 -0.013213 -0.010809 -0.009197 -0.006159
## calc         0.359149  0.539247  0.644828  0.757547  0.999458
```

From the summary, we can say that the probability that the posterior parameters are between the 2 percentages (in the 2.5% and 97.5% columns) is 95%.



We can see each coefficient's trace plots above. The "cond" and "calc" trace plots seemed to converge fairly well, while "osmo" and "urea" did not.

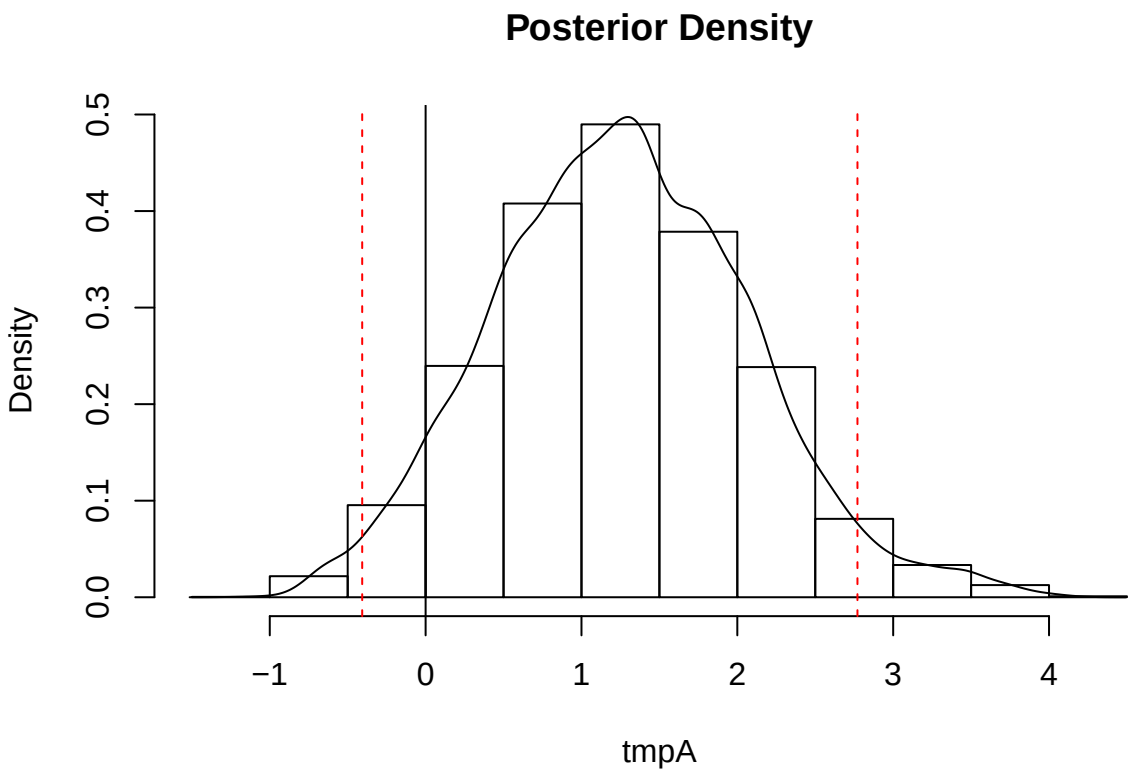
```
## (Intercept)        osmo        cond        urea        calc
##   39.223129    3.944692   39.260279   11.015128  169.883126
```

The effective sample size for each covariate can be seen above. We can interpret effective sample size as the estimate of number of independent MC samples necessary to give the same precision as the MCMC sample.

```
## (Intercept)        osmo        cond        urea        calc
##  0.02258104  0.07120467  0.02257035  0.04261085  0.01085025
```

The MC standard errors are shown above. Compared to the frequentist regression model, the Bayesian model has lower standard error for the "cond" and "calc" predictors.

We can use the samples collected to estimate the posterior distribution of the probability of calcium oxalate crystals present. We will create a histogram and density plot of the posterior density, with mean values substituted for each predictor used.



We can see that the plot is approximately normally distributed. The red, vertical lines indicate the 95% posterior credibility region.

IV. Conclusion

In the logistic regression model, we maximized the likelihood function to produce the maximum likelihood estimators using the glm() function. The coefficients could then be substituted into the model. In the Bayesian analysis, we considered a prior. In this case, the prior is weak, so the prior distribution was wide. Therefore, the likelihood was much more influential in the posterior distribution.

One of the advantages of the Bayesian perspective is that it allows us to make credible interval statements, as opposed to the confidence interval in frequentist methods. In the Bayesian framework, we believe the interval to truly contain the true population parameter. On the other hand, a 95% confidence interval covers the true parameter value in 95% of cases.