Executive summary/abstract

In this homework, I use the weight data of chickens under different recipes to determine which recipe has the best effect on the growth of chickens, and answer some questions related to probability. I answered the above questions by constructing an ANOVA Bayesian model, and I came to the conclusion that recipe No. 3 has the strongest effect on chicken growth. For the conclusions of other probability problems, see the results section.

Introduction

Understand the problem

In the link provided at the bottom of the course, I found data on chicken growth. The data includes the weight changes of a total of 50 chickens under 4 different diets within 21 days. Through this data, we can study the impact of different recipes on chicken growth, so as to select the best recipe. At the same time, through the Bayesian model learned in the course, we can explore the probability of the influence of different recipes on chicken weight. For example, what is the probability that the chickens that use each recipe will grow better (that is, have more weight) than the chickens that use other recipes? In order to solve the above problems, we will first perform data preprocessing, data cleaning, and data exploration. The above work will be carried out in Python. After preparing the data, we will use Jags to build the Bayesian model in R Studio. According to the data, we can judge that we need to use the ANOVA Bayesian model.

Data

Plan and properly collect relevant data

The selected data contains the growth weight data of the chickens under different diets, so we can judge the influence of the diet on the growth of the chickens according to the final or maximum weight of the chickens in different groups, so as to select the best recipe and The question of the probability of chicken growth in recipes.

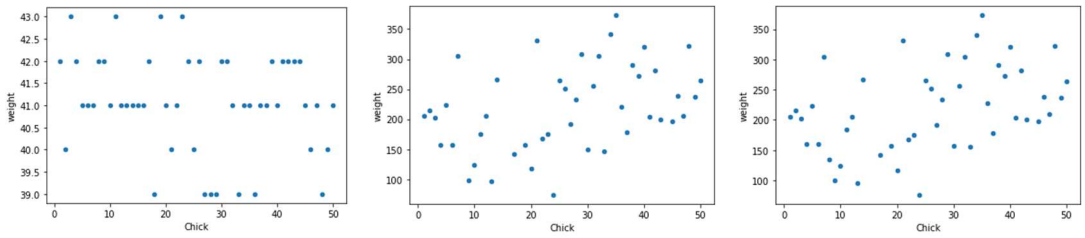
The data comes from the link below the course. The data on the page is described in detail as follows: The ChickWeight data frame has 578 rows and 4 columns from an experiment on the effect of diet on early growth of chicks. The body weights of the chicks were measured at birth and every second day thereafter until day 20. They were also measured on day 21. There were four groups on chicks on different protein diets.

After drawing the picture using the code provided on the page and manually observing the 578 rows of data, we found that most chickens have 21-day weight data, a small number of chickens only have 20-day weight data, and a few chickens have less 20-day weight data. We need to screen the chicken data, and choose to use 20 or 21 days or the final weight data in history as the evaluation index for chicken growth. At the same time, we have to consider whether to take the weight of the chicken as a reference, and only use the change in weight as the criterion for the growth of the chicken.

Explore data

First, we explore and visualize the number of days in the data to see if the final number of days for each chicken is the same. If it is different, which day should be selected as the final weight, or the maximum weight as the final weight. At the same time, we have to observe whether the birth weight of each chicken is similar. If it is similar, we can ignore the influence of the initial weight of the chicken. If they are not similar, we need to take the difference between the final body weight and the initial body weight as the criterion for judging the chicken's growth.

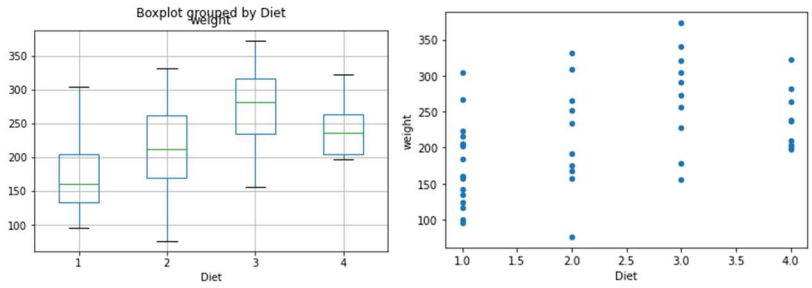
First, we explored the final days of the chickens and found that 4 chickens had final measurement days less than 20 days, 1 chicken had the end of the measurement days and 20 days, and the weight measurements of the other 45 chickens all ended in 21 days.



From left to right are the weight of the chicken on day 0, the weight of chicken on day 21, and the maximum weight of chicken

Subsequently, we made a scatter plot of the initial body weight of the chicken, and made a scatter plot of the body weight and the maximum body weight of the chicken on the 21st day. It is found that the difference between the initial body weight of the chicken and the final body weight is not much different, so we can directly use the final body weight as the standard for the growth of the chicken. At the same time, we can find through statistics that among the 44 chickens whose measurements ended on day 21, a total of 35 chickens reached the maximum weight on day 21, and the maximum weight of the other 9 chickens was not on day 21. Therefore, we can choose the maximum weight of the chicken as the standard for growth.

Finally, we made a box plot of the maximum weight of each group of chickens, and it can be seen that the maximum weight of the third group of chickens was significantly higher than that of the other groups. However, we cannot make probabilistic judgments based on this alone. We also need to fit the model to the data. At the same time, we made a scatter plot of chicken weight between different groups. Through the scatter plot, it can be considered that the weight of the chickens in the group is roughly normally distributed.



Box plots and scatter plots of chicken weights under different recipes

Model

Postulate a model

For this data, I chose to use the ANOVA Bayesian model. Because our problem is to explore whether different recipes will affect the growth of chickens, chickens using different recipes belong to different groups, so the ANOVA model is appropriate. Through this model, we can get the final average weight of different groups and their posterior probability distribution, which can determine which group has the best effect on chicken growth and answer related probability questions. As mentioned above, the obtained probability distribution mean and posterior probability distribution can answer the probability question of which recipe is the best for chicken growth and that between different groups.

For this model, we can think that the weight of chickens presents a normal distribution within the group, and there are different expectations between different groups. We can first assume that the variances between different groups are the same, and then assume that the variances between different groups are different, and then calculate the DIC values of different models to determine which model is better.

For the prior probability distribution of the model, we can choose a prior probability distribution with little information, so that the model can choose the best posterior probability distribution.

To explore the effect of within-group variance on the results, I used two models. The variances of all groups in the first model are the same, and the variances within all the groups in the second model are different. Judging by the DIC value that the first model is better, that is, the variance within the group is considered the same.

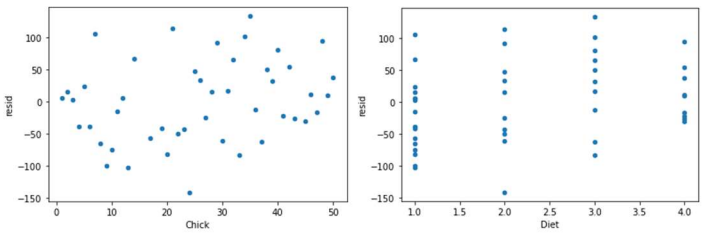
```
mod_string = " model {
for (i in 1:length(Chick)) {
  weight[i] ~ dnorm(theta[Diet[i]], 1.0 / sigma^2)
}

for (j in 1:max(Diet)) {
  theta[j] ~ dnorm(mu, 1.0 / tau^2)
}

mu ~ dnorm(0, 1.0 / 1e6)

tau_prec ~ dgamma(1.0/2.0, 3*1.0/2.0)
tau = sqrt(1.0 / tau_prec)

sigma_prec ~ dgamma(1.0/2.0, 2*1.0/2.0)
sigma = sqrt(1.0 / sigma_prec)

} "
```

```
> gelman.diag(mod_sim)
Potential scale reduction factors:

       Point est. Upper C.I.
mu           1.04       1.09
sigma        1.08       1.25
tau          1.24       1.75
theta[1]     1.23       1.66
theta[2]     1.02       1.03
theta[3]     1.32       1.88
theta[4]     1.12       1.38

Multivariate psrf

1.29
```

```
mod_string = " model {
for (i in 1:length(Chick)) {
  weight[i] ~ dnorm(theta[Diet[i]], 1.0 / (sigma[Diet[i]])^2)
}

for (j in 1:max(Diet)) {
  theta[j] ~ dnorm(mu, 1.0 / tau^2)

  sigma_prec[j] ~ dgamma(1.0/2.0, 2*1.0/2.0)
  sigma[j] = sqrt(1.0 / sigma_prec[j])
}

mu ~ dnorm(0, 1.0 / 1e6)
tau_prec ~ dgamma(1.0/2.0, 3*1.0/2.0)
tau = sqrt(1.0 / tau_prec)

} "
```

```
> dic
Mean deviance:  517.9
penalty 6.199
Penalized deviance: 524.1
> dic2
Mean deviance:  517
penalty 5.974
Penalized deviance: 523
```

From left to right are the structure of model 1, the fitting result of model 1, the structure of model 2, and the DIC value of model 1 and 2.



From left to right are the residuals of model 1 and the residual scatter plots of each group

Results & Conclusions

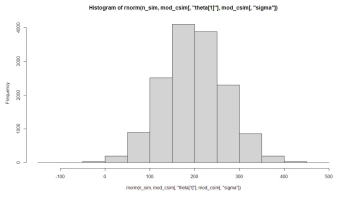
```
> summary(mod_sim)

Iterations = 2001:7000
Thinning interval = 1
Number of chains = 3
Sample size per chain = 5000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean     SD Naive SE Time-series SE
mu      220.46 18.935  0.15460         0.6124
sigma    66.73  8.046  0.06569         0.2423
tau      21.47 23.000  0.18779         1.3339
theta[1] 199.11 19.148  0.15634         1.2899
theta[2] 217.34 15.044  0.12283         0.5692
theta[3] 239.69 24.722  0.20185         1.8071
theta[4] 226.63 17.672  0.14429         0.9056

2. Quantiles for each variable:

          2.5%     25%     50%     75%  97.5%
mu      187.194 210.124 219.39 229.72 259.87
sigma    52.822  61.039  66.14  71.65  84.25
tau       1.541   3.359  16.41  31.43  76.18
theta[1] 160.317 185.381 200.87 213.69 231.66
theta[2] 187.645 207.948 217.16 226.45 248.65
theta[3] 201.573 219.741 236.17 258.16 290.56
theta[4] 197.479 214.269 224.38 237.18 266.00
```

Summary of posterior probability of model parameters

Theta1~4 represent the average chicken weight of the four recipes, sigma represents the variance of chicken weight, and mu and tau represent the mean and variance of the normal distribution

controlling theta distribution.

```
# The weight distribution of chicken recipe No. 1
hist(
  rnorm(n_sim, mod_csim[, 'theta[1]'], mod_csim[, 'sigma'])
)
```



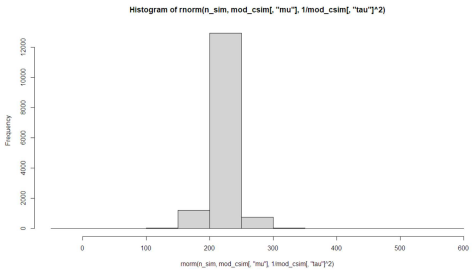

```
# The probability that the average weight of the chickens on the No. 2 recipe is higher than the average weight of the chickens on the No. 1 recipe
mean(mod_csim[, 'theta[2]'] > mod_csim[, 'theta[1]'])
```

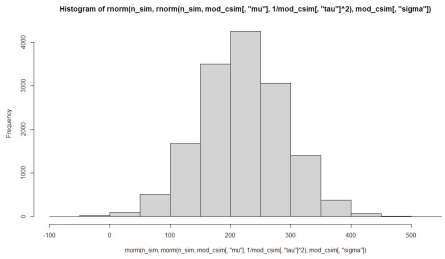
0.7944

```
# The probability that a chicken with recipe No. 3 will weigh more than a chicken with recipe No. 1
mean(rnorm(n_sim, mod_csim[, 'theta[3]'], mod_csim[, 'sigma']) > rnorm(n_sim, mod_csim[, 'theta[1]'], mod_csim[, 'sigma']))
```

0.652

```
# Adopt a new recipe, the average weight distribution of chickens and the weight distribution of a chicken
hist(
  rnorm(n_sim, mod_csim[, 'mu'], 1 / mod_csim[, 'tau']^2)
)
hist(
  rnorm(n_sim,
        rnorm(n_sim, mod_csim[, 'mu'], 1 / mod_csim[, 'tau']^2),
        mod_csim[, 'sigma']
  )
)
```





Drawback: Using the normal distribution will cause the weight of the chicken to be 0 or even a negative number, which is not consistent with common sense, and this is also the shortcoming of the model. Perhaps other distributions should be used to fit the body weight of the chicken.