Project 4 Template

Bayesian Analysis of Question 4 from the Master's Proficiency Exam

Jesse Troy, PhD

January 03, 2024

Contents

1 Background **1**

2 Question 1: Write the Bayesian Logistic Regression Model **2**
2.1 Write the data model . . . 2
2.2 Write the model for the prior . . . 2
2.2.1 Prior for Beta0 . . . 2
2.2.2 Prior for Beta1 . . . 2

3 Question 2: Simulation of the Priors **2**

4 Question 3: Update your prior belief using the OSTEO data **3**

5 Question 4: Including Age in the Model **3**
5.1 Writing the Data Model . . . 3
5.2 Prior Distributions for the Regression Parameters . . . 3
5.3 Posterior . . . 3

6 Question 5: Evaluating Interaction Between Age and a Prior History of Fracture **4**
6.1 Write the Data Model . . . 4
6.2 Prior Distributions for the Regression Parameters . . . 4
6.3 Posterior . . . 4

7 Question 6: Predicting Fracture Risk Based on Prior History and Age **4**

8 Question 7: Posterior Predictive Check **4**

9 Question 8: Identifying Women at High Risk of Fracture **4**

1 Background

The enclosed dataset called OSTEO.CSV includes the results of a prospective cohort study that followed 500 women age 55 or older for 1 year after diagnosis with osteoporosis. The primary outcome of the study was bone fracture (yes/no) within the 1st year of osteoporosis diagnosis. The investigators hypothesized that women who had a history of bone fractures prior to their osteoporosis diagnosis would be at higher risk of having an osteoporosis-related fracture than those who had no such prior history. A potentially important prognostic variable was age at diagnosis of osteoporosis, which was recorded in years. You have already done a frequentist analysis of this dataset. Now, you will analyze the same dataset using a Bayesian logistic regression model. You will complete the analysis by answering the following questions.

2 Question 1: Write the Bayesian Logistic Regression Model

Write a simple Bayesian logistic regression model for fracture at 1 year post-diagnosis with osteoporosis based only on a history of prior fracture. Follow the example notation used in your textbook in Chapter 13, equation 13.5. In addition to specifying the model in mathematical notation, give a plain English description of what the notation means. Do this separately for the data model and the priors as indicated below.

2.1 Write the data model

Write the data model here.

2.2 Write the model for the prior

Specify "vague" normal prior distributions for the parameters in your data model as follows.

2.2.1 Prior for Beta0

First, assume your prior information says that 20% of women will have fractures at 1 year when they don't have a prior history of fracture. Your level of uncertainty about this is quite high. A prior that has a 95% credible interval ranging from about 12% to 31% would capture your uncertainty well. What are the parameters of a normally distributed prior for β_0 that express this uncertainty?

2.2.2 Prior for Beta1

Assume that you have prior information that suggests a history of fractures is expected to increase the probability of fracture by 10% over not having a prior history (i.e., a 30% probability of fracture at 1 year in women who have a prior history of fracture vs. 20% probability in women without a prior history). But your prior information is highly uncertain. In fact, it seems plausible that women with a prior history of fracture may have as little as a 10% probability of fracture at 1 year (the same risk in women without a prior history!) or as high as a 62% risk of fracture at 1 year (a very large increase in risk over women without a prior history). Based on this information, specify a normally distributed prior for β_1 on the log-odds scale and write an interpretation of the mean and 95% credible interval with respect the clinical problem (i.e., whether a prior history promotes or reduces the risk of fracture at 1 year).

State the mean and 95% credible interval for the β_1 prior on the odds ratio scale and write an interpretation.

3 Question 2: Simulation of the Priors

Simulate the priors for your model. Plot the theoretical priors as well as the MCMC simulated priors and verify the simulations are correct. State the mean, median, and 95% credible intervals for your priors and write interpretations for these.

HINT: Follow the instructions in the stan_glm() help pages and use a model without an intercept. Include a term for no prior history of fracture instead of the intercept:

"If you prefer to specify a prior on the intercept without the predictors being auto-centered, then you have to omit the intercept from the formula and include a column of ones as a predictor, in which case some element of [prior] specifies the prior on it, rather than [prior_intercept]."

More information is here.

Following example 13.2.2 in your textbook, simulate 100 models that represent your prior belief about the effect of a prior history of fracture on the risk of fracture at 1 year. HINT: use add_linpred_draws() to look at the impact on the log-odds scale. Describe what you see and how it relates to the theoretical prior distributions you plotted above.

4 Question 3: Update your prior belief using the OSTEO data

Now, use the OSTEO data set to update your priors. Plot the posterior distribution for the change in log-odds of fracture at 1 year associated with a prior history of fracture. Interpret the median and 95% highest posterior density credible interval with respect to the research question about whether prior history of fracture is a risk factor for fracture at 1 year after diagnosis with osteoporosis. Repeat these steps on the odds ratio scale.

What is the posterior probability that the odds ratio is greater than 1? What does this imply about the risk of fracture at 1 in women with vs. without a prior history of fracture?

What is the posterior probability that the odds ratio is greater than 2?

What is the value of the odds ratio at which the posterior probability exceeds 80%?

Based on the above posterior probabilities, write a summary sentence that describes the strength of the evidence you now that that a prior history of fracture is a risk factor for fracture 1 year after diagnosis with osteoporosis.

5 Question 4: Including Age in the Model

5.1 Writing the Data Model

Write the data model when the interaction term is included. Use a similar format to what you used to write the data model in question 1. In your notation, use β_0, β_1, and β_2 as the intercept, prior history of fracture, and age respectively and explain what each parameter means. Remember, rstanarm gives model results without centering (even though the prior for the intercept is based on the other predictors being centered; see the next question).

5.2 Prior Distributions for the Regression Parameters

So far you have used informative but vague priors for the risk of fracture in women with and without a prior history of fracture. The priors are vague in the sense that they suggest an expected difference in the population comparing women with and without a prior history, but the level of uncertainty is very high.

Suppose now that you want to include age in the model since you already know age is a confounder in the population and your descriptive analysis demonstrates age is a confounder in your dataset. Suppose however that you don't have any pre-study information about the risk of fracture in women with and without a prior history according to age. So, instead of using vague, informative priors you will not let rstanarm select weakly informative priors scaled to match your data.

Create an MCMC simulation of weakly informative priors for a model that includes prior history of fracture and age as covariates. Write the priors that rstanarm has identified.

5.3 Posterior

Now update the priors with the data in the OSTEO dataset. Plot the posterior distribution for prior history of fracture and state the median and 95% highest posterior density credible interval on the odds ratio scale. What has happened to the posterior distribution for β_1 now that age is included in the model (as compared to the first model you ran without age)? Is this consistent with what you expected would happen? Why or why not?

6 Question 5: Evaluating Interaction Between Age and a Prior History of Fracture

6.1 Write the Data Model

Write the data model. Use a similar format to what you used to write the data model in question 1. In your notation, use β_0, β_1, and β_2 as the intercept, prior history of fracture, and age respectively and explain what each parameter means. Remember, rstanarm gives model results without centering (even though the prior for the intercept is based on the other predictors being centered; see the next question).

6.2 Prior Distributions for the Regression Parameters

Create an MCMC simulation of weakly informative priors for a model that includes prior history of fracture and age as covariates, as well as the interaction between these two factors. Write the priors that rstanarm has identified.

6.3 Posterior

Now update the priors with the data in the OSTEO dataset. Plot the posterior distribution for the interaction term and interpret the median and 95% highest posterior density credible interval. What impact does older age have on the relationship between fracture risk in those with and without prior history?

Write a single sentence that summarizes the strength of evidence you have that age interacts with prior history of fracture to increase the risk of fracture 1 year after diagnosis with osteoporosis?

7 Question 6: Predicting Fracture Risk Based on Prior History and Age

Make a plot that shows the posterior predicted probability of fracture at 1 year post-diagnosis with osteoporosis by age at diagnosis (55 to 80 years, in increments of 5 years) and prior history of fracture. HINT: Your plot should show the age at diagnosis on the X-axis and the posterior predicted probability on the Y axis. There should be two lines on the plot: one for women without a prior history of fracture and another for women who have a prior history of fracture.

8 Question 7: Posterior Predictive Check

Perform a posterior predictive check examining how consistent the model is with the original data. Use 100 datasets. Explain how this posterior predictive check works and what you conclude from the results.

9 Question 8: Identifying Women at High Risk of Fracture

Suppose that in clinical practice an intervention would be warranted in women who have a posterior predicted probability of fracture at 1 year that exceeds 40%. First, use this cutoff for predicting a woman is at high risk to evaluate the model based on the data you have. State the sensitivity, specificity, overall accuracy, and positive predictive value and interpret each of these in light of the clinical problem of identifying women eligible for an intervention to prevent fracture within 1 year of diagnosis with osteoporosis. HINT: rstanarm doesn't give you the positive predictive value by default but you can use Bayes' theorem to find positive predictive value based on sensitivity, specificity and prevalence.

Then, use cross validation to evaluate the future performance of the model. Again, state the sensitivity, specificity, overall accuracy, and positive predictive value and interpret each of these.