

hw08

1 BR 13.7

Plans change. Hotel room bookings get canceled. In the next exercises, you'll explore whether hotel cancellations might be predicted based upon the circumstances of a reservation. Throughout, utilize weakly informative priors and the `hotel_bookings` data in the `{bayesrules}` package. Your task is to predict `is_cancelled` (Y , whether or not the booking was cancelled) based on the following predictors:

- `lead_time`: number of days between the booking and scheduled arrival (X_1)
 - `previous_cancellations`: number of previous times the guest has cancelled a booking (X_2)
 - `is_repeated_guest`: whether or not the booking guest is a repeat customer at the hotel (X_3)
 - `average_daily_rate`: the average per day cost of the hotel (X_4)
- (a) Using formal mathematical notation, specify an appropriate Bayesian regression model of Y by predictors X_1, X_2, X_3, X_4
 - (b) Simulate the posterior model of the regression parameters $(\beta_0, \beta_1, \dots, \beta_4)$.
 - (c) Construct trace plots, density plots, and a `pp_check()` of the chain output.
 - (d) Report the posterior median model of hotel cancellations on each of the log(odds), odds, and probability scales.
 - (e) Construct 80% posterior credible intervals for your model coefficients. Interpret those for β_2 and β_3 on the odds scale.
 - (f) Among the four predictors, which are significantly associated with hotel cancellations, both statistically and meaningfully? Explain.

2 BR 13.9

A guest that is new to a hotel and has only canceled a booking 1 time before, has booked a \$100 per day hotel room 30 days in advance.

- (a) Simulate, plot, and discuss the posterior predictive model of Y , whether or not the guest will cancel this booking.
- (b) Come up with the features of another fictitious booking that's more likely to be canceled than the booking in part a. Support your claim by simulating, plotting, and comparing this booking's posterior predictive model of Y to that in part a.

3 BR 13.8

- (a) How good is your model at anticipating whether a hotel booking will be canceled? Evaluate the classification accuracy using both the in-sample and cross-validation approaches, along with a 0.5 probability cut-off.
- (b) Are the cross-validated and in-sample assessments of classification accuracy similar? Explain why this makes sense in the context of this analysis.
- (c) Interpret the cross-validated overall accuracy, sensitivity, and specificity measures in the context of this analysis.
- (d) Thinking like a hotel booking agent, you'd like to increase the sensitivity of your classifications to 0.75. Identify a probability cut-off that you could use to achieve this level while maintaining the highest possible specificity.

4 Social conformity model

The code chunk below fits the strategy model that we discussed in class on Wednesday. Start by running the code and make sure you understand each piece. Then, answer the following questions.

- (a) How sensitive are the results to the choice of prior? Fit the model using (i) a very strong Dirichlet prior and (ii) a weaker Dirichlet prior to justify your answer
- (b) How sensitive are the results to the strategy options defined? Fit the model using a different set of strategies (you should include at least 1 new strategy, but can remove as many as you wish) and compare to the results from class to justify your answer.

Note: This model takes a long time to run! Feel free to run your models locally, and then save any images you want to include. Then add `#l eval: false` to your chunk when rendering and include your images separately. This will greatly speed up the time it takes to render your final PDF file. I've included an example here.

```
library(tidyverse)
library(rstan)
library(bayesplot)
```

```
Boxes <- read_delim("https://raw.githubusercontent.com/rmcelreath/rethinking/refs/heads/master/
```

```

Boxes_model <- "data{
  int N;
  int y[N];
  int majority_first[N];
}
parameters{
  simplex[5] p;
}
model{
  vector[5] phi;

  // prior
  p ~ dirichlet( rep_vector(4,5) );

  // probability of data
  for ( i in 1:N ) {
    if ( y[i]==2 ) phi[1]=1; else phi[1]=0; // majority
    if ( y[i]==3 ) phi[2]=1; else phi[2]=0; // minority
    if ( y[i]==1 ) phi[3]=1; else phi[3]=0; // maverick
    phi[4]=1.0/3.0; // random
    if ( majority_first[i]==1 ) // follow first
      if ( y[i]==2 ) phi[5]=1; else phi[5]=0;
    else
      if ( y[i]==3 ) phi[5]=1; else phi[5]=0;

    // compute log( p_s * Pr(y_ils )
    for ( j in 1:5 ) phi[j] = log(p[j]) + log(phi[j]);
    // compute average log-probability of y_i
    target += log_sum_exp( phi );
  }
}"

data_list = list(
  N = nrow(Boxes),
  y = Boxes$y,
  majority_first = Boxes$majority_first
)

model_fit <- stan(model_code = Boxes_model,
  data = data_list,
  chains = 4,
  iter = 2*5000,
  seed = 110525)

```

```
as.data.frame(model_fit) |>
  mcmc_areas(regex_pars = "p\\\[", prob = .8) +
  theme_minimal()
```

