GR5065 Homework 5

Ben Goodrich

Due April 11, 2022 at 4PM

```
# call the set.seed function once here to make the knitting conditionally deterministic
# but remember to also pass a seed to any Stan-related function that draws from the posterior
```

1 Employment

For this problem, you are going to need to download some Current Population Survey (CPS) data. The easiest way to do that is to go to

http://ceprdata.org/cps-uniform-data-extracts/cps-outgoing-rotation-group/cps-org-data/

and download the compressed file for the year that you were born to the same directory as your RMarkdown file. Then, unzip that file in the same directory to produce a Stata formatted file, which will have a dta file extension. The Stata formatted file can be loaded into R properly with something like

```
library(haven)
CPS <- as_factor(read_dta(dir(pattern = "^cepr_.*dta$")))
# as_factor changes the categorical variables in Stata to R factors
```

Finally, filter the `CPS` data.frame down to the month that you were born using the `month` variable so that it is not too big.

A brief description of the variables in `CPS` and the values they take (if categorical) can be obtained by

```
defs <- sapply(CPS, FUN = attr, which = "label")
vals <- sapply(CPS, FUN = attr, which = "levels")
```

Additional documentation of these variables can be found at http://ceprdata.org/cps-uniform-data-extracts/cps-basic-programs/cps-basic-documentation/ or the from the links on that page but note that `CPS` does not include the household-level variables and recodes / combines / renames some of the individual-level variables. If you are familiar with Stata, it might be helpful to look at the dofiles that create the dataset which can be found at http://ceprdata.org/cps-uniform-data-extracts/cps-outgoing-rotation-group/cps-org-programs/ .

1.1 Selection

One of the criticisms labor microeconometricians make of models of whether an individual is (un)employed is that they take whether the person is in the labor force as an exogenous known instead of attempting to develop a model that explains both why some people are (not) in the labor force and why some people in the labor force are (not) employed. Treating whether someone is in the labor force as an exogenous known is harmful if any of the unmeasured variables that affect whether someone is in the labor force also affect the probability that such people could find jobs if they sought them.

Draw $S = 1000$ times from the prior predictive distribution of a model that avoids this criticism. This is much easier to do well with centered predictors that you can choose among those in the CPS dataset. For each iteration s up to S, you should

1. Draw a realization of the intercept and coefficients from their respective prior distributions and use those realizations to form a linear predictor of size N (the number of people in the CPS that month) for being in the labor force
2. Draw a realization of the intercept and coefficients from their respective prior distributions and use those realizations to form a linear predictor of size N (the number of people in the CPS that month) for having a job if that person *were* in the labor force, irrespective of whether they actually *are* in the labor force.
3. Draw a correlation for the errors that will get added to these two linear predictors below
4. For each person n up to N, draw errors from a bivariate standard normal distribution (i.e. both expectations are 0 and both standard deviations are 1) with correlation $\tilde{\rho}$ that was drawn in the previous step. See HW2 if you have forgotten how to do this.
5. Add the first error term to the first linear predictor for each person n. If this sum is negative, then person n is simulated to be not in the labor force, which is coded as -1. Conversely, if this sum is positive, then person n is simulated to be in the labor force, in which case you next need to consider the sum of the second linear predictor and the second error term. If the latter sum is negative, then person n is simulated to be unemployed, which is coded as 0, and otherwise person n is simulated to have a job, which is coded as 1. You are going to need `if` and `else` clauses to handle this.

The end result should be a matrix with S rows and N columns, where each cell is either -1, 0, or 1.

1.2 Prior Predictive Checking

Check that your prior beliefs about the data are reasonable in various respects. For example, about two-thirds of adults should be in the labor force historically and among those, about five percent should be unemployed. In addition, every individual should have a decent chance of being in the labor force but the probability should not be to close to 100% for anyone.

1.3 Probit Model

Use rstanarm to estimate a model for the labor force participation part of the above process, where `1 - nilf` is the outcome variable, using `family = binomial(link = "probit")`.

1.4 Interpretation

For which coefficients in the previous model are you now reasonably certain about their sign?

2 Auto Insurance

Read this paper about the recent use of "telematic" apps that auto insurance companies incentivize drivers to use in order to collect data on how good or bad they are driving in order to determine how high their auto insurance premiums are. Such datasets are never publicly available, so the authors describe how they created a synthetic dataset that mimics the characteristics of a dataset that was given to them privately by a Canadian auto insurance company.

You can load the authors' synthetic dataset via

```
UBI <- readr::read_csv("http://www2.math.uconn.edu/~valdez/telematics_syn-032021.csv")
```

Although splitting the data into training and testing is not necessary or recommended for a Bayesian analysis, in this case there are 100,000 observations, so for computational reasons, we might want to split the data like

```
training <- UBI[1:50000, ]
testing  <- UBI[-(1:50000), ]
```

2.1 SMOTE

In section 2.1, the paper uses an algorithm called Synthetic Minority Oversampling Technique (SMOTE) to create a synthetic dataset that has the same proportions of a count (or categorical) outcome as the original genuine dataset. Essentially, SMOTE is like K nearest neighbors where a $K + 1$th neighbor is drawn from a conditional probability distribution. How does this process differ from what we have been doing when we draw from a prior predictive distribution and see if the result looks reasonable?

2.2 Bayesian Optimization

In section 2.2, the paper uses something called Bayesian Optimization

https://en.wikipedia.org/wiki/Bayesian_optimization

to estimate the parameters of a neural network. Based on the discussion in the paper and in the Wikipedia article, in what ways is Bayesian optimization Bayesian and in what ways is it not Bayesian? In what ways is Bayesian optimization optimal and in what ways is it not optimal?

2.3 Count Models

Use rstanarm with a Poisson and a negative binomial likelihood in the training data for the outcome called `NB_claim`, which is the number of auto insurance claims that the person had during this time period. You should be using `offset = log(Duration)` in order to model the rate of claims per day.

Note that some of the predictors that you might want to use are in percentages and sum to 100%. While this does not prevent using them to draw from the posterior distribution, you are usually better off excluding one of the components and formulating your priors on the coefficients relative to the omitted predictor. You may want to use fewer than the default number of iterations (2000) and chains (4) in order to keep the subsequent computations tractable on old laptops by specifying the optional `iter` or `chains` arguments to a somewhat lower value.

2.4 Posterior Predictive Checking of Count Models

Do either the Poisson or the negative binomial model indicate a problem with zero-inflation, which is defined as a model that predicts substantially fewer zeros than what is exhibited in the (training) data?

2.5 PSISLOOCV

Use `loo` to estimate the expected log predictive density (ELPD) for $50,000$ future observations. Which of the Poisson and the negative binomial model is expected to predict future data better? Which, if any, observations are inconsistent with the assumption that omitting that observation would have a small effect on the posterior distribution?

2.6 ELPD

Use the `log_lik` function in rstanarm with `newdata = testing` to compute a matrix with rows equal to the number or draws and columns equal to the number of testing observations for each of the Poisson and negative binomial models. Is the average of the rowsum of each of these matrices substantially different from the ELPD values that you calculated in the previous subproblem?

2.7 Claim Models

The authors also refer to gamma regression model for the average dollar amount of insurance claims among people who make at least one insurance claim. The default link function in R for a Gamma log-likelihood

```
Gamma()
```

```
##
## Family: Gamma
## Link function: inverse
```

is $\frac{1}{\mu_n} = \eta_n = \alpha + \sum_{k=1}^{K} x_{nk}\beta_k$, which is not a choice that anyone should ever make, particularly when using Bayesian methods because if η_n is negative then μ_n will also be negative, which is impossible for an outcome variable that is restricted to be positive. In addition, it is virtually impossible to think about reasonable priors for α and each β_k under this multiplicative inverse link function. A far better, albeit not default, choice would be `family = Gamma(link = "log")`, in which case $\ln \mu_n = \eta_n$ like in a count model, in which case small coefficients can be interpreted as percentage changes.

Thus, the expected value of a claim is modeled as a function of predictors and the shape parameter of the Gamma distribution is treated as an unknown scalar to be estimated with a default prior that is exponential with a rate of 1.

First, draw from the prior distribution without conditioning on the (training) data by specifying `prior_PD = TRUE`. Does the resulting predictive distribution seem reasonable in light of the fact that cars are usually valued in tens of thousands of dollars, but luxury cars can be very expensive and occasionally one car accident can involve more than two cars?

Second, condition on the subset of the training data where `NB_claim > 0` and model `AMT_claim / NB_claim` with `prior_PD = FALSE`. How would you interpret the posterior distribution of your beliefs about the effect of one telematic predictor?

2.8 Premiums

In an ideal insurance market, the policy holder should be willing to pay a premium that is equal to the expected value of their insurance claims or perhaps a little higher if the policy holder is risk adverse, in which case the risk-neutral insurance company would expect to make a profit. Of course, there are no perfect insurance markets for a variety of reasons. But why would the posterior predictive distribution implied by the previous submodel *not* be appropriate for setting premiums on drivers in the training data?