GR5065 Assignment 3

Due Thursday March 1, 2018 by 4:00 PM

Wages

Download the file called dataset.rds from Canvas to your working directory, which contains so called “person record” data on individuals who have worked during 2016 from a randomly selected state. There is documentation Specifically, the data come from the 2016 ACS 1-year PUMS subjects. No “housing record” variables are included, some variables (such as those starting with “F” or “PWGTP”) have been deleted, and the location and income-related variables have been filled with `NA`s. You can load it with

```
dataset <- readRDS("dataset.rds")
```

Prior Predictive Distribution

Write a Stan function that draws *once* from the prior predictive distribution of wages / salary (WAGP) under a generative model that is linear in its parameters, although you can apply any transformation to any variable, include any interaction or polynomial, etc. Your function may have to input several `vector`s as exogenous knowns from the columns of `dataset` and should return a `vector` of the same size. You can use any distributions that you want for the priors and conditional distribution of the outcome.

Expose your Stan function to R and call it a few times to demonstrate that the distribution of predictions is reasonable overall.

R^2

Now suppose you were going to estimate a model using the same predictors but with `stan_lm`. What would you choose to characterize your beliefs about the R^2 under a beta distribution with first shape parameter equal to $\frac{K+1}{2}$ and second shape parameter free? By default, the $R2$ function inputs a prior *mode* (if $K > 1$) but you could also specify its second argument as `what = "median"` or `what = "mean"` if that makes things easier for you.

Median Wage

How would you go about describing your beliefs about the median wage / salary (among those with a job) in this state under your model after seeing the data? How would you go about describing your beliefs about the difference in median wage / salary (among those with a job) between men and women under your model after seeing the data?

You cannot estimate the model because you do not have the outcome but you should write the code to answer the above questions as if you had estimated the model.

Twitter

Download (once) the tweets.csv and users.csv files from the bottom of this NBC website to your working directory. You can load and merge them into R with

```
library(readr)
tweets <- read_csv("tweets.csv")
tweets$created_at <- NULL
```

```r
tweets$retweeted <- NULL
tweets$posted <- NULL
tweets <- tweets[!is.na(tweets$retweet_count), ]

users <- read_csv("users.csv")
colnames(users)[1] <- "user_id"
russia <- merge(tweets, users, by = "user_id")
```

Is a Tweet Retweeted?

Use `stan_glm` to draw from the posterior distribution of a good model where the outcome is `TRUE` if the tweet is ever retweeted and `FALSE` if the tweet is never retweeted; in other words `retweet_count > 0`. You may need to create additional variables from other columns in the data, such as

```r
russia$Clinton <- grepl("Clinton", russia$text, ignore.case = TRUE)
```

which returns a logical vector indicating whether Hilary (or Bill or Chelsea) Clinton was referenced in the text of the tweet. Post on Piazza if there is a variable you want to create but do not know how to do so.

For at least one predictor, how do you believe the probability of retweeting changes as the predictor changes, after conditioning on the data?

Sidenote

When writing a RMarkdown file that uses Stan, it is best to proceed like this:

First, load the relevant package(s) and tell it to use all available cores: `library(rstanarm); options(mc.cores = parallel::detectCores())`

Then, make a chunk that has the following metadata: `{r, TAG, cache = TRUE, results = "hide"}` This says to store (on your disk) but not show (in the HTML / PDF) the output of the commands in the chunk, which would ordinarily involve drawing from a posterior distribution. You should change `TAG` to some descriptive tag for the chunk, which has to be unique among tagged chunks in a RMarkdown file. If you do not change the body of this chunk, RStudio will read the results off the disk when knitting, rather than re-estimating the model. If you do change the body of this chunk, RStudio will re-estimate it.

Finally, make a regular R chunk like `{r}` and put things like `print(post)`, etc. inside it, which will be rendered in the HTML / PDF.

How Frequently Is a Tweet Retweeted?

Use `stan_glm.nb` to draw from the posterior distribution of a good model where the outcome is the number of times that a tweet is retweeted, given that it is retweeted at least once. Thus, you need to specify `subset = retweet_count > 0` when you call `stan_glm.nb`. Technically, this is not the correct way to estimate a "zero-inflated" count model since the two model components — is a tweet retweeted and if so, how many times — should be combined into a single likelihood function but **rstanarm** currently does not support that. Based on the results, is there evidence of overdispersion in the outcome (relative to that in a Poisson model)? How do you know?