Bayesian Inference Assignment 1

Due Monday April 15 by 8:45 AM

1 Poisson Processes with Outliers

Suppose that you are about to collect N observations on a count variable. Of those, N_1 are generated by a Poisson distribution with expectation μ_1 and N_2 are generated by a Poisson distribution with expectation $\mu_2 \neq \mu_1$, such that $N = N_1 + N_2$. However, you are assuming all N observations are generated according to the *same* Poisson distribution with expectation μ, in which case you could think of the N_2 observations as unrecognized outliers (with a particular form).

For this problem, you should show your work but you can refer to lemmas we used in Week 1 without having to prove everything from scratch. Also, you can use LaTeX but you can just write the math as text or write it out with a pen and include a cell-phone picture of it in the knitted RMarkdown file.

1.1 Sampling Distribution of the Sample Mean

What is the asymptotic distribution — i.e. as $N_1 \uparrow \infty$ and $N_2 \uparrow \infty$ at the same rate — of the sample mean (before you collect the data)?

1.2 Power of the Test

A consistent test is one whose power to reject a false null hypothesis approaches 1 asymptotically. Suppose you assume a value for μ to use for a null hypothesis and it is the case that $\mu = \mu_1$. Under the true data generating process outlined above with $\mu_1 \neq \mu_2$ is the test consistent?

1.3 Posterior Distribution

Now suppose your hypothesized value for μ in the previous subproblem is given by $\frac{a}{b}$ where a is the shape parameter of a Gamma prior distribution and b is the rate parameter of a Gamma prior distribution. What is the posterior distribution conditional on the N observations that are generating according to the process outlined above, but you assume it to be a homogenous Poisson data-generating process with a single expectation?

1.4 Summary

Based on your answers to the previous subproblems, what would you conclude about the effect of (these) outliers to Frequentist and Bayesian methods in this particular case?

2 Current Population Survey

2.1 Frequentist Perspective

If one wanted to make the case for applying conventional Frequentist techniques to the unemployment rate, it might go something like this:

- Each month, the Bureau of Labor Statistics conducts the Current Population Survey (CPS) to estimate the unemployment rate, which is defined as the ratio of people who are unemployed divided by the number of people in the labor force. In turn, the number of people in the labor force is defined as the number of people who have jobs plus the number of people who do not have jobs but have actively looked for a job in the past few weeks.
- The CPS is *not* a simple random sample from the population of adults who live in the United States, but with complicated weighting schemes it is possible to make the samples (collectively) representative of the population.
- If the unemployment rate in the population (U_t) in month t is at (U^*) — the non-accelerating inflation rate of unemployment (NAIRU) — then the distribution of the estimated unemployment rate in each month's CPS is binomial with probability (U^*) and size equal to the number of people in the labor force in the CPS.
- Thus, in any particular month, one could perform a test of the null hypothesis that $U_t = U^*$ using the `binom.test` function in R with `x` being the number of (weighted) unemployed people in the CPS, `n` being the number of people in the labor force in the CPS, and `p` being U^*
- We would expect such a test to fail to reject the null hypothesis in 19 out of 20 months when $U_t = U^*$ at the $\alpha = 0.05$ level.

For some research area that you are familiar with, make the best case you can for applying conventional Frequentist practice to some other statistic besides the unemployment rate. Then, criticize your arguments from an applied statistics perspective.

2.2 Getting CPS Data

For the rest of this problem, you are going to need to download some 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/ .

2.3 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.

Write a short (about 30 lines) function in the Stan language that draws from the prior predictive distribution for a model that avoids this criticism. It should return a matrix with S rows (the number of simulated draws) and N columns (the number of individuals) where each cell is one of the following values:

- -1 if the person is simulated to be not in the labor force
- 0 if the person is simulated to be in the labor force but simulated to not have a job
- 1 if the person is simulated to be in the labor force and simulated to have a job

In doing so, you should assume that the two error terms — one that affects the probability that the person is (not) in the labor force and the other that affects the probability that the person does (not) have a job conditional on them being in the labor force — are distributed bivariate standard normal with exogenous unknown correlation ρ (but $\mu_1 = 0 = \mu_2$ and $\sigma_1 = 1 = \sigma_2$). There was a Stan function to draw from a bivariate normal distribution in Week 1. You can use whatever variables you like from the `CPS` dataset to form the linear predictors and you can use whatever priors you like for the intercept and slope(s) that comprise those linear predictors.

For each iteration s up to the S that you specify, you should draw a realization of the exogenous unknowns from their respective prior distributions and use those realizations to form a linear predictor for (not) being in the labor force and another linear predictor for (not) being employed if that person *were* in the labor force. Note that these two linear predictor vectors are both of size N because each person has both propensities irrespective of whether or not they actually *are* in the labor force. Then, on iteration s, for each person n up to N (the number of people in the CPS that month), you should draw two error terms from the bivariate standard normal distribution with correlation $\tilde{\rho}$ and add these two error terms to the respective linear predictors for person n. If the sum of the first linear predictor for person n and their error realization is less than zero, then the person is simulated to be not in the labor force and you can just put a -1 into the s-th row and n-th column of the output matrix and move onto the next person. Conversely, if the sum of the first linear for person n and their error realization is greater than zero, then person is simulated to be in the labor force in which case you need to consider the sum of the second linear predictor for person n and the second error term. If that is less than zero, then the person is simulated to be unemployed, and if that is greater than zero, then the person is simulated to have a job. You are going to need `if` and `else` clauses inside the double loop of your Stan program to handle this, but those work the same way in Stan as in R and almost all other computer languages.

Be sure to upload your Stan program along with your RMarkdown file and the HTML or PDF file it generates when you turn in your homework.

2.4 Prior Predictive Checking

Use the `expose_stan_functions` function in the **rstan** package to translate the Stan function you wrote in the previous subproblem to C++, compile it, and make it available to R. Call your function to produce the $S \times N$ matrix of -1, 0, and 1 values and check that it is reasonable in various respects. For example, about two-thirds of adults should be in the labor force and among those, about five percent should be unemployed.

2.5 Estimating a Probit Model

For this subproblem, we are going to treat the individual's decision to be not in the labor force (`nilf`) as an exogenous known — i.e. do what labor microeconometricans criticize — because the `stan_glm` function is only capable of estimating univariate outcome models, which in this case is for the `empl` variable. Call `stan_glm` in the **rstanarm** package with following arguments:

```
data = CPS, family = binomial(link = "probit"), subset = nilf == 0, QR = FALSE
```

You should specify the `formula`, `prior`, and `prior_intercept` values to be like those you used in the previous subproblem for the "second" part of the model for (not) being employed conditional on being in the labor force. In this problem, you presumably want to call the functions you use for `prior` and `prior_intercept` with `autoscale = FALSE` so that they are interpreted in their raw units rather than in terms of standard deviations.

2.6 Inference

Call the `as.matrix` function on the result of `stan_glm` in the previous subproblem in order to obtain a $S \times K$ matrix where K is the number of parameters estimated including the intercept (relative to uncentered predictors). For each parameter, estimate the posterior probability that the parameter is positive from its proportion of positive draws. How does this differ from a frequentist test of a null hypothesis that a parameter is zero against the one-sided alternative hypothesis that it is positive?