GR5065 Homework 2

Ben Goodrich

Due February 21, 2022 at 4PM

1 Gross Domestic Product

Gross Domestic Product can be conceptualized as the total output of an economy, and is used in essentially every social science model where countries are the unit of analysis. However, no country "counts" up its Gross Domestic Product. Rather, the Gross Domestic Product "data" that gets reported is the output of a model, and different models produce different predictions of the same concept. In addition, there are (at least) two different ways of modeling it, either as total expenditures or as total income received. To better understand these two methods, closely read

https://obamawhitehouse.archives.gov/sites/default/files/docs/gdo_issue_brief_final.pdf

On Thursday, February 24th, at 8:30AM the U.S. Bureau of Economic Analysis will release data on real *growth* of Gross Domestic Income (GDI) and Gross Domestic Product (GDP) for the fourth quarter of 2021, i.e. the percentage increase after removing inflation expressed as an annual rate, $100\left((1+g)^4-1\right)$.

To reduce confusion, let's use to GDP* to refer to the *concept* of Gross Domestic Product growth and GDP to refer to its estimate via the expenditure approach. In this problem, we are going to prepare for a Bayesian analysis conditional on this widely-anticipated data release. You can download these data series up through the third quarter (starting on July 1 and ending on September 30) of 2021 via

```r
library(readr)
library(dplyr)
FRED <- "https://fred.stlouisfed.org/graph/fredgraph.csv"
dataset <- read_csv(paste0(FRED, "?id=A261RL1Q225SBEA,A191RL1Q225SBEA"), na = ".") %>%
  na.omit %>%
  rename(quarter_startdate = DATE, GDI = A261RL1Q225SBEA, GDP = A191RL1Q225SBEA)
```

As can be seen from

```r
tail(dataset, n = 7)
```

```
## # A tibble: 7 x 3
##   quarter_startdate   GDI   GDP
##   <date>            <dbl> <dbl>
## 1 2020-01-01         -0.8  -5.1
## 2 2020-04-01        -32.7 -31.2
## 3 2020-07-01         24.4  33.8
## 4 2020-10-01         19.6   4.5
## 5 2021-01-01          6.3   6.3
## 6 2021-04-01          4.3   6.7
## 7 2021-07-01          5.8   2.3
```

these two estimates have been rather different since the start of the pandemic, except in the first quarter of 2021. But they typically differ by a couple of percent each quarter even going back to 1947. In recent decades, outside of recessions, real GDP growth in the United States tends to be between 2% and 3% per year.

1.1 Normal Prior

Let $\mu_I = \mu = \mu_P$, which is to say that *reported* GDI and GDP growth have the same expectation, and that expectation is the unknown true GDP* growth. If you think back to the fourth quarter of 2021, the unemployment rate dropped from 4.6 percent in October to 3.9 percent in December, but people became increasingly worried about supply chain backlogs leading to inflation, which increased by one percent over those three months. Also, the Omicron variant started to spread in coastal cities during the last two weeks of the year. In principle, μ can be any real number but the chances of it being negative in the fourth quarter of 2021 are negligible because the economy was still catching up to where it was before the pandemic.

Describe your beliefs about μ (the true annualized GDP* growth) during the fourth quarter of 2021 with a normal distribution that has expectation m and standard deviation s (which you should choose values for). Remember that Bayesian inference is predicated on the fact that you do not know μ. However, Bayesian inference requires that you know what you *believe* about μ and express that belief in terms of some probability distribution that you can subsequently update. Choose m and s and somehow plot your normal prior in R (but you do not have to use ggplot if you are not that familiar with it).

1.2 Bivariate Normal Likelihood

Let's further assume that $\sigma_I = \frac{7}{3} = \sigma_P$, which is to say that each measured series is within $\mp\frac{7}{3}$ of μ in about two out of every three quarters. Finally, assume that the correlation between the between the errors in each series is $-\frac{1}{10}$. Ordinarily, we would want to estimate the standard deviations and correlation along with μ (or better μ_q for each quarter), but we have not yet learned about how to obtain the posterior distribution when there are multiple unknown parameters. With these simplifying assumptions, reported GDI and GDP growth can be modeled as bivariate normal with unknown μ (and thus unknown μ_I and μ_P) but known σ_I, σ_P and ρ.

Write a likelihood function in R with this signature

```
L <- function(mu, GDI, GDP) {
  # fill in the bivariate normal likelihood of mu,
  # (to be) evaluated at the values of GDI and GDP growth for the 4th quarter of 2021
  # using sigma_I = 7 / 3, sigma_P = 7 / 3, and rho = -1 / 10
}
```

Recall that the likelihood function is the same expression as the bivariate normal PDF for GDI and GDP growth conditional on μ (and the other parameters) but now taken to be a function of the unknown μ. Thus, it will not integrate to 1 over the parameter space of μ.

1.3 Simulating Reported GDI and GDP Growth

Take one million random draws from your normal prior distribution for μ (given m and s). For each of those one million realizations of μ, take one random draw of both GDI and GDP growth under the assumptions of the previous subproblem. Create some useful plot of the distribution of what you believe the US government will report for GDI and GDP growth on February 24th. You will notice that the estimated standard deviation over one million draws of each series exceeds $\frac{7}{3}$. Explain why.

1.4 Posterior

Write a function in R to evaluate the numerator of Bayes rule for the conditional distribution of μ given reported GDI and GDP growth (as well as the assumed values for the other parameters). Note that this R function must work with a *vector* of μ values and return a vector of the same size containing trivariate density values between μ, GDI growth, and GDP growth.

Then write code to evaluate the denominator of Bayes rule. Note that you will not be able to actually evaluate this denominator until the values of GDI and GDP growth are revealed on February 24th, but you should write the code now and specify `eval = FALSE` in the header of the R code chunk so that your RMarkdown file will successfully knit to a PDF.

1.5 Gross Domestic Output

The paper linked above advocates calculating Gross Domestic Output (GDO), which is just the average of GDI and GDP growth, as a better measure of GDP* than either GDI or GDP is individually. Figure 4 of the paper also shows that a simple average is close to the optimal weighted average, which would put slightly more weight on GDI growth.

How does GDO conceptually compare and contrast with the posterior distribution of GDP* conditional on GDI and GDP growth?

2 Bowling

If your working directory is Assignments/HW2/, execute

```r
# defines Pr(x, n = 10, theta = 1) and Omega as 0:10
source(file.path("..", "..", "Week04", "bowling.R"))
```

The `Pr` function is equivalent to

$$\Pr\left(x \mid n, \theta\right) = \frac{\log_{n + 1 + \theta}\left(1 + \frac{1}{n + \theta - x}\right)}{1 - \log_{n + 1 + \theta}\left(\theta\right)}$$

like we used at the end of Week04. The function body looks different because it tries to be more numerically accurate for small or large values of the inability parameter, $\theta \geq 0$, but it should yield the same probabilities as in the mathematical definition above.

2.1 Prior

Think about what you believe θ to be for a woman who is bowling in the Olympics, if bowling were an Olympic sport (which it is not because it is too boring). In particular, choose an exponential prior for θ with expectation m (different from the m in the previous problem). Somehow plot your exponential prior in R.

2.2 Marginal Probability of a Frame

Create an 11×11 matrix whose cells contain the marginal probability of knocking down x_1 pins on the first roll (the row index) of a frame of bowling and x_2 pins on the second roll (the column index) of that same frame of bowling, irrespective of θ but taking into account your exponential prior beliefs about θ for this woman. As in class, if the bowler gets a strike on the first roll of a frame, then presume the bowler is guaranteed to knock down zero pins on the second roll of that frame (which does not actually happen in genuine bowling).

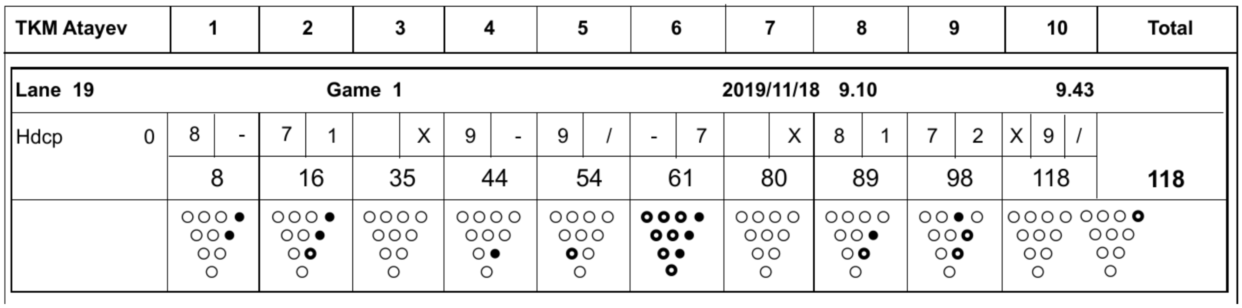
What is the sum over all the elements of this matrix, what should the sum be in theory, and what accounts for the difference?

2.3 Data

Once you have completed the previous two subproblems, ask Ben on Ed Discussions for a link to a bowling scorecard for an elite woman bowler. The scorecard will look like

There are several things to note about such scorecards:

- The column headings indicate the frame and there are ten frames in a game



TKM Atayev	1		2		3		4		5		6		7		8		9		10			Total
Lane 19					Game 1								2019/11/18 9.10						9.43			
Hdcp 0	8	-	7	1		X	9	-	9	/	-	7		X	8	1	7	2	X	9	/	
	8		16		35		44		54		61		80		89		98		118			118

Figure 1: a bowling scorecard

- The two boxes immediately below the frame number indicate the number of pins knocked down on the first and (in the event of a non-strike) second roll of that frame
  - An `X` for a frame indicates a strike on the first roll even though it is usually marked in the second box
  - A `\` in the second box for a frame indicates a "spare", meaning that *all* of the remaining pins were knocked down on the second roll of that frame
  - A `-` in either box means that zero pins were knocked down on that roll
- The circles around some numbers in the first box for a frame are not meaningful in this homework. They merely indicate that the pins that remain upright after the first roll of that frame were not totally adjacent to each other.
- The integer below the two boxes for the number of pins knocked down on each roll of that frame is the bowler's score *through* that frame. Scoring in bowling is somewhat convoluted and defined in the last subproblem.
- The triangles at the bottom indicate (in black) which pins remained upright after the first roll of that frame of bowling and which were knocked down on the second roll (with a white dot), but the particular pins that were or were not knocked down are not relevant on this homework
- In the tenth frame only, if the bowler gets a strike on the first roll, then the bowler is awarded two additional rolls starting with $n = 10$ pins upright again. If the bowler also gets a strike on the first additional roll, then $n = 10$ pins are put upright for the second additional roll. In the tenth frame only, if the bowler does not get a strike but gets a spare on the second roll, then the bowler is awarded one additional roll with $n = 10$ pins upright. Thus, it is possible to get zero, one, two, or three strikes on the tenth frame, and the above figure the bowler got a strike on the first roll and a spare between the two additional rolls.

For each of the six scorecards you receive the link to, create a separate matrix with 10 rows and 3 columns that records the number of pins knocked down on that frame and that roll. Note that the third column should be `NA` except perhaps for the tenth frame because no other frame has a third roll. Also, now you should use `NA` for any second roll when the first roll is a strike. For the above scorecard, this matrix would look like

```r
game_1 <- matrix(c(
   8,  2, NA,
   7,  1, NA,
  10, NA, NA,
   9,  0, NA,
   9,  1, NA,
   0,  7, NA,
  10, NA, NA,
   8,  1, NA,
   7,  2, NA,
  10,  9,  1),
  nrow = 10, ncol = 3, byrow = TRUE)
```

2.4 Posteriors

For *your* bowler and prior, obtain the posterior distribution of θ given the entire first game of bowling, assuming that frames are independent of each other. This would be similar to what we did at the end of Week03, except now θ is a continuous parameter. Plot this posterior distribution of θ.

Then, for each of the next five bowling games that you have for your bowler, plot the posterior distribution of θ conditioning on all the data *through* that game of bowling. So, there will be six plots in total for this subproblem. Note that it is presumably easier in this case not to use an intermediate posterior distribution as a prior for θ when conditioning on the next game. Rather, stick with your original exponential prior for θ but condition on multiple games of bowling as you go along. Both approaches will yield the same answer, but the latter is easier to implement. How does the final posterior distribution compare to your prior distribution from the first subproblem?

2.5 Simulate Score Distribution

Although it is plausible to assume that pins knocked down are independent across frames of bowling, the scoring rules for bowling are not that simple. The "traditional" scoring rules are described at

https://en.wikipedia.org/wiki/Ten-pin_bowling#Traditional_scoring

but basically they say that if you get a strike on frame j, then your score for frame j is 10 plus the number of pins knocked down on your next two rolls (for a maximum of 30). And if you get a spare on frame j, then your score for frame j is 10 plus the number of pins knocked down on the first roll of frame $j+1$ (for a maximum of 20). The tenth frame is scored as described above, and the additional rolls make the scoring consistent with the previous nine frames. A perfect game would achieve the maximum possible score of 300 with twelve consecutive strikes.

Write a R function called `score` that takes a matrix with 10 rows and 3 columns like that above that indicates how many pins were knocked down on that frame and that roll and returns the bowler's score at the end of the game according to "traditional" scoring rules.

Then, take 1000 draws from your exponential prior (not posterior) for θ. For each of those 1000 realizations of θ, simulate one entire game of bowling (represented by an 11×3 matrix), and pass that matrix to your `score` function. In the end, you should have 1000 integers between 0 and 300. Create a histogram out of these 1000 integers. What is the mode of the bowler's score under your prior beliefs about θ?