APSTA-GE 2123 Assignment 1

Due by 1:45 PM on April 20, 2022

1 On p-values and statistical significance

This problem requires you to read four short journal articles:

(A) Wasserstein and Lazar (2016)
(B) Wasserstein, Schirm, and Lazar (2019)
(C) Kafadar (2021)
(D) Kafadar, et al. (2021)

The context for these four papers is as follows. In 2015, the American Statistical Association (ASA) organized a group of about 20 researchers — most of whom were statisticians or from related fields and some of whom were Bayesian — to make a statement about p-values and statistical significance. In 2016, the researchers agreed to a statement, which was then approved by the Executive Committee of the ASA. Wasserstein and Lazar (2016) summarizes that history and includes the text of the statement, which was interpreted as being against the prevailing use of p-values but stopped well short of recommending that applied researchers use Bayesian methods. Naturally, almost all of the 20 researchers also published articles (that you do not have to necessarily read) in that same issue of *The American Statistician* explaining their disagreements with the statement that they agreed to.

Wasserstein, Schirm, and Lazar (2019) responded to a criticism that the ASA statement was opposed to the prevailing use of p-values but did not include enough constructive recommendations for applied statisticians to follow. Indeed, that issue of *The American Statistician* in 2019 included 43 other papers (that you do not have to necessarily read) with such recommendations, which is perhaps too many. Although Wasserstein and Lazar were among the 20 or so researchers involved in the 2016 ASA statement, in their 2019 article with Schirm, they were writing for themselves rather on behalf of any particular group.

Nevertheless, some applied researchers may have perceived Wasserstein, Schirm, and Lazar (2019) or the other 43 articles in that issue of *The American Statistician* to be some sort of an extension of the policy that the Executive Committee of the ASA voted to approve in 2016. As a result, Kafadar (2021) organized yet another panel of researchers — most of whom were statisticians or were from related fields but only one of whom was part of the 2016 panel — that agreed to a brief statement in Kafadar, et al. (2021) that was interpreted as being much more sympathetic to the use of p-values. Kafadar also happens to be the editor of *The Annals of Applied Statistics* where it was published, which sort of makes it the prevailing policy of that journal but not the ASA.

Your task is to write a few paragraphs outlining how applied statisticians could conduct a valuable analysis that is, to the extent possible, consistent with the four papers above and could be considered Frequentist, even if it differs somewhat with what has historically been considered Frequentist. In particular, you should address what value is added to an otherwise decent empirical paper by including a p-value if that p-value is not being referred to some threshold for statistical significance, such as 0.05. In addition, in light of the fact that the 2016 statement is an official policy recommendation of the ASA and the articles in the 2019 issue of the *The American Statistician* are not, do you think the ASA should be issuing official recommendations and if so, do you think that an additional one is warranted in light of the developments on this front since 2016?

2 Economic Growth

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 the two methods that are used by the U.S. government, read

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

To reduce confusion, let's use μ to refer to the *concept* of Gross Domestic Product growth and GDP to refer to its estimate via the expenditure approach, in contrast to Gross Domestic Income (GDI) as its estimate via the income approach. You can download illustrative data on GDI and GDP growth from 1947 through the end of 2021 via

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

although you do not need to do anything with that data on this subproblem except to note that GDI and GDP growth estimates are rarely in agreement with each other in any particular quarter, despite being estimates of the same concept.

In this problem, we are going to consider GDI and GDP data — that have not been released yet — for the first quarter of 2022. Recall that in January of 2022, Omicron was widespread and although businesses tended not to be completely shut down, many workers were out sick and many customers were either sick or limiting their movements. As a result, the unemployment rate was 4.0%. February was more typical, until the Russian invasion of Ukraine, which caused energy prices to increase rapidly and disrupted trade involving both countries. Nevertheless, the unemployment rate dropped to 3.6% during March. You can find on the internet many predictions of μ for the first quarter of 2022 to guide your analysis.

2.1 Normal Prior

Describe your beliefs about μ 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.

2.2 Prior Predictive Distribution

Assume that *reported* GDI and GDP growth are distributed bivariate normal with the same expectation, and that their respective expectations, μ_I and μ_P, are both equal to the unknown true growth, μ, in the first quarter of 2022. 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 errors in each series is $-\frac{1}{10}$. Ordinarily, we would express prior beliefs about σ_I, σ_P, and ρ but assuming values for them simplifies the analysis by making the only unknown parameter be μ.

Draw $100,000$ times from the univariate normal prior for μ that you specified in the previous subproblem and for each of those realizations of μ, draw from the bivariate normal distribution of GDI and GDP growth for the first quarter of 2022 using the aforementioned values of σ_I, σ_P, and ρ. Make a scatterplot of the $100,000$ GDI realizations on the horizontal axis and the $100,000$ GDP realizations on the vertical axis.

Why does the standard deviation of these $100,000$ realizations of GDI and GDP for the first quarter of 2022 substantially exceed $\frac{7}{3}$?

2.3 Posterior Distribution of μ

Data on GDI and GDP growth for the first quarter of 2022 do not yet exist, but we can prepare for the data release by writing a function (of μ in the mathematical sense) to compute the posterior PDF of μ given data on GDI and GDP growth, as well as m, s, σ_I, σ_P, and ρ.

First, write a function called `numerator` that takes three separate arguments: `mu`, `GDI`, and `GDP`. `GDI` and `GDP` will be real scalars. Although μ is conceptually a scalar, your `numerator` function should treat it as a vector of possible values for economic growth and return a vector of the same size that computes the numerator of Bayes Rule under the assumptions of the previous subproblems. Ordinarily when doing numerical computations, we would work with the natural logarithm of the numerator of Bayes Rule, but in this subproblem you should not take logarithms and just work with density units.

Second, write a function called `denominator` that takes two separate arguments: `GDI` and `GDP`. The `denominator` should internally call a numerical integration function to calculate the area under the curve implied by the `numerator` function, evaluated at the same values of `GDI` and `GDP`. Although you do not know the reported values of `GDI` and `GDP` yet, you can test your function with plausible guesses for them. In other words, your `denominator` function defines a bivariate PDF for future `GDI` and `GDP` irrespective of μ but conditional on m and s that you drew from in the previous subproblem.

Third, make a plot with μ on the horizontal axis and the ratio of `numerator` to `denominator` on the vertical axis, evaluated at some plausible guesses for `GDI` and `GDP`. If the reported GDI and GDP growth turned out to be those values, how would you describe your posterior uncertainty about μ?

2.4 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 μ 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 μ conditional on GDI and GDP growth?

3 Linear Model

Read the Wikipedia article on the Capital Asset Pricing Model (CAPM)

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

which implies a linear statistical model

$$\underbrace{R_{it} - R_{ft}}_{y_t} = \alpha + \beta_i \left(\underbrace{R_{mt} - R_{ft}}_{x_t} \right) + \epsilon_{it}$$

where α is an intercept, β is a coefficient, $\epsilon_{it} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma)$, σ is the standard deviation of the errors, and

- R_{it} is the return of the stock of company i since period $t-1$
- R_{ft} is the return of a "risk-free asset", usually proxied by short-term U.S. government bonds since period $t-1$
- R_{mt} is the return of "the market", usually proxied by some broad index of stocks of major companies

In this subproblem, we are going to consider weekly returns from one Wednesday to the next since mid-2020, which comprises $N = 95$ observations. Note that for simplicity, we assume that the errors are independent over time even though that might not be the case. You can download and process data for (proxies of) R_{ft} and R_{mt} with

```
library(tidyquant) # you may have to first install this package

# short-term U.S. government bonds are essentially "risk-free"
R_f <- tq_get("SGOV", from = "2020-06-01", to = "2022-04-01") %>%
  filter(weekdays(date) == "Wednesday") %>%
  transmute(R_f = (adjusted - lag(adjusted)) / lag(adjusted)) %>%
  na.omit %>%
  pull

# the S&P500 is essentially "the market"
R_m <- tq_get("SPY", from = "2020-06-01", to = "2022-04-01") %>%
  filter(weekdays(date) == "Wednesday") %>%
  transmute(R_m = (adjusted - lag(adjusted)) / lag(adjusted)) %>%
  na.omit %>%
  pull
```

but you also need to choose one public company that is listed on the U.S. stock exchanges but is *not* among those that comprise the S&P500. Put its stock market symbol into the following code chunk to define `R_i`

```
R_i <- tq_get(stop("put a stock symbol here"), from = "2020-06-01", to = "2022-04-01") %>%
  filter(weekdays(date) == "Wednesday") %>%
  transmute(R_i = (adjusted - lag(adjusted)) / lag(adjusted)) %>%
  na.omit %>%
  pull
```

A list of the S&P500 companies (that you should *not* use for `R_i`) can be found at

https://en.wikipedia.org/wiki/List_of_S%26P_500_companies

In the CAPM, the outcome and the predictor are already expressed relative to R_{ft} so in this situation, you should not further center the predictor once you construct it. Thus, α can be interpreted as the expected "excess" return of company i beyond the risk-free asset when the "excess" return of the market beyond the risk-free asset is zero.

3.1 Priors

To express your prior beliefs about α, β, and σ, first `source` in the GLD_helpers.R file from Week02. Then, solve for the asymmetry and steepness hyperparameters of the Generalized Lambda Distribution that are consistent with your prior quantiles on each of these three parameters. If you would like to use some *previous* data to help you figure out reasonable priors, just call `tq_get` again with dates before the covid pandemic.

3.2 Prior Predictive Distribution

First, draw 1000 realizations of α, β, and σ from their respective Generalized Lambda Distributions under your priors.

Second, for each of these 1000 parameter realizations, draw from the predictive distribution of R_{it} for $t = 1$, $t = 2$, ..., $t = 95$ by adding realizations of the errors to the realization of the conditional expectation implied by the realizations of α and β along with the data on R_{mt} and R_{ft}.

Third, make a histogram of the marginal distribution of $R_{it} - R_{ft}$ to verify that it looks reasonable.

How and why does this prior predictive distribution differ from a normal distribution?

3.3 Posterior Distribution

Use Stan to draw from the posterior distribution of α, β, and σ conditional on your GLD priors and data. There will also be draws of p_α, p_β, and p_σ that correspond to the CDF values of each of these parameters, but you can largely ignore those because they are of no substantive interest once you have posterior draws of α, β, and σ.

Make sure to adjust the `adapt_delta`, `max_treedepth`, and `iter` settings as necessary to avoid any warnings from Stan about difficulties sampling from your posterior distribution.

Create a `pairs` plot of each two-dimensional plane through the posterior distribution (excluding planes involving p_α, p_β, and p_σ). On the basis of this `pairs` plot, how would you describe the marginal distributions and bivariate dependence in your beliefs about the three parameters after having conditioned on the data?

3.4 Interpretation

How would you describe the consistency, or lack thereof, between your posterior beliefs about the parameters and the CAPM?

3.5 Posterior Predictive Distribution

Draw 4000 times from the posterior predictive distribution of $R_{it} - R_{ft}$ for each of $t = 1$, $t = 2$, ..., $t = 95$ using your 4000 realizations of α, β, and σ from the posterior distribution.

What proportion of the $N = 95$ observations fall

- Below the $\frac{1}{3}$ quantile of their posterior predictive distribution
- Above the $\frac{2}{3}$ quantile of their posterior predictive distribution
- Between the $\frac{1}{3}$ and $\frac{2}{3}$ quantiles of their posterior predictive distribution

and what does this tell you about how well the CAPM fits these data?