APSTA-GE 2123 Assignment 1 Answer Key

Due by 1:45 PM on April 20, 2022

1 On p-values and statistical significance

The purpose of this problem is to think critically about what should constitute a Frequentist analysis. Doing so helps to understand why Bayesians object to the Frequentist framework.

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)

There are many ways that you could go in responding to these four articles. Let's first summarize their main points:

- Wasserstein and Lazar (2016) enumerated six principles regarding p-values and statistical significance:
    1. P-values can indicate how incompatible the data are with a specified statistical model
    2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone
    3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold
    4. Proper inference requires full reporting and transparency
    5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result
    6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis
- Wasserstein, Schirm, and Lazar (2019) went farther than the ASA statement that was reported in Wasserstein and Lazar (2016) and attempted to be more constructive about what applied statisticians should do besides not using the phrase (or the concept) "statistically significant" or otherwise attempting to dichotomize a continuous concept. Nevertheless, Wasserstein, Schirm, and Lazar (2018) recommends that p-values be calculated, reported, and interpreted correctly.
- Kafadar (2021) expressed concern that Wasserstein, Schirm, and Lazar (2019) might be misinterpreted as official ASA policy when none of the disparate recommendations made therein had been voted upon by the ASA committee that approved Wasserstein and Lazar (2016). After making the hardly controversial point that p-values and statistical significance are widely used in scientific communication and legal judgments, Kafadar (2021) notes that the original proponents of p-values wanted to see a collection of independent studies that test the same null hypothesis and find small p-values.
- Kafadar, et al. (2021) enumerates a competing set of principles to that in Wasserstein and Lazar (2016) without providing much in the way of a justification for them:
    1. Capturing the uncertainty associated with statistical summaries is critical
    2. Dealing with replicability and uncertainty lies at the heart of statistical science
    3. The theoretical basis of statistical science offers several general strategies for dealing with uncertainty
    4. Thresholds are helpful when actions are required

5. In summary, P-values and significance tests, when properly applied and interpreted, increase the rigor of the conclusions drawn from data

There is some overlap in the recommendations of Wasserstein, Schirm, and Lazar (2019) and those of Kafadar, et al. (2021) in that an empirical paper could report all p-values and provide the correct interpretation of them as the conditional probability of observing a particular function of the data that is more extreme than that observed in the present dataset, given that the null hypothesis is true. Of course, interpreting this p-value correctly requires that one conceptualize probability as the proportion of randomly-sampled datasets that yield a more extreme test statistic in the limit as the number of randomly-sampled datasets goes to ∞.

However, the fourth principle in Kafadar, et al. (2021) that thresholds, such as 0.05 are helpful when actions are required seems to be in conflict with the third principle in Wasserstein and Lazar (2016). Kafadar, et al. (2021) elaborates that "If thresholds are deemed necessary as a part of decision-making, they should be explicitly defined based on study goals, considering the consequences of incorrect decisions." But even that would be insufficient to satisfy Wasserstein and Lazar (2016)'s principles, which were approved by the ASA.

One concern I have is that the fourth principle in Kafadar, et al. (2021) could be invoked by applied statisticians and teachers who simply do not want to change the status quo and do not want to consider the tenets of (mostly Bayesian) decision theory. From their perspective, the attractive part about thresholds is that they serve as a convention whereby a researcher proceeds as if $\theta = \theta_0$, where θ_0 is the null hypothesized value of a parameter θ if the null hypothesis is not rejected at the 0.05 level. Furthermore, the established convention is that the researcher is allowed to proceed as if $\theta = \widehat{\theta}$ if and only if the null hypothesis that $\theta = \theta_0$ is rejected in favor of the alternative hypothesis that $\theta \neq \theta_0$. But it does not follow from the fact that $\theta \neq \theta_0$ that $\theta = \widehat{\theta}$. No Frequentist has or can advocate for these convention because these conventions because they are at odds with Frequentist definitions and its conception of $\widehat{\theta}$ as a random variable. Wasserstein and Lazar (2016) and Wasserstein, Schirm, and Lazar (2019) are, in essence, criticizing these conventions as being scientifically indefensible.

All four articles highlight the importance of communicating uncertainty without noting the obvious corollary that the Bayesian approach offers the most encompassing measure of uncertainty about the parameters, functions thereof (such as predictions), and the ordering of models. None attempt to dispel the widespread misconception that standard errors, confidence intervals, etc. provide a measure of uncertainty about the parameters, when in fact they are statements about the uncertainty of the estimator conditional on the parameters. A p-value is a unitless measure of evidence against the point null hypothesis but does not provide any measure of certainty about what the parameter involved in the null hypothesis is likely to be. As such, it provides limited value-added to an analysis but perhaps would not hurt if it were interpreted correctly.

It is debatable whether the ASA should be setting rules for statistical analysis or whether any rules the ASA might set would be followed given that there is so much disagreement among applied statisticians as to how a statistical analysis should be conducted. Nevertheless, given that the ASA voted to affirm the principles in Wasserstein and Lazar (2016), it would seem prudent for the ASA to provide more clarity in light of the tension between Wasserstein, Schirm, and Lazar (2019) and Kafadar, et al. (2021) as to whether significance thresholds should be used at all.

2 Economic Growth

The purpose of this problem is to think about a prior predictive distribution, which is an important concept in Bayesian analysis that has no analogue in Frequentist analysis. If the prior predictive distribution of an observable is reasonable, then the priors on the unobserved parameters cannot be that bad. Researchers often find it easier to assess the reasonableness of an observable than an unobservable. The prior predictive distribution simply represents the researcher's beliefs about the unknown outcomes, under the model, before observing them, which in this case is not possible yet because the U.S. government will not release GDP data until April 28th and GDI data about a month later. This question also highlights the fact that a Bayesian analysis can proceed with one data point, or one data point at a time, unlike Frequentist analyses that are usually only justified as the number of observed data points approaches ∞.

2.1 Normal Prior

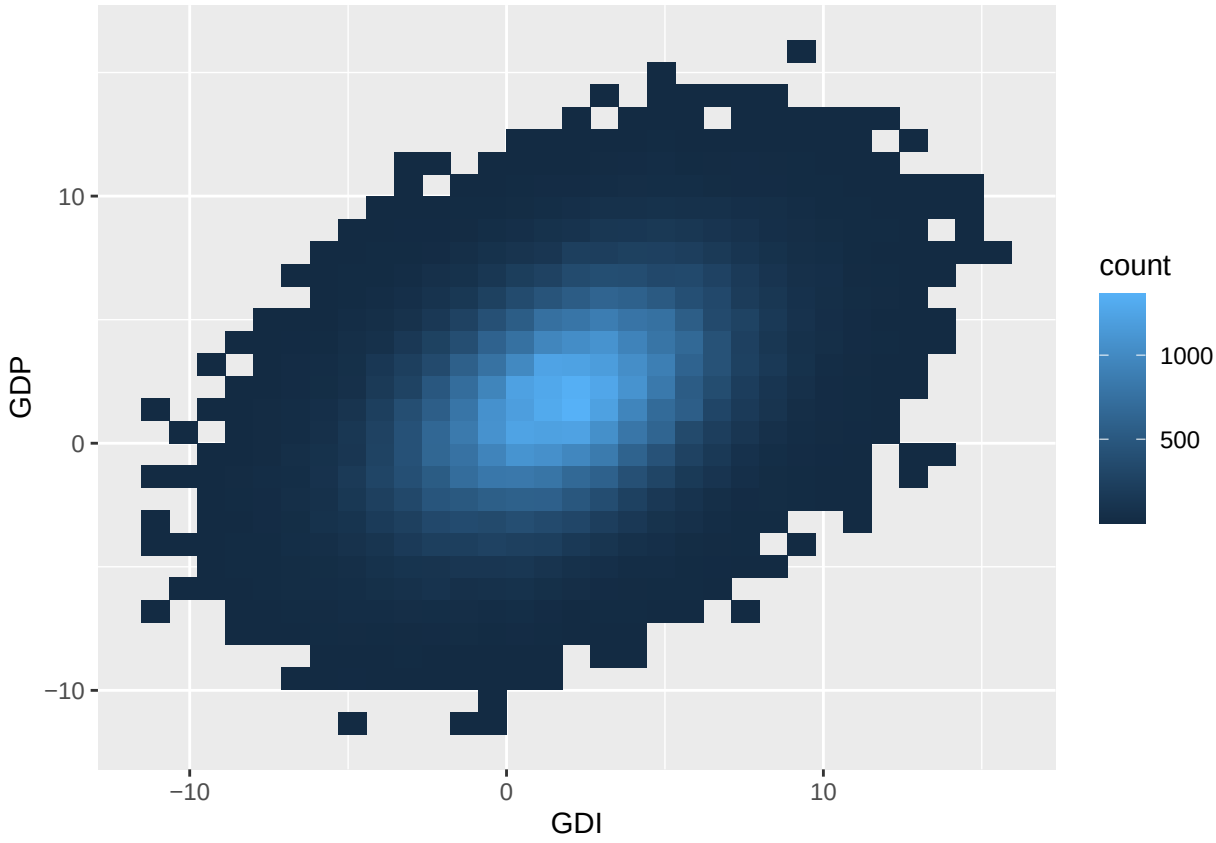
The median of the Survey of Professional Forecasters concluded in late February was for economic growth of 1.8%, although that might not fully reflect the impact of the conflict in Ukraine. That same survey also put a one-in-five chance that economic growth would be negative. A normal prior that is consistent with those facts is

```
m <- 1.8
s <- 2.15
pnorm(0, m, s)
```

```
## [1] 0.2012375
```

2.2 Prior Predictive Distribution

```
library(dplyr)
library(ggplot2)
sigma <- 7 / 3
rho <- -1 / 10
tibble(mu = rnorm(10^5, mean = m, sd = s),
       GDI = rnorm(10^5, mean = mu, sd = sigma),
       GDP = rnorm(10^5, mean = mu + rho * (GDI - mu),
                   sd = sigma * sqrt(1 - rho^2))) %>%
  ggplot() + geom_bin_2d(aes(x = GDI, y = GDP))
```



The variance in both GDP and GDI under the prior predictive distribution is much higher than $\sigma = \frac{7}{3}$ because the prior predictive distribution also incorporates uncertainty about μ itself. In other words, $\sigma = \frac{7}{3}$ is the assumed standard deviation in GDP and GDI, *given* μ. When μ is not, in fact, given, then our uncertainty in

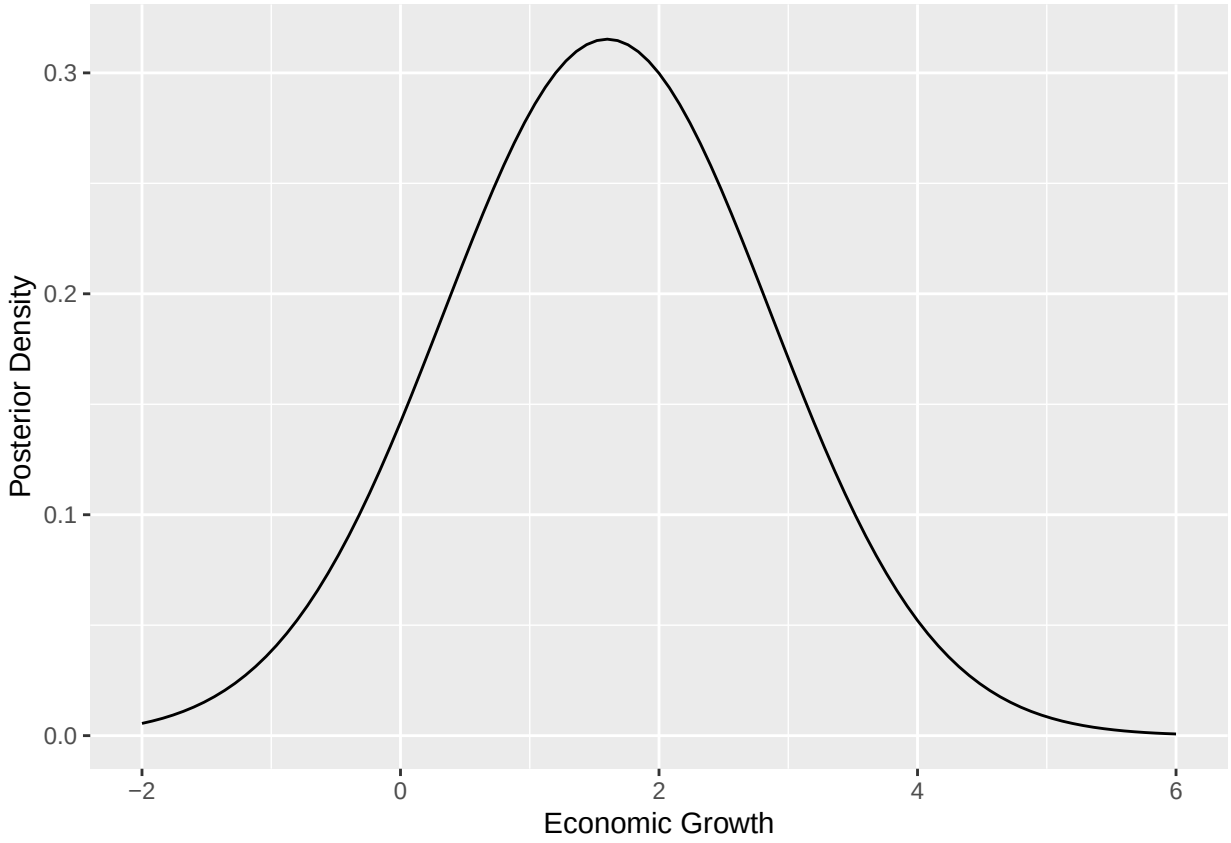
what will be reported for GDP and GDI depends both on our uncertainty about μ and on our uncertainty about how well the government's model and preliminary data capture μ in their reported GDP and GDI.

2.3 Posterior Distribution of μ

```
numerator <- function(mu, GDP, GDI) {
  dnorm(mu, mean = m, sd = s) *
    dnorm(GDP, mean = mu, sd = sigma) *
    dnorm(GDI, mean = mu + rho * (GDP - mu), sd = sigma * sqrt(1 - rho^2))
}

denominator <- function(GDP, GDI) {
  integrate(numerator, lower = -Inf, upper = Inf, GDP = GDP, GDI = GDI)$value
}
```

```
ratio <- function(mu, GDP, GDI) numerator(mu, GDP, GDI) / denominator(GDP, GDI)
ggplot() + geom_function(fun = ratio, args = list(GDP = sqrt(2), GDI = pi / 2)) +
  xlim(-2, 6) + xlab("Economic Growth") + ylab("Posterior Density")
```



2.4 Gross Domestic Output

GDO is a point estimate of μ that has lower variance than either GDP or GDI alone because GDO divides their sum by two. The variance of the numerator is equal to the sum of the variances of GDP and GDI plus double the covariance (which is small). And dividing a random variable by a constant divides the variance by the square of the constant. The posterior distribution of μ is not a point estimate but it is conditional on GDP and GDP, as well as m and s plus the assumed values of $\sigma = \frac{7}{3}$ and $\rho = -\frac{1}{10}$.

Although it was not necessary for this subproblem, the correspondence with GDO can be clarified by deriving the posterior density analytically, which is a normal distribution with updated expectation m^* and updated standard deviation s^*. Conditional on GDP but not GDI, $m^* = m\frac{\sigma^2}{s^2+\sigma^2} + \text{GDP}\frac{s^2}{s^2+\sigma^2}$ and $s^* = s\sigma\sqrt{\frac{1}{s^2+\sigma^2}}$. Conditional on both GDP and GDI, $m^* = m\frac{(1+\rho)\sigma^2}{2s^2+(1+\rho)\sigma^2} + \text{GDO}\frac{2s^2}{2s^2+(1+\rho)\sigma^2}$ and $s^* = s\sigma\sqrt{\frac{1+\rho}{2s^2+(1+\rho)\sigma^2}}$. As $s \uparrow \infty, m^* \to \text{GDO} = \frac{\text{GDP}+\text{GDI}}{2}$. Thus, the GDO can be seen as an estimate of the posterior mean under an improper prior for μ, which is somehow less controversial than acknowledging that economists have very informative priors about μ each quarter.

We can also check our previous answers that were obtained numerically:

```
numerator(1.5, GDP = sqrt(2), GDI = pi / 2) / denominator(GDP = sqrt(2), GDI = pi / 2)
```

```
## [1] 0.3143007
```

```
GDO <- (sqrt(2) + pi / 2) / 2
m_star <- m * (1 + rho) * sigma^2 / (2 * s^2 + (1 + rho) * sigma^2) +
  GDO * 2 * s^2 / (2 * s^2 + (1 + rho) * sigma^2)
s_star <- s * sigma * sqrt((1 + rho) / (2 * s^2 + (1 + rho) * sigma^2))
dnorm(1.5, mean = m_star, sd = s_star)
```

```
## [1] 0.3143007
```

3 Linear Model

The purpose of this problem is to gain experience drawing from a posterior distribution and using the results to evaluate a theoretical model in the social sciences. Doing so requires the researcher to specify prior distributions for the parameters, which in this case are required to be in the family of Generalized Lambda Distributions (GLD).

The GLD is the closest thing to a "universal" probability distribution, in the sense that it can be used for random variables that are unbounded, lower bounded only, upper bounded only, or bounded on both sides, depending on the values of its four hyperparameters. Moreover, the four hyperparameters are relatively easy to specify because they depend on prior quantiles rather than prior moments that are sensitive to the exact nature of the tail behavior. Although most Bayesian analyses use prior distributions with only two parameters, such distributions do not have the flexibility to express thhe wide range of beliefs researchers may have about the parameters. Specifying four prior quantiles to run the quantile function of the GLD through forces you to critically examine your beliefs about unknown parameters, rather than merely choosing admissible hyperparameters.

```
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
```

```r
# GameStop was a highly volitile stock during this period
R_i <- tq_get("GME", from = "2020-06-01", to = "2022-04-01") %>%
  filter(weekdays(date) == "Wednesday") %>%
  transmute(R_i = (adjusted - lag(adjusted)) / lag(adjusted)) %>%
  na.omit %>%
  pull
```

3.1 Priors

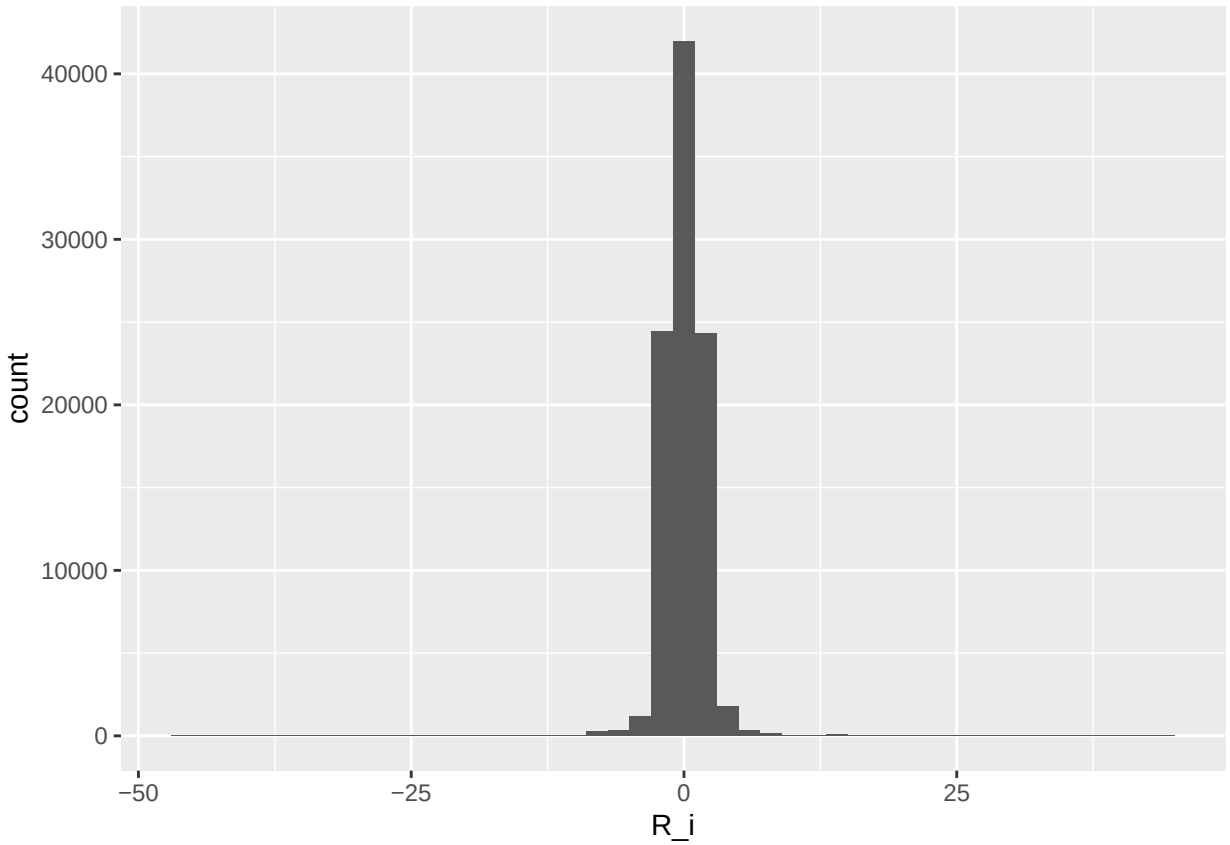
```r
source(file.path("..", "..", "Week2", "GLD_helpers.R"))
a_s_alpha <- GLD_solver(lower_quartile = -1, median = 0, upper_quartile = 1,
                        other_quantile = 2, alpha = 0.95)
a_s_beta  <- GLD_solver_LBFGS(lower_quartile = 0.5, median = 1,
                              upper_quartile = 2, other_quantile = 0, alpha = 0)
a_s_sigma <- GLD_solver_LBFGS(lower_quartile = 0.25, median = 0.5,
                              upper_quartile = 0.9, other_quantile = 0, alpha = 0)
```

3.2 Prior Predictive Distribution

```r
alpha <- qgld(runif(1000),
              median = 0, IQR = 2, asymmetry = a_s_alpha[1], steepness = a_s_alpha[2])
beta  <- qgld(runif(1000),
              median = 1, IQR = 2.5, asymmetry = a_s_beta[1], steepness = a_s_beta[2])
sigma <- qgld(runif(1000),
              median = 0.5, IQR = 0.65, asymmetry = a_s_sigma[1], steepness = a_s_sigma[2])
```

```r
prior_PD <- t(sapply(1:1000, FUN = function(s) {
  alpha[s] + beta[s] * (R_m - R_f) + R_f + rnorm(1, mean = 0, sd = sigma[s])
}))
```

```r
ggplot(tibble(R_i = c(prior_PD))) + geom_histogram(aes(x = R_i), binwidth = 2)
```

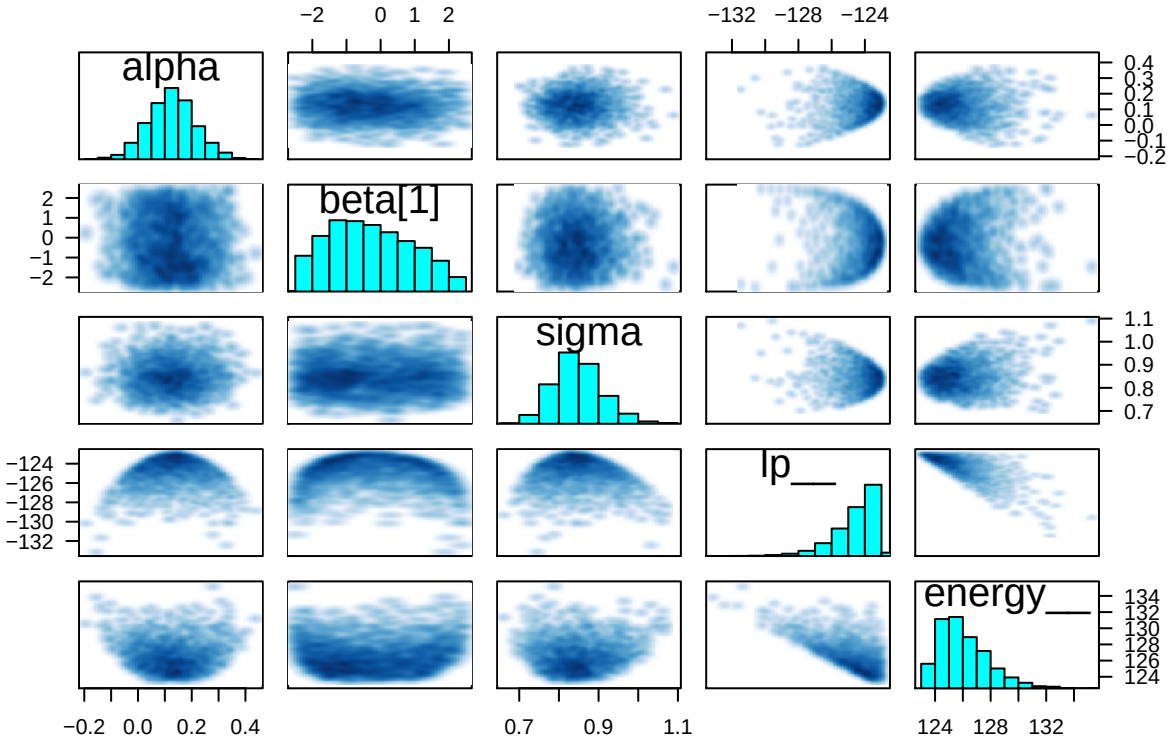


This plot has much heavier tails than a normal distribution with an occasional realization of greater than 10 in magnitude. Also, the distribution is more peaked than a normal distribution. The reason for this is the same as in question 2.2, namely that the prior predictive distribution composes a distribution for the unknown parameter with an error from a normal distribution and thereby incorporates two sources of uncertainty. In addition, the standard deviation of the normal error has a prior distribution instead of being a fixed value, which also affects the marginal distribution of our beliefs about R_i before observing it.

3.3 Posterior Distribution

```
library(rstan)
post <- stan(file.path("..", "..", "Week3", "linear.stan"),
             data = list(N = NROW(R_i), K = 1, X = as.matrix(R_m - R_f),
                         y = R_i - R_f, prior_only = FALSE,
                         m = c(alpha = 0, beta = 1, sigma = 0.5),
                         r = c(alpha = 2, beta = 2.5, sigma = 0.65),
                         a = c(a_s_alpha[1], a_s_beta[1], a_s_sigma[1]),
                         s = c(a_s_alpha[2], a_s_beta[2], a_s_sigma[2])))
```

```
pairs(post, pars = "p", include = FALSE, las = 1)
```



The posterior distributions of α and σ are roughly normal and fairly independent of each other. The posterior distribution of β is right-skewed but not very dependent on other other parameters.

3.4 Interpretation

The posterior distribution of β puts considerable probability on it being negative, although the third quartile of the posterior distribution is positive

```
print(post, pars = "p", include = FALSE)
```

```
## Inference for Stan model: linear.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##            mean se_mean   sd    2.5%     25%     50%     75%   97.5% n_eff Rhat
## alpha      0.13    0.00 0.09   -0.04    0.07    0.13    0.18    0.30  2978    1
## beta[1]   -0.29    0.02 1.21   -2.25   -1.25   -0.37    0.64    2.06  4035    1
## sigma      0.85    0.00 0.06    0.74    0.81    0.84    0.88    0.98  3552    1
## lp__    -124.54    0.03 1.30 -127.92 -125.17 -124.22 -123.57 -123.01  1942    1
##
## Samples were drawn using NUTS(diag_e) at Sat Apr 23 15:53:04 2022.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

If β were negative, that would be inconsistent with the CAPM (but perhaps would not entirely refute it) because it assumes that stocks on average move with "the market" and thus for an average stock, $\beta = 1$. However, the posterior distribution also puts some probability on the β for GameStop being near or greater than 1, so this posterior distribution alone does not provide decisive evidence on the CAPM (no individual stock, observed over a short period of time, would).

3.5 Posterior Predictive Distribution

```
alpha <- rstan::extract(post, pars = "alpha")[[1]]
beta  <- rstan::extract(post, pars = "beta")[[1]]
sigma <- rstan::extract(post, pars = "sigma")[[1]]
post_PD <- t(sapply(seq_along(alpha), FUN = function(s) {
  alpha[s] + beta[s] * (R_m - R_f) + R_f + rnorm(1, mean = 0, sd = sigma[s])
}))
low  <- apply(post_PD, MARGIN = 2, FUN = quantile, probs = 1 / 3)
high <- apply(post_PD, MARGIN = 2, FUN = quantile, probs = 2 / 3)
c(too_low = mean(R_i <= low),
  just_right = mean(R_i > low & R_i < high),
  too_high = mean(R_i > high))
```

```
##    too_low just_right   too_high
## 0.04210526 0.88421053 0.07368421
```

The fit of this model is poor in the sense that far too many observations are in the central interval when only $\frac{1}{3}$ should be. In essence, the model has overfit the data being conditioned on and is unlikely to provide good predictions of future data because the model is not extreme enough. This suggests that the CAPM may be inappropriate for GameStop and / or that the assumption of a normally distributed error term is inappropriate and an error distribution with heavier tails (such as a Student t) could be better.