Bayesian logistic regression and post-stratification for estimating the outcome of 2020 Presidential Election: Will Donald Trump win?

Ruoning Guo Student No.1004772114 Fanzhen Meng Student No.1002812824
Yifan Zhu Student No.1006345849 Kaiou Guo Student No.1004151865

11/02/2020

Model

Since we are interested in whether or not the voters prefer to vote for Trump, vote_trump can be considered as a response variable for these data and a logistic regression analysis could be carried out to determine the effect of gender and ange_group expected on the odds of voting for Trump. Under classic logistic regression, the estimator of β_1 would have a fixed value regardless of the value of predictors. However, in reality, people with different gender might have different β_1. This problem could be solved using the Bayesian approach, by incorporating our initial beliefs into the model (Kana, 2020). Therefore, Bayesian logistic regression model is the most appropriate model for this analysis.

Model Specifics

Fit two logistic regression models to these data, both with gender and age_group as predictor variables, where:

- "age group" is recorded as a categoricial variable which is classified as 4 categories "ages18to29", "ages30to44", "ages45to59" and "ages60plus".
- "gender" is a binary variable: female(0) and male(1).

The Baysian logistic model is:

$$log(p_i/(1-p_i)) = \beta_0 + \beta_1 X_{Male,i} + \beta_2 X_{18to29,i} + \beta_3 X_{30to44,i} + \beta_4 X_{45to59,i} + \beta_5 X_{60plus,i} + \epsilon$$

Where:

- X_{Male} = 1 if the ith voter is male, and 0 if female;
- X_{18to29} = 1 if the i-th voter's age between 18 to 29 years old, and 0 if not 18 to 29 years old;
- X_{30to44} = 1 if the i-th voter's age between 30 to 44 years old, and 0 if not 30 to 44 years old;
- X_{45to59} = 1 if the i-th voter's age between 45 to 59 years old, and 0 if not 45 to 59 years old;
- X_{60plus} = 1 if the i-th voter's age above 60 years old, and 0 if not above 60 years old;
- p_i is the probability of having a preference for voting Donald Trump.
- Coefficients β_1, β_2, β_3, β_4, β_5 represent change in log odds.For example, β_1 coefficient represents change in log odds for every one unit increase in X_{Male}.

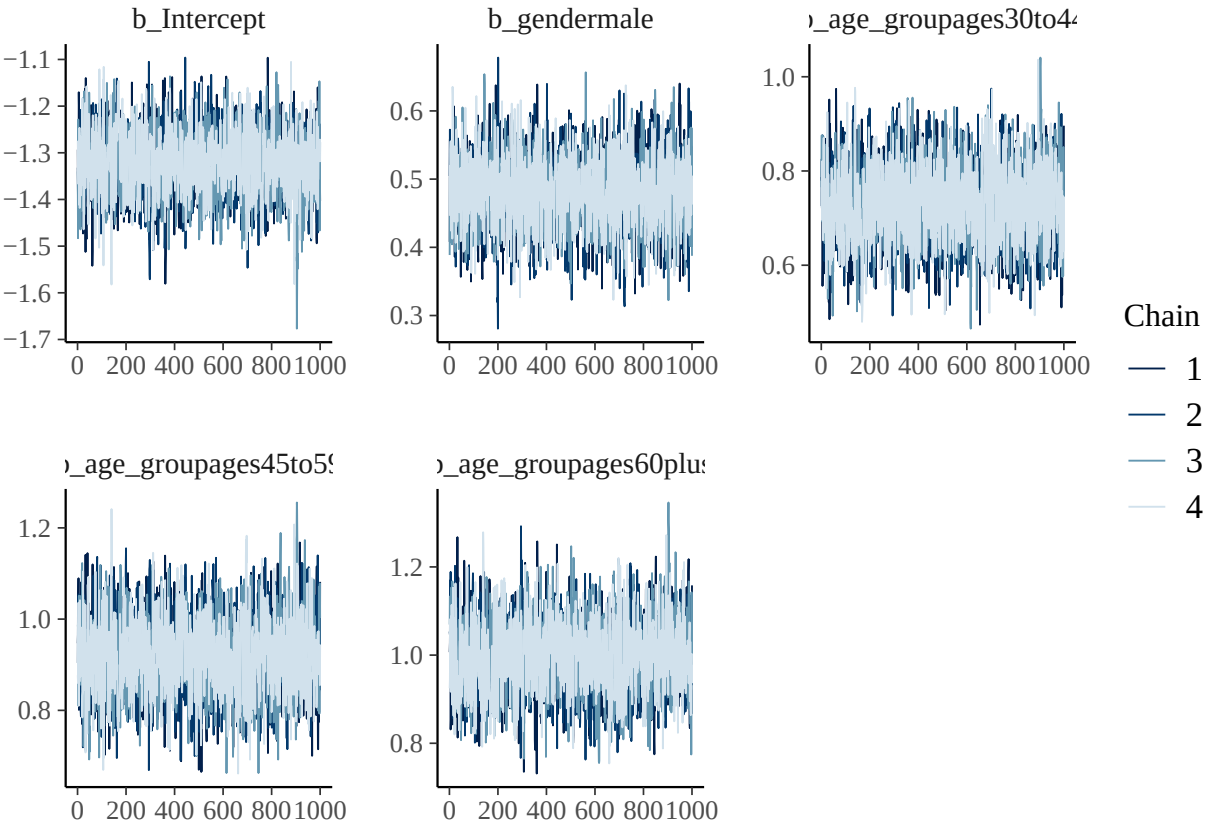
Post-Stratification

In survey sampling, stratification is often used to reduce the variance in order to obtain a more accurate estimate of the parameter of interest (Reilly et al, 2001). In the context of our study, the population can be put into different strata based

on their employment status: employed, unemployed, and not in the labor force. Since we know the size of each strata, if we are able to find a better estimate within each strata, then our estimation will be more accurate for the whole population (Reilly et al, 2001). Using the brm model described in the previous sub-section, the proportion of voters will be estimated in each age bin. Then each proportion estimate(within each bin) will be weight by the respective population size of that bin and sum those values and divide that by the entire population size. (Notice: Becasuse the raw dataset has population of more than 300w, we have only sampled 10000 observations, and eventually 9710 observations after dropping NA.)

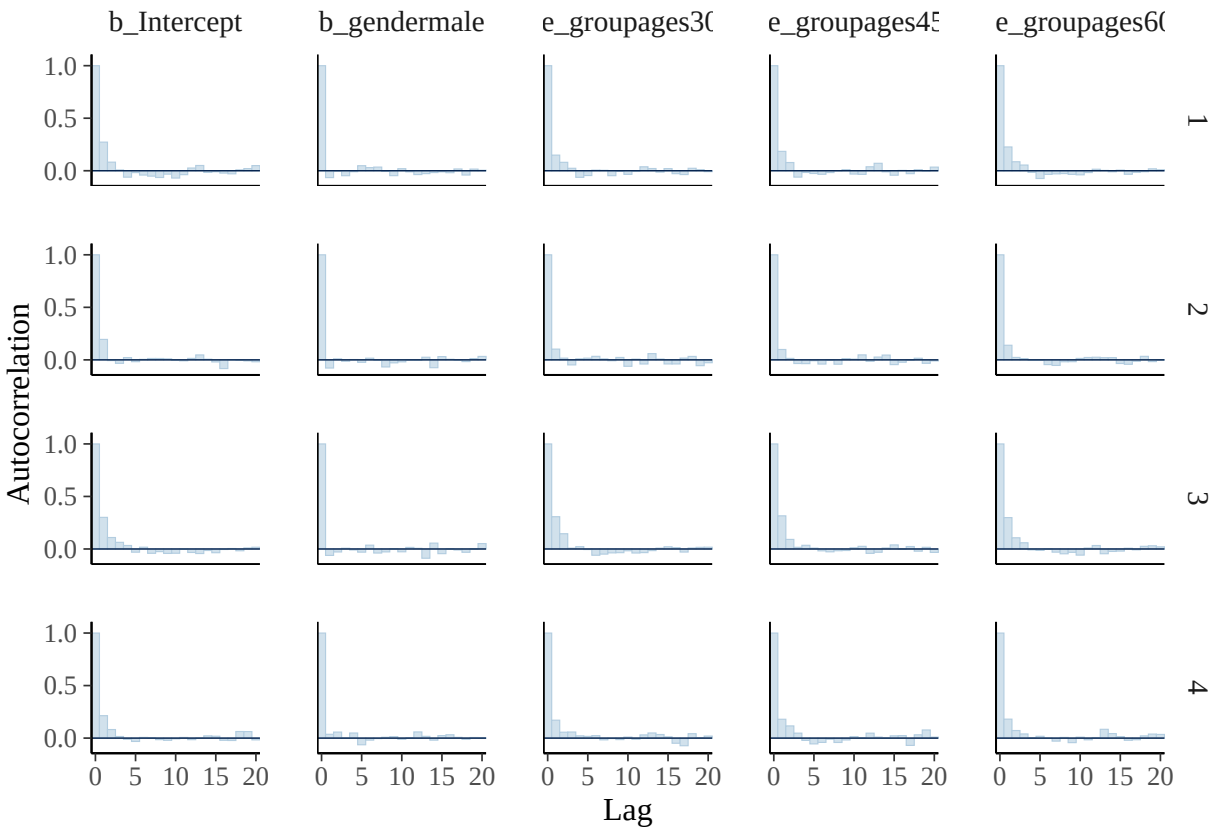
Results

Figure 1: plot model convergence to check whether there is evidence of non-convergence for the four chains.



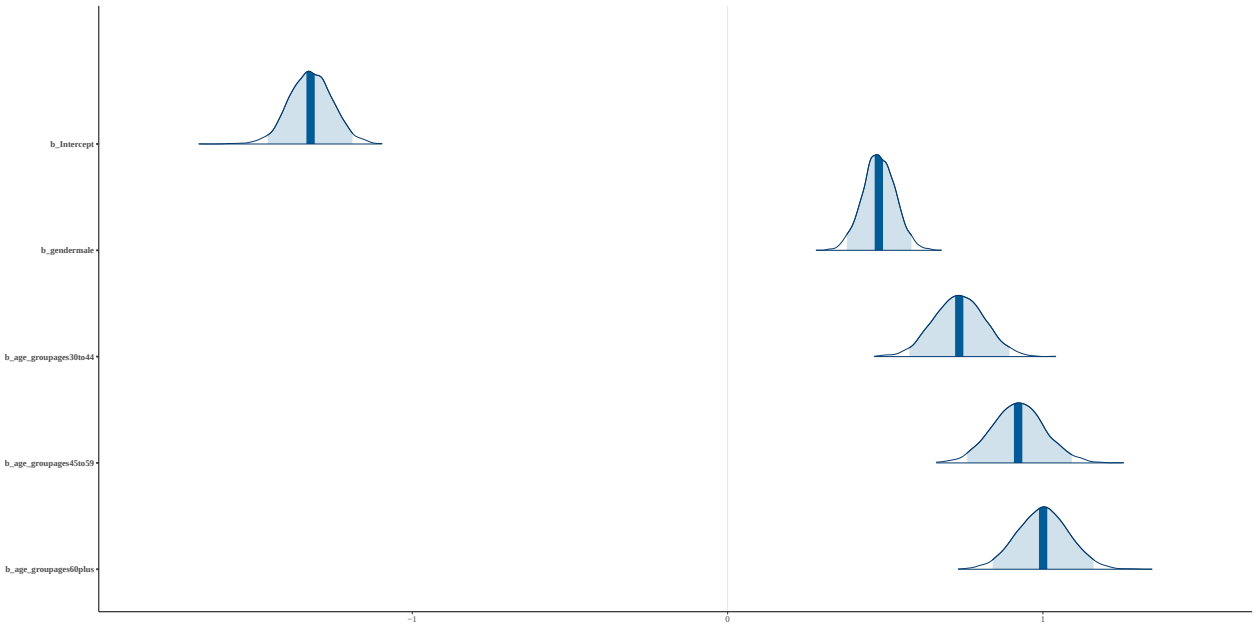
To check non-convergence, we use the stanplot function from the brms package to plot the caterpillar plot for each parameter of interest. From the series of plot above, we find the four chains mix well.

Figure 2: check autocorrelation due to the fact that the presence of strong autocorrelation would bias variance estimates.



The plots above show that the autocorrelation parameters all diminish to around zero.

Figure 3: visualise the point estimates and their associated 95%-CI



In the Bayesian model, the 95% confidence interval states that there is 95% chance that the true value falls within this interval. Since the 95%-CI does not contain zero, the respective model parameters are likely meaningful.

Table 1: Model summary

```
summary(modelbrm)
```

```
##  Family: bernoulli
##   Links: mu = logit
## Formula: vote_trump ~ gender + age_group
##    Data: survey_data (Number of observations: 6452)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Population-Level Effects:
##                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept              -1.32      0.07    -1.46    -1.19 1.00     2473     2344
## gendermale              0.48      0.05     0.38     0.58 1.00     4331     2440
## age_groupages30to44     0.73      0.08     0.58     0.89 1.00     2632     2542
## age_groupages45to59     0.92      0.08     0.76     1.09 1.00     2677     2387
## age_groupages60plus     1.00      0.08     0.84     1.16 1.00     2443     2397
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

The output shows above is our partial regression output from the Bayesian model, where R uses "genderfemale" and "age 19to29" as the level of the reference. Now we have the "estimates" values which are equivalent to beta coefficients in the logistic formula. A rough probability of voting for trump can be calculated based on the voter's gender and age_group.

```
# An estimate for each employment status based on their demographic features.
head(usa_data)
```

```
## # A tibble: 6 x 5
## # Groups:   gender, age_group [2]
##   gender age_group  empstat               n cell_prop_of_division_total
##   <chr>  <chr>      <chr>             <dbl>                       <dbl>
## 1 female ages18to29 employed            586                      0.104
## 2 female ages18to29 not in labor force  254                      0.0672
## 3 female ages18to29 unemployed           38                      0.138
## 4 female ages30to44 employed            811                      0.143
## 5 female ages30to44 not in labor force  229                      0.0606
## 6 female ages30to44 unemployed           35                      0.127
```

Table 2: Estimates in terms of emplyment status

```
post_stratified_estimates
```

```
## # A tibble: 3 x 4
##   empstat             mean  lower upper
##   <chr>              <dbl>  <dbl> <dbl>
## 1 employed           0.410 0.0891 0.748
## 2 not in labor force 0.433 0.0365 0.856
## 3 unemployed         0.382 0.0543 0.761
```

We estimate that the proportion of voters whose not in the labor force in favour of voting for trump to be 0.432, which is the highest; Employed voters in favour of voting for trump to be 0.410; Unemployed voters in favour of voting for trump to be 0.379, which is the lowest. This is based off our post-stratification analysis of the proportion of voters in favour of voting for trump modeled by a Bayesian logistic regression model, which accounted for age_group and gender.

We got the $\hat{y}$, now calculate for $\hat{y}^{PS}$.

```
#Number of "employment": 5655
#Number of "not in labour force": 3779
#Number of "unemployment": 276

(0.4103455*5655+0.4321664*3779+0.3787699*276)/(5655+3779+276)
```

```
## [1] 0.4179404
```

Discussion

Checking the evidence of non-convergence and autocorrelation for the four chains(age_group).

- Figure 1: there seems no evidence of non-convergence since the four chains mix well.
- Figure 2: As the autocorrelation parameters all diminish to around zero, there is no evidence of autocorrelation for all model variables in our four chains.
- Figure 3: Gender and age group are meaningful predictors as their 95% confidence intervals do not contain zero. In addition, gender is a more credible predictor in the constructed model since its density plot is relatively narrower.

Thus, we can safely illustrate the model output.

Model summary

In survey data, the population of male voters is almost consistent with female voters(respectively 3152 and 3300), which also matches with the current population structure in the U.S. (World Population Review, 2020). Since CIs for all Population-Levels do not contain zero(means p-value<0.05), there is strong statistical evidence to show that preference for voting Trump is associated with both gender and age_group of the voters. To be more specific, for males, 47.18%(1487/3152) voting for Trump while for females, only 34.48%(1138/3300) voting for Trump. Also, as the age group increases, the coefficient tends to increase on average. In other words, if holding everything else constant, turnout voters who belong to the higher age group are more likely to vote for Trump.

Relating model with the world

Probabilistic estimates for each strat are 0.410, 0.432, 0.379 for three employment statuses (employed, not in the labour force, and unemployed respectively table.3). It shows slight differences between strats while people not in the labour force are more likely to vote for Trump as re-elected president than those who are employed or unemployed. One thing that might explain this result is the policy that two parties claimed to the public. For example, House Democrats passed a 2.2 trillion coronavirus bill in early October that contained another round of direct payments, federal unemployment benefits,

and small-business aid (Zeballos-Roig,2020) to replace the $600-a-week benefit that expired in July (Buchwald, 2020). Especially for the people who are not in the workforce, for workers without pensions or 401(k)s, the "Biden Plan" pledges near-universal access to "an 'automatic 401(k),' which provides the opportunity to easily save for retirement at work. (Forbes, 2020) It is designed to put millions of middle-class families on the path to a secure retirement. On the contrary, Trump's administration published a proposed rule in September 2020 that would stop proxy votes in favor of social or political positions that don't advance the financial interests of retirement plan participants (Miller, 2020).

Eventually, we forecast that Trump has a 41.6% chance to win the 2020 United States presidential election based off our post-stratification estimates weighted by the three strats(calculation can be found at the end of the appendix).

Weaknesses

Since our research model contains only three variables, including two variables (gender and age groups) in the primary model and one variable(employment status) in the stratification section. One drawback might be the lack of cells since there are only 24 cells. We probably need to add more categories or variables to see whether the brm result may change or not. Besides, the allocation of sample units to strata cannot be controlled since the variation of sample sizes according to strata could cause too few samples to categorize a reliable stratum mean and standard error (Westfall, Patterson& Coulston, 2011).

Another drawback of a lack of variables in the model is "OVB" which is a bias that stems from the absence of relevant variables in a model (S, 2020). To solve this, we need to add more variables and test each of their robustness to eliminate an underfit model. We might also need to compare with the actual election results and do a post-hoc analysis (or at least a survey) of how to better improve estimation in future elections, because of the nature of the large uncertainty of our post stratified estimates. In addition, the drawback of surveys and census themselves may exist.

Next Steps

The next steps of our research may contain:

- adding layers to do a Multilevel Regression Poststratification, therefore to achieve the adjustment for selection bias
- correcting for imbalances in sample composition, even when the dataset involves a large number of variables– together with multilevel regression (Si, 2020).

References

Buchwald, E. (2020, September 30). Democrats are pushing for another round of stimulus checks and $600 in weekly unemployment benefits. Retrieved from https://www.marketwatch.com/story/democrats-are-making-a-renewed-push-to-bring-back-the-600-unemployment-benefit-11601412769?mod=the-ratings-game

Downes, M., & Carlin, J. (2020). Multilevel regression and poststratification for estimating population quantities from large health studies: A simulation study based on US population structure. Journal of Epidemiology and Community Health. doi:10.1136/jech-2020-214346

GENDER DIFFERENCES IN VOTER TURNOUT - CAWP. (n.d.). Retrieved from https://cawp.rutgers.edu/sites/default/files/resources/genderdiff.pdf

Ghilarducci, T. (2020, August 17). How Democrats And Republicans Would Tackle The Enduring Retirement Security Crisis. Retrieved from https://www.forbes.com/sites/teresaghilarducci/2020/08/14/how-democrats-and-republicans-would-tackle-the-enduring-retirement-security-crisis/?sh=6959f7d2342f

Si, Y. (2020). Multilevel Regression and Poststratification. Retrieved from http://www-personal.umich.edu/~yajuan/files/MrPbrownbag-YAJUANSI.pdf

Kana, M. (2020, February 21). Introduction to Bayesian Logistic Regression. Retrieved from https://towardsdatascience.com/introduction-to-bayesian-logistic-regression-7e39a0bae691

Reilly, C., Gelman, A., & Katz, J. (2001). Poststratification Without Population Level Information on the Poststratifying Variable With Application to Political Polling. Journal of the American Statistical Association, 96(453), 1-11. doi:10.1198/016214501750332640

Alexander, R. (2019, December 03). Getting started with MRP. Retrieved from https://rohanalexander.com/posts/2019-12-04-getting_started_with_mrp/

S, T. (2020, May 06). What is statistical bias and why is it so important in data science? Retrieved from https://towardsdatascience.com/what-is-statistical-bias-and-why-is-it-so-important-in-data-science-80e02bf7a88d

Miller, S. (2020, November 02). DOL Proposes Limits on Proxy Voting by Retirement Plan Fiduciaries. Retrieved from https://www.shrm.org/ResourcesAndTools/hr-topics/benefits/pages/dol-proposes-limits-on-proxy-voting-by-plan-fiduciaries.aspx

(this is the source of database) IPUMS. (n.d.). U.S. CENSUS DATA FOR SOCIAL, ECONOMIC, AND HEALTH RESEARCH. Retrieved from https://usa.ipums.org/usa/

United States Population 2020 (Live). (n.d.). Retrieved from https://worldpopulationreview.com/countries/united-states-population

Westfall, J. A., Patterson, P. L., & Coulston, J. W. (2011). Post-stratified estimation: Within-strata and total sample size recommendations. Canadian Journal of Forest Research, 41(5), 1130-1139. doi:10.1139/x11-031

Zeballos-Roig, J. (2020, October 01). House Democrats pass their $2.2 trillion stimulus plan, which includes $600 federal unemployment benefits and direct payments. Retrieved from https://www.businessinsider.com/house-democrats-stimulus-plan-congress-unemployment-economy-pass-vote-bill-2020-9

Appendix

```
table(survey_data$gender, survey_data$vote_trump)
```

```
##
##             0    1
##   female 2162 1138
##   male   1665 1487
```

```
table(survey_data$age_group, survey_data$vote_trump)
```

```
##
##                 0    1
##   ages18to29 1044  339
##   ages30to44 1153  846
##   ages45to59  790  665
##   ages60plus  840  775
```

```
library(srvyr)
data(api, package="survey")
# compute for each cell
by_vs_am <- usa_sample %>%
            group_by(gender, age_group,empstat)
by_vs <- by_vs_am %>% summarise(n = n())
by_vs %>% summarise(n = sum(n))

# To removing grouping, use ungroup
```

```r
by_vs %>%
  ungroup() %>%
  summarise(n = sum(n))

# compute population in 3 employment status
library(data.table)
as.data.table(by_vs)[, sum(n), by = empstat]

empN<- by_vs$empstat
empN <- as.data.frame(empN)
as.data.table(by_vs)[,sum(n),by=empstat]
empN$empN[empN$empN %in% "employed"] <- "5655"
empN$empN[empN$empN %in% "not in labor force"] <- "3779"
empN$empN[empN$empN %in% "unemployed"] <- "276"
N <- cbind(by_vs,empN)

N$n = as.numeric(as.character(N$n))
N$empN = as.numeric(as.character(N$empN))

cell_prop_of_division_total <-N$n/N$empN
cell_prop <-data.frame(cell_prop_of_division_total)
usa_data <- cbind(N,cell_prop)

usa_data <- subset(usa_data, select = -5 )
```

```r
head(usa_data)
```

```
## # A tibble: 6 x 5
## # Groups:   gender, age_group [2]
##   gender age_group  empstat               n cell_prop_of_division_total
##   <chr>  <chr>      <chr>             <dbl>                       <dbl>
## 1 female ages18to29 employed            586                      0.104
## 2 female ages18to29 not in labor force  254                      0.0672
## 3 female ages18to29 unemployed           38                      0.138
## 4 female ages30to44 employed            811                      0.143
## 5 female ages30to44 not in labor force  229                      0.0606
## 6 female ages30to44 unemployed           35                      0.127
```

```r
##Post-stratification
set.seed(2020)
# We're just going to do some rough forecasts. For each gender and age_group we want the relevant coeff
post_stratified_estimates <-
  modelbrm %>%
  tidybayes::add_predicted_draws(newdata = usa_data) %>%
  rename(trump_predict = .prediction) %>%
  mutate(trump_predict_prop = trump_predict*cell_prop_of_division_total) %>%
  group_by(empstat,.draw) %>%
  summarise(trump_predict = sum(trump_predict_prop)) %>%
  group_by(empstat) %>%
  summarise(mean = mean(trump_predict),
            lower = quantile(trump_predict, 0.025),
            upper = quantile(trump_predict, 0.975))
```

```
post_stratified_estimates
```

```
## # A tibble: 3 x 4
##   empstat             mean  lower upper
##   <chr>              <dbl>  <dbl> <dbl>
## 1 employed           0.410 0.0891 0.748
## 2 not in labor force 0.433 0.0365 0.856
## 3 unemployed         0.382 0.0543 0.761
```

```
count(usa_sample$empstat)
```

```
##                    x freq
## 1           employed 5655
## 2 not in labor force 3779
## 3         unemployed  276
```

```
(0.4103455*5655+0.4321664*3779+0.3787699    *276)/(5655+3779+276)
```

```
## [1] 0.4179404
```

Link to the associated GitHub repo: https://github.com/YvonneYifanZhu/STA304_PS3_Submission