GR5065 HW4

1 Wages

We are going to consider an individual's (not their household's) income from wages and salary in the 2021 American Community Survey (ACS), which is reasonably representative.

You should choose a state that is smaller than New York to download the data for, using the following pattern but changing the `FILE` to be "csv_p*.zip" where the asterisk is the two-letter postal abbreviation for that state (in lower case). See the vector `state.abb` in R if you are unsure of the postal codes.

```
library(dplyr)
if (!any(list.files() == "ACS.zip")) {
  ROOT <- "https://www2.census.gov/programs-surveys/acs/data/pums/2021/1-Year/"
  FILE <- "csv_pny.zip" # change this to a state that is not ny
  download.file(paste0(ROOT, FILE), destfile = "ACS.zip")
  unzip("ACS.zip")
}
ACS <- readr::read_csv(list.files(pattern = "csv$"), show_col_types = FALSE) %>%
  select(RT:WAOB) %>%
  filter(WAGP > 0) %>%
  mutate(censored = WAGP == max(WAGP))
```

The variable in question is called `WAGP` and we have only kept observations for people who have a paying job. As can be seen, the wage data are "top-coded", which is to say that a small percentage of the largest values have been recorded as some state-specific number, c,

```
summarize(ACS, c = max(WAGP), n = sum(censored)) # your state will be somewhat different
```

```
# A tibble: 1 x 2
      c     n
  <dbl> <int>
```

```
1 634000  1219
```

that is less than their actual wages to prevent them from being individually identifiable.

It is commonplace in economics to model the *logarithm* of a worker's wages, which is thought to be conditionally normal given the worker's characteristics (but without censoring). You need to construct a plausible model for log-wages in your chosen state that takes into account the fact that the observed data are top-coded at $\ln c$. To do so, you will need to read through the data dictionary to see how the potential predictors are coded, but keep in mind that these data do not include any household-level variables, only person-level variables (starting with `AGEP` and excluding any "allocation flags" or "person weights").

1.1 Generative Model

Use your chosen predictors in a plausible generative model where log-wages have Gaussian errors (with unknown standard deviation σ). From there, you can transform the generated log-wages to wages (in units of dollars) and apply the censoring at whatever c is for your state. Make sure to center the predictors you use, after doing any transformations to them (such as creating dummy variables from a factor).

Use `geom_density` in the ggplot2 package to plot the prior predictive distribution of wages up to c. In what sense is the plot of your prior predictive beliefs about wages reasonable? How would you respond if someone pointed out that this distribution does not look Gaussian?

1.2 Posterior Distribution

Use the `brm` function in the brms package to draw from the posterior distribution of the parameters in your generative model above for log-wages using a formula that starts like `brm(log(WAGP) | cens(censored) ~ .`

Use the `hypothesis` function to evaluate one substantive hypothesis that you might have about one or more of the parameters in your model.

1.3 Pareto k Values

Which, if any, observations in the data (after dropping any observations that are missing on variables that are included in the model, so it is better to look at the result of `model.frame` being applied to object produced by `brm`) have values that are problematic for estimating the ELPD with the plain PSISLOOCV estimator? What characteristics do these observations seem to have in common?

Call `loo` again with either `moment_match = TRUE` or `reloo = TRUE` to deal with any problematic observations and specify `save_psis = TRUE` to retain the new Pareto k estimates.

1.4 Posterior Predictive Checks

Evaluate how well your model fits via some plots that use the leave-one-out concept (but are estimated with Pareto Smoothed Importance Sampling). Provide a brief explanation.

1.5 Posterior Predictive Distribution

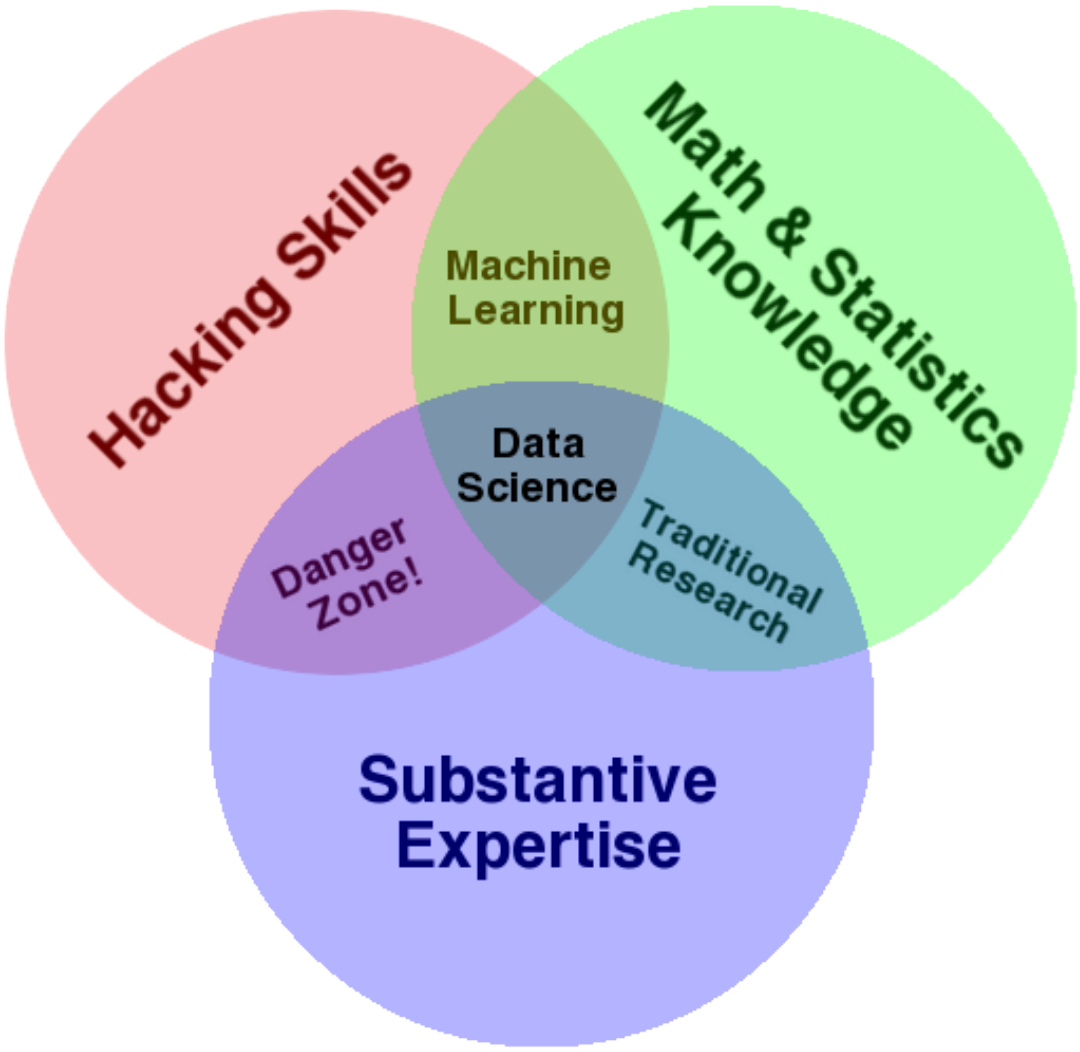
Use the `posterior_predict` function to draw from the posterior predictive distribution of log-wages (with error but without censoring) and then apply the `exp` function to obtain the posterior predictive distribution of wages (in dollar units). This will be a $R \times N$ matrix and you should apply then apply `rowMeans` function to it to obtain a vector of size R with draws of your beliefs about average wages in your state. You may need to specify the `ndraws` argument to `posterior_predict` to be some value less than 4000 in order to not run out of RAM.

Use the `geom_density` function in the ggplot2 package to plot the distribution of your beliefs about average wages. Use the `geom_vline` function to add a vertical line at `mean(ACS$WAGP)`. Describe why the average of `WAGP` is an underestimate of average wages and to what degree?

2 Culture

Read Leo Breiman's (2001a) article, along with the short responses by Cox, Efron, Hoadley, Parzen, and the rejoinder to those responses by Breiman (2001b) at the end of the PDF. Write an essay that addresses at least the following questions:

1. Breiman draws contrasts between the "two cultures in the use of statistical modeling to reach conclusions from data". Are Bayesians a third culture, a subset of one of the two cultures Breiman mentions, an intersection of the two cultures, or something different?
2. How would you characterize these two cultures with regard to the four or five sources of uncertainty that were mentioned on the first day of GR5065?
3. In what ways does the Expected Log Predictive Density (ELDP) that is estimated by `loo` (albeit 15 years after Breiman's paper) fit into the discussion?
4. Take one of the arguments made by Cox, Efron, Hoadley, or Parzen that you think is important but not addressed in enough depth by Breiman's (2001b) rejoinder. How do you think Breiman should have responded?
5. One of the two cultures has definitely become more influential in the twenty years since Breiman's (2001a) paper. Read also this blog post by Drew Conway that includes this Venn Diagram for Data Science:



Where do you think each of the two cultures that Breiman's describes lie within Conway's Venn diagram?

6. If we were to equate or specialize Math & Statistics Knowledge to "probability", Substantive Expertise with the ability to specify "priors" on the parameters (as well as the other parts of a generative model), and Hacking Skills with "use of MCMC algorithms", then the intersection would characterize Bayesian analysis. And yet, Bayesian analysis is not taught very much in Data Science programs or used very much by their graduates, despite the fact that Bayesian analysis started to become practical in 1990 and Data Science programs started to pop up around 2010. Why do you think that is?