Project Proposal

Ranen Oomen-Danckert

15/03/2024

For my project, which I will be working on individually, I would like to compare Bayesian regression or Classification and Frequentist Regression or Classification on a wine quality data set. My Github repo is here https://github.com/rdanckert/STAT447-Project. The first potential dataset is for Spanish wines from Kaggle, it has a response variable which is a rating from 4 to 5, and potential predictors acidity, body, grape varietal, price, region, and number of reviews.

```
spain_wine_ratings <- read_csv("C:/Users/ranen/Downloads/wines_SPA.csv")
```

```
head(spain_wine_ratings, n = 3)
```

```
## # A tibble: 3 x 11
##   winery wine  year  rating num_reviews country region price type   body acidity
##   <chr>  <chr> <chr> <dbl>       <dbl> <chr>   <chr>  <dbl> <chr> <dbl>   <dbl>
## 1 Teso ~ Tinto 2013    4.9          58 Espana  Toro     995 Toro~     5       3
## 2 Artadi Vina~ 2018    4.9          31 Espana  Vino ~   314. Temp~     4       2
## 3 Vega ~ Unico 2009    4.8        1793 Espana  Riber~   325. Ribe~     5       3
```

The second potential dataset is also from Kaggle, this dataset doesn't restrict its wines to spain and has more scientific potential predictor variables, like residual sugar and pH level.

```
wine_quality <- read_csv("C:/Users/ranen/Downloads/WineQT.csv")
```

```
head(wine_quality, n = 3)
```

```
## # A tibble: 3 x 13
##   `fixed acidity` `volatile acidity` `citric acid` `residual sugar` chlorides
##             <dbl>              <dbl>         <dbl>            <dbl>     <dbl>
## 1             7.4                0.7             0              1.9     0.076
## 2             7.8               0.88             0              2.6     0.098
## 3             7.8               0.76          0.04              2.3     0.092
## # i 8 more variables: `free sulfur dioxide` <dbl>,
## #   `total sulfur dioxide` <dbl>, density <dbl>, pH <dbl>, sulphates <dbl>,
## #   alcohol <dbl>, quality <dbl>, Id <dbl>
```

This project interests me because I want to improve my understanding of what makes different wines higher quality. I could compare Lasso regularization or variable selection methods and a Bayesian approach to determine if predictor coefficients are non-zero.