Main Assignment

Mukesh Arambakam - 19301497

Introduction

The Wine Dataset taken from Kaggle (thoutt, 2018) contains about 130k (to be exact 129971) observations and 14 variables namely: country, description, designation, points, price, province, region_1, region_2, taster name, taster twitter handle, title, variety, winery. These columns provide information regarding *country* that the wine is from, *description* of the wine, *designation* of the vineyard within the winery, the number of *points* the wine enthusiast rated it on a scale of 1-100, *price* of a bottle of wine, the location details of where the wine is produced from like *province, region_1* and *region_2*, And also details like *taster name and taster twitter handle* about the wine enthusiast who tasted the wine and rated it. And more details about the wine in itself like the *title* of the wine that was reviewed, *variety* of grapes used and also the *winery* that made the it. The dataset in general describes about wines through blind tasting like a master sommelier would. But to answer the question at hand we will need to preprocess and analyse the data to obtain the required useful information from it.

Problem
1 (a). My wife likes Sauvignon Blanc from South Africa. My mother-inlaw likes Chardonnay from Chile. Both agree that $15 is the right amount to spend on a bottle of wine.
(i) Which type of wine is better rated? How much better?
(ii) Suppose I buy a South African Sauvignon Blanc and a Chilean Chardonnay, both priced $15. What is the probability that the Sauvignon Blanc will be better?

Data Handling for Q1 (a)

To answer Question 1 (a), a little bit of preprocessing is required, which involves removing unwanted columns like: description, designation, province, region_1, region_2, taster name, taster twitter handle and winery. Which leaves the columns: country, points, price, and variety. Since the we are required to work particular types of wine, a filter on the data is applied to obtain the same, namely: Sauvignon Blanc from South Africa and Chardonnay from Chile both priced exactly at $15.

```
wine_df = within(wine_df, remove('X', 'description', 'designation',
    'taster_name', 'taster_twitter_handle', 'title'))

wine_df = within(wine_df, remove('province', 'region_1', 'region_2',
    'winery'))

# applies filter for Sauvignon Blanc Wine from South Africa, Priced at $15
wine_sa_sb_15 = filter(wine_df, variety=='Sauvignon Blanc' & country ==
    'South Africa' & price == 15)

# applies filter for Chardonnay Wine from Chile, Priced at $15
wine_ch_ch_15 = filter(wine_df, variety=='Chardonnay' & country == 'Chile'
    & price == 15)
```

Analysis for Q1 (a)

The Figure 1 shows a Jitter Plot of the points from a sample of wines which the wine enthusiast rated of two different varieties from different countries. Thirty-sevens wine samples from Chardonnay Wine from Chile, Priced at $15 (from here on simply referred to as "Chardonnay Wine" unless explicitly stated) and 14 samples from Sauvignon Blanc Wine from South Africa, Priced at $15 (from here on simply referred to as "Sauvignon Blanc" unless explicitly stated) were randomly selected for for this analysis in the dataset.

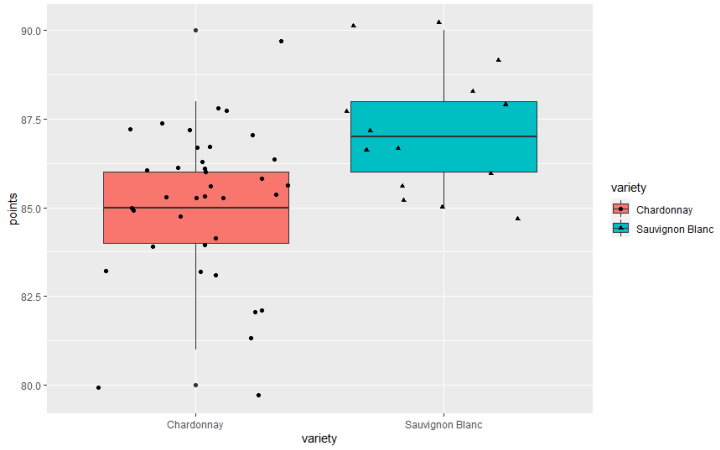


Figure 1: Wine points compare

Analysis of Q1 (a) i

From this we can observe that per group means for Chardonnay Wine ($\overline{y_1}$) is 85.05 and for Sauvignon Blanc Wine ($\overline{y_2}$) is 87.21. On performing **t-test** on the same we obtain a p-value of 0.00203 which is less than 0.05 which indicates that θ_1 and θ_2 are different. With this information we can also answer the question which wine is rated better and by how much (**1.a.i**), for which we can clearly say that **Sauvignon Blanc Wine from South Africa** is rated better by an average of 2.13321 points.

Analysis of Q1 (a) ii

Gibbs sampler is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations which are approximately from a specified multivariate probability distribution, when direct sampling is difficult. It is commonly used as a means of statistical inference. In this case we use this to compare means of different wine varieties. Considering the entire dataset the points wine enthusiast rated it, the mean of points for the complete dataset is 88.447414 and a standard deviation is 3.03973. Reasonable prior parameters can be based on this information. μ and σ^2, we can take μ_0 = 88.45 and σ_0^2 = 3.04^2, although these values might be an overestimate (considering the entire dataset), but we can consider them as prior knowledge.

On constructing a Gibbs sampler to approximate the posterior distribution of the wines p($\mu, \delta, \sigma^2|y_1, y_2$), with the R-Code provided in the Hierarchical Models Case-Study[1] and making slight changes to the variables related to the prior knowledge and plotting the results obtained using *as.mcmc(fit)* we get Figure 2a. and for the output of the *Auto-Correlation Function* we get Figure 2b. And the results for Raftery And Lewis's Diagnostic are as follows:

```
> raftery.diag(as.mcmc(fit))
```

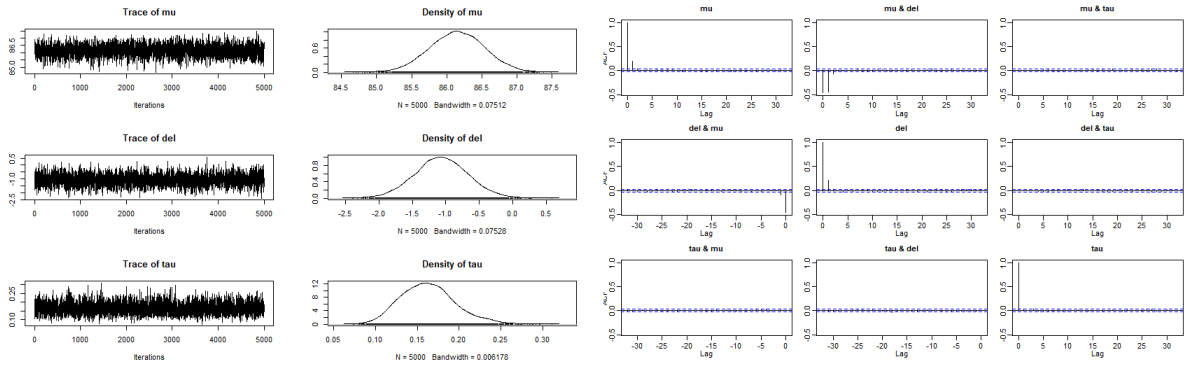
[1]https://www.scss.tcd.ie/ arwhite/Teaching/CS7DS3/Hierarchical_Models_case_study.html

```
Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

    Burn-in Total Lower bound Dependence
     (M)    (N)  (Nmin)     factor (I)
 mu  3      4198 3746       1.12
 del 3      4267 3746       1.14
 tau 2      3803 3746       1.02
```

The mean and standard deviation of the MCMC model are fellows:

```
> apply(fit, 2, mean)
       mu       del       tau
86.2359814 -1.0639598 0.1634477
> apply(fit, 2, sd)
      mu      del      tau
0.3852922 0.3820694 0.0316206
```



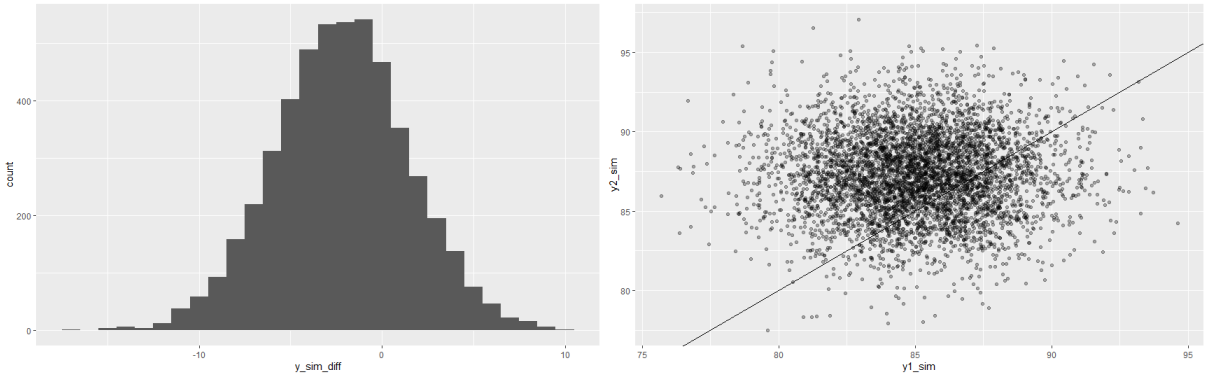
(a) MCMC Fit

(b) Auto-Correlation Function

Figure 2: Gibbs Sampler Output

Using the model's output from the Gibbs Sampler, we can make predictions about unobserved data, like say "I buy a South African Sauvignon Blanc and a Chilean Chardonnay, both priced $15. What is the probability that the Sauvignon Blanc will be better?" (like answer Question **1.a.ii**). This can be done by **BLAH BLAH BLAH**. For this the output is as below. And Figure 3 represents graphs of the simulated data using the results from the Gibbs sampler.

```
> y1_sim <- rnorm(5000, mean = fit[, 1] + fit[, 2], sd = 1/sqrt(fit[, 3]))
> y2_sim <- rnorm(5000, mean = fit[, 1] - fit[, 2], sd = 1/sqrt(fit[, 3]))
> mean(y2_sim > y1_sim)
[1] 0.7294
```



(a) Point Difference vs Count (b) Scatter Plot

Figure 3: Graphs of Simulated Data

From the results above we can see that when random wine samples selected from Chardonnay (*y1_sim*) and Sauvignon Blanc (*y2_sim*) the probability that **Sauvignon Blanc** is better is almost 73%.

Problem
1 (b). Consider the Italian wines in the dataset. Which regions produce better than average wine? Limit your analysis to wines costing less than $20 and to regions which have at least four such reviews.

Data Analysis for Q1 (b)

Here, the questions requires us to compare Wines from various regions in Italy, and build a model where we would be able to tell which type(s) of wines are better than average. To filter the data as per the question's requirement, we can achieve this by using *sqldf* and *dplyr* libraries as follows:

```
# Italian Wine, Priced less than $20 with regions which have at least 4
    reviews
wine_it_lt20 = filter(wine_df, country == 'Italy' & price < 20 & region_1
    != "" & price != 0 & price != "")
wine_regs <- sqldf("SELECT region_1, count(*) FROM wine_it_lt20 GROUP BY
    region_1 HAVING count(*) >= 4")$region_1
wine_it_lt20 = wine_it_lt20[wine_it_lt20$region_1 %in% wine_regs, ]
wine_it_lt20 = na.omit(wine_it_lt20)
```

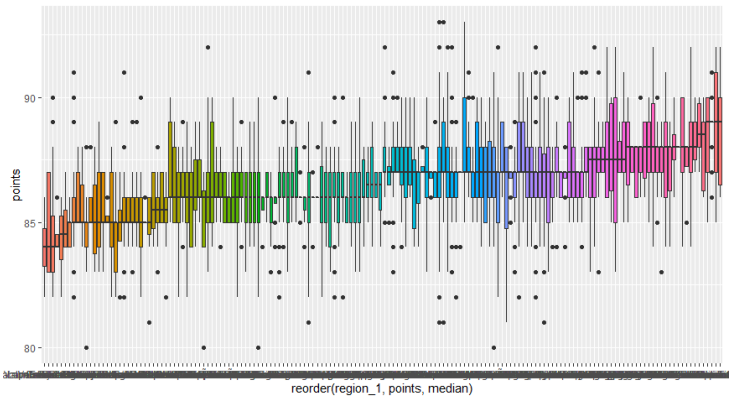


Figure 4: Jitter plot of regions

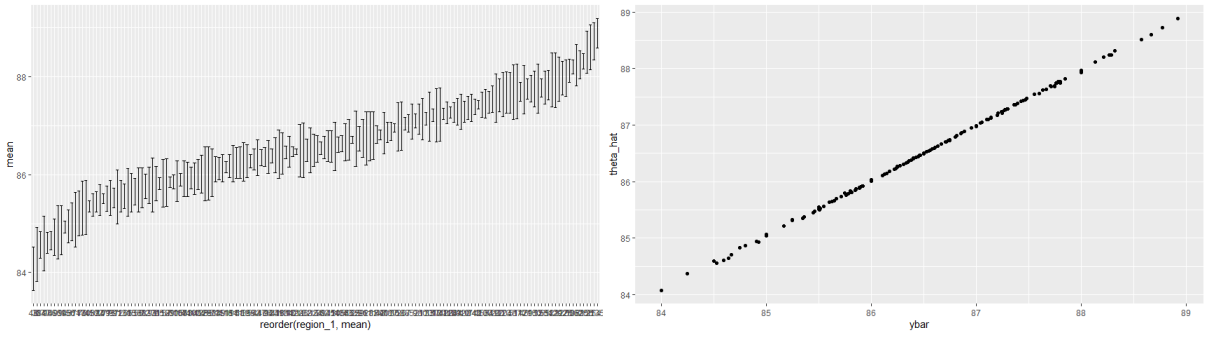
On filtering the data as above we get about 4700 records from 162 regions in Italy that match this criteria and plotting a Jitter plot for the regions, the results are as seen in Figure 4.

Analysis for Q1 (b)

The average of this data is 86.58. On applying the *compare_m_gibbs* function from the Hierarchical Models Case-Study we get vital information about the wines like there total predicted mean and standard deviation which are are follows 86.55 and 3.06 respectively. We can also gain information about how the wine's points vary within (between) each region (groups), the mean value of this variation is 0.10 and the standard deviation is 0.21. And using the output from the results of the code as below we see that there are **77 regions** among the 162 regions in Italy which produce wine which is better than average.

```
 ## get basic posterior summary
theta_hat <- apply(fit2$theta, 2, mean)
## keep track of different regions
names(theta_hat) <- wine_regs
## which regions did best and worst?
sort(theta_hat, decreasing = TRUE)
## and displaying regions which produced better wine than average.
abv_avg <- theta_hat[sort(theta_hat, decreasing = TRUE) >
    mean(wine_it_lt20$points)]
## to give a count of all the regions which produce wine above average
sum(complete.cases(abv_avg))
```

We can also obtain information like how the wines are spread within each region by drawing box plots with there 2.5% and 97.5% quarterlies and ignoring the outliers, which will give us and idea about the treat of wine's points with respect to each region and but sorting these values we can clearly see how some regions produce wines better than other regions. See Figure 5a. Apart from this information we can also see how our Gibbs Sampler performs (theta_hat) when compared to the true value (y_bar), see Figure 5b, the almost straight formed by plotting the two values indicates the performance of the Gibbs Samplers when compared to the actual data, this represents the accuracy of our prediction.



(a) Region Box Plot

(b) theta_hat vs y_bar

Figure 5: Graphs of Simulated Data

References

thoutt, z. (2018). Wine reviews. *Kaggle.Com, https://www.kaggle.com/zynicide/wine-reviews.*.