Thesis Simulation Results Analysis for Chapter 4

Dasha Asienga

2024-04-01

Contents

This file is intended to synthesize and analyze the results from the simulation.

Reading in the Result Data Sets

Logistic Regression

The results data set has 200 observations for each of the simulation trials, 50 from each sample size: $n = 500, 1000, 2500, 5000$.

```
lr_500 <- read.csv("/home/dasienga24/Statistics-Senior-Honors-Thesis/R/Simulation/LogisticRegression/Re
lr_1000 <- read.csv("/home/dasienga24/Statistics-Senior-Honors-Thesis/R/Simulation/LogisticRegression/R
lr_2500 <- read.csv("/home/dasienga24/Statistics-Senior-Honors-Thesis/R/Simulation/LogisticRegression/R
lr_5000 <- read.csv("/home/dasienga24/Statistics-Senior-Honors-Thesis/R/Simulation/LogisticRegression/R

lr_500 <- lr_500 |>
  mutate(sample_size = 500) |>
  dplyr::select(-X)

lr_1000 <- lr_1000 |>
  mutate(sample_size = 1000) |>
  dplyr::select(-X)
```

```r
lr_2500 <- lr_2500 |>
  mutate(sample_size = 2500) |>
  dplyr::select(-X)

lr_5000 <- lr_5000 |>
  mutate(sample_size = 5000) |>
  dplyr::select(-X)

logistic_results <- rbind(lr_500, lr_1000, lr_2500, lr_5000)
glimpse(logistic_results)
```

```
## Rows: 200
## Columns: 5
## $ dataset_id        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1~
## $ lr_convergence    <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ lr_accuracy       <dbl> 0.790, 0.738, 0.778, 0.758, 0.788, 0.822, 0.776, 0.7~
## $ lr_discrimination <dbl> 0.2020, 0.2402, 0.2730, 0.3123, 0.2663, 0.2171, 0.31~
## $ sample_size       <dbl> 500, 500, 500, 500, 500, 500, 500, 500, 500, 500, 50~
```

Seldonian Solutions

The results data set has 200 observations for each of the simulation trials, 50 from each sample size: $n = 500, 1000, 2500, 5000$.

```r
seldonian_results <- read.csv("/home/dasienga24/Statistics-Senior-Honors-Thesis/Python/COMPAS Simulatio
seldonian_results <- distinct(seldonian_results) #remove duplicate rows
glimpse(seldonian_results)
```

```
## Rows: 200
## Columns: 14
## $ sample_size       <int> 1000, 1000, 1000, 2500, 1000, 1000, 1000, 2500, 500,~
## $ dataset_id        <int> 25, 10, 22, 17, 34, 38, 36, 2, 49, 17, 7, 41, 8, 33,~
## $ passed_safety_02  <chr> "True", "True", "True", "True", "True", "True", "Tru~
## $ passed_safety_01  <chr> "True", "True", "True", "True", "True", "True", "Tru~
## $ passed_safety_005 <chr> "True", "True", "True", "True", "True", "True", "Tru~
## $ passed_safety_001 <chr> "True", "True", "True", "False", "True", "True", "Tr~
## $ sa_02_accuracy    <dbl> 0.6420, 0.6410, 0.6370, 0.7832, 0.5520, 0.6180, 0.73~
## $ sa_01_accuracy    <dbl> 0.5560, 0.5030, 0.5200, 0.4844, 0.4930, 0.5190, 0.49~
## $ sa_005_accuracy   <dbl> 0.5330, 0.5030, 0.5460, 0.4844, 0.4930, 0.5200, 0.49~
## $ sa_001_accuracy   <dbl> 0.5430, 0.5030, 0.5830, 0.4844, 0.4930, 0.5190, 0.49~
## $ sa_02_disc_stat   <dbl> 0.1791, 0.0995, 0.0948, 0.1428, 0.0787, 0.1081, 0.22~
## $ sa_01_disc_stat   <dbl> 0.0345, 0.0000, 0.0355, 0.0000, 0.0000, 0.0000, NA, ~
## $ sa_005_disc_stat  <dbl> 0.0184, 0.0000, 0.0496, 0.0000, 0.0000, NA, 0.0000, ~
## $ sa_001_disc_stat  <dbl> 0.0252, 0.0000, 0.0347, 0.0000, 0.0000, 0.0000, 0.00~
```

Combine the Data Sets from Both Simulations

```r
sim_results <- inner_join(logistic_results, seldonian_results,
                          by = c("sample_size", "dataset_id"))

glimpse(sim_results)
```

```
## Rows: 200
```

```
## Columns: 17
## $ dataset_id         <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1~
## $ lr_convergence     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ lr_accuracy        <dbl> 0.790, 0.738, 0.778, 0.758, 0.788, 0.822, 0.776, 0.7~
## $ lr_discrimination <dbl> 0.2020, 0.2402, 0.2730, 0.3123, 0.2663, 0.2171, 0.31~
## $ sample_size        <dbl> 500, 500, 500, 500, 500, 500, 500, 500, 500, 500, 50~
## $ passed_safety_02   <chr> "True", "True", "True", "True", "True", "True", "Tru~
## $ passed_safety_01   <chr> "True", "True", "False", "True", "True", "False", "T~
## $ passed_safety_005 <chr> "True", "True", "False", "True", "True", "False", "T~
## $ passed_safety_001 <chr> "True", "True", "False", "True", "True", "False", "T~
## $ sa_02_accuracy     <dbl> 0.526, 0.510, 0.528, 0.520, 0.512, 0.508, 0.608, 0.5~
## $ sa_01_accuracy     <dbl> 0.522, 0.510, 0.528, 0.520, 0.496, 0.508, 0.482, 0.5~
## $ sa_005_accuracy    <dbl> 0.522, 0.510, 0.528, 0.520, 0.496, 0.508, 0.756, 0.5~
## $ sa_001_accuracy    <dbl> 0.522, 0.510, 0.528, 0.520, 0.496, 0.508, 0.482, 0.5~
## $ sa_02_disc_stat    <dbl> NA, 0.0000, 0.0000, 0.0000, 0.0429, 0.0000, 0.0106, ~
## $ sa_01_disc_stat    <dbl> 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.00~
## $ sa_005_disc_stat   <dbl> 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.13~
## $ sa_001_disc_stat   <dbl> 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.00~
```

Probability of a Solution

This section assesses what proportion of the trials returned a solution. It is expected that all logistic regression trials will return a solution. However, for the Seldonian algorithms, while all trials will return a candidate solution (based on the logistic regression as a starting point), it is expected that not all candidate solutions will pass the safety test. The table below records the number of Seldonian solutions that passed the safety test in each sample size.

```r
reps <- nrow(sim_results)/4

sim_results |>
  group_by(sample_size) |>
  summarise(LR = 100*count(lr_convergence == 1)/reps,
            `SA (0.2)` = 100*count(passed_safety_02 == "True")/reps,
            `SA (0.1)` = 100*count(passed_safety_01 == "True")/reps,
            `SA (0.05)` = 100*count(passed_safety_005 == "True")/reps,
            `SA (0.01)` = 100*count(passed_safety_001 == "True")/reps) |>
  rename("Sample Size" = sample_size) |>
  kable(caption = "Probability of a Solution")
```

Table 1: Probability of a Solution

Sample Size	LR	SA (0.2)	SA (0.1)	SA (0.05)	SA (0.01)
500	100	100	90	86	70
1000	100	100	100	100	88
2500	100	100	100	100	52
5000	100	100	100	96	28

Accuracy

```r
# filter accuracy data out for only seldonian solutions that converged
sim_results_converged_accuracy <- sim_results |>
  mutate(sa_02_accuracy = ifelse(passed_safety_02 == "True", sa_02_accuracy, NA),
```

```
         sa_01_accuracy = ifelse(passed_safety_01 == "True", sa_01_accuracy, NA),
         sa_005_accuracy = ifelse(passed_safety_005 == "True", sa_005_accuracy, NA),
         sa_001_accuracy = ifelse(passed_safety_001 == "True", sa_001_accuracy, NA))
```

```
# table
sim_results_converged_accuracy |>
  group_by(sample_size) |>
  summarise(LR = round(100*mean(lr_accuracy, na.rm = TRUE),2),
            sd_lr = round(sd(lr_accuracy, na.rm = TRUE),2),
            `SA (0.2)` = round(100*mean(sa_02_accuracy, na.rm = TRUE),2),
            sd_02 = round(sd(sa_02_accuracy, na.rm = TRUE),2),
            `SA (0.1)` = round(100*mean(sa_01_accuracy, na.rm = TRUE),2),
            sd_01 = round(sd(sa_01_accuracy, na.rm = TRUE),2),
            `SA (0.05)` = round(100*mean(sa_005_accuracy, na.rm = TRUE),2),
            sd_005 = round(sd(sa_005_accuracy, na.rm = TRUE),2),
            `SA (0.01)` = round(100*mean(sa_001_accuracy, na.rm = TRUE),2),
            sd_001 = round(sd(sa_001_accuracy, na.rm = TRUE),2)) |>
  rename("Sample Size" = sample_size,
         "sd" = sd_lr,
         "sd " = sd_02,
         "sd  " = sd_01,
         "sd   " = sd_005,
         "sd     " = sd_001) |>
  kable(caption = "Accuracy of Convergent Solutions")
```

Table 2: Accuracy of Convergent Solutions

Sample Size	LR	sd	SA (0.2)	sd	SA (0.1)	sd	SA (0.05)	sd	SA (0.01)	sd
500	78.22	0.02	55.26	0.07	52.32	0.07	52.27	0.06	51.09	0.04
1000	77.89	0.01	58.21	0.08	52.24	0.05	51.55	0.04	51.54	0.05
2500	77.85	0.01	71.71	0.07	57.24	0.06	51.08	0.02	50.44	0.01
5000	78.04	0.01	71.12	0.07	65.31	0.09	55.71	0.06	50.04	0.01

```
# visual
```

make a visual

Discrimination

```
# filter discrimination data out for only seldonian solutions that converged
sim_results_converged_disc <- sim_results |>
  mutate(sa_02_disc_stat = ifelse(passed_safety_02 == "True", sa_02_disc_stat, NA),
         sa_01_disc_stat = ifelse(passed_safety_01 == "True", sa_01_disc_stat, NA),
         sa_005_disc_stat = ifelse(passed_safety_005 == "True", sa_005_disc_stat, NA),
         sa_001_disc_stat = ifelse(passed_safety_001 == "True", sa_001_disc_stat, NA))
```

```
# table
sim_results_converged_disc |>
  group_by(sample_size) |>
  summarise(LR = round(100*mean(lr_discrimination, na.rm = TRUE),2),
            sd_lr = round(sd(lr_discrimination, na.rm = TRUE),2),
```

```r
            `SA (0.2)` = round(100*mean(sa_02_disc_stat, na.rm = TRUE),2),
            sd_02 = round(sd(sa_02_disc_stat, na.rm = TRUE),2),
            `SA (0.1)` = round(100*mean(sa_01_disc_stat, na.rm = TRUE),2),
            sd_01 = round(sd(sa_01_disc_stat, na.rm = TRUE),2),
            `SA (0.05)` = round(100*mean(sa_005_disc_stat, na.rm = TRUE),2),
            sd_005 = round(sd(sa_005_disc_stat, na.rm = TRUE),2),
            `SA (0.01)` = round(100*mean(sa_001_disc_stat, na.rm = TRUE),2),
            sd_001 = round(sd(sa_001_disc_stat, na.rm = TRUE),2)) |>
  rename("Sample Size" = sample_size,
         "sd" = sd_lr,
         "sd " = sd_02,
         "sd  " = sd_01,
         "sd   " = sd_005,
         "sd     " = sd_001) |>
  kable(caption = "Discrimination of Convergent Solutions")
```

Table 3: Discrimination of Convergent Solutions

Sample Size	LR	sd	SA (0.2)	sd	SA (0.1)	sd	SA (0.05)	sd	SA (0.01)	sd
500	24.03	0.08	5.07	0.07	2.34	0.07	1.80	0.06	0.65	0.03
1000	24.20	0.05	8.31	0.07	2.21	0.05	1.19	0.04	0.97	0.04
2500	24.13	0.04	16.48	0.07	6.79	0.06	1.27	0.02	0.06	0.00
5000	24.88	0.03	13.57	0.06	13.17	0.08	5.60	0.06	0.02	0.00

```r
# visual
```

make a visual

```r
# table
sim_results_converged_disc |>
  group_by(sample_size) |>
  summarise(`SA (0.2)` = round(100*count(sa_02_disc_stat < 0.2)/count(passed_safety_02 == "True"),2),
            `SA (0.1)` = round(100*count(sa_01_disc_stat < 0.1)/count(passed_safety_01 == "True"),2),
            `SA (0.05)` = round(100*count(sa_005_disc_stat < 0.05)/count(passed_safety_005 == "True"),2
            `SA (0.01)` = round(100*count(sa_001_disc_stat < 0.01)/count(passed_safety_001 == "True"),2
  rename("Sample Size" = sample_size) |>
  kable(caption = "Satisfaction of the Behavioral Constraint by Convergent Solutions")
```

Table 4: Satisfaction of the Behavioral Constraint by Convergent Solutions

Sample Size	SA (0.2)	SA (0.1)	SA (0.05)	SA (0.01)
500	84	82.22	90.70	91.43
1000	88	84.00	84.00	90.91
2500	70	70.00	82.00	96.15
5000	82	30.00	52.08	100.00

Accuracy-Discrimination Trade-Off (means)

Visual displaying the trade-off

Convergent-Discrimination Trade-Off (counts)

visual displaying the relationship

NSF

compare these solutions with logistic regression