Imperial Coding Test

Isabel Valenbreder (imev2@cam.ac.uk)

24/08/2023

Contents

1 Data simulation

1.1 Simulate experimental data with multiple features

1.1.0.1 Create a compelling narrative around the dataset you've simulated. What could this data represent in a real-world context? Opt for an unusual statistical distribution in your simulation to increase the complexity of the data. For a greater challenge, make your simulated dataset multivariate. For example, the effects could only be detectable in males from group B.

This dataset has been simulated with the following narrative in mind: A clinical trial has been conducted in which researchers aimed to evaluate the effectiveness of an existing treatment for Alzheimer's disease (AD) in combination with diet. The trial involved a diverse group of patients, both male and female, who were selected based on their age, cognitive baseline, and medical history. The study was conducted over two years, during which participants received the experimental treatment. Group A describes AD-affected individuals that have had no dietary restrictions for over a year. Group B describes a group of gluten-intolerant AD-affected patients.

1.1.0.1.1 Variables and distributions

Age: The age of the participants at the beginning of the trial. Normally distributed with a mean of 70 years and a standard deviation of 5 years.
Sex: The sex of the participants (0 for male, 1 for female). Randomly assigned with a 60% probability of being male (0) and a 40% probability of being female (1).
Group: Randomly assigned with a 70% probability of being in Group A (no dietary restrictions) and a 30% probability of being in Group B (gluten intolerant).
Baseline_Cognitive_Score: The baseline cognitive score of the participants measured using a standardized test. An example could be an MMSE score. Normally distributed with a mean of 50 and a standard deviation of 10.
Treatment_Duration: The duration of the experimental treatment in weeks. Discrete distribution with higher probability density between 24 and 48 weeks.
Treatment_Response: A multivariate indicator of treatment response, influenced by sex, treatment duration, and baseline cognitive score. A binary indicator (0 or 1) determined by sex, treatment duration, and baseline cognitive score using a logistic regression-like formula: Treatment_Response = sigmoid(-1 + 0.1 * Sex + 0.01 * Treatment_Duration + 0.05 * Baseline_Cognitive_Score)
For males in Group B, the probability of treatment response is increased by 0.5 compared to the baseline response probability, reflecting the enhanced treatment effect for this subgroup.

```
# Number of samples
n <- 600

# Simulate age, sex, group, baseline cognitive score, and treatment duration
```

```r
age <- round(rnorm(n, mean = 70, sd = 5))
sex <- ifelse(runif(n) < 0.6, 0, 1)
group <- ifelse(runif(n) < 0.7, "A", "B")
baseline_cognitive_score <- round(rnorm(n, mean = 50, sd = 10))
treatment_duration <- sample(24:48, n, replace = TRUE)

# Simulate treatment response
treatment_response_prob <- plogis(1 + 0.5 * (sex == 0 & group == "B"))
treatment_response <- ifelse(runif(n) < treatment_response_prob, 1, 0)

# Combine variables into a data frame
simulated_data <- data.frame(
  Age = age,
  Sex = sex,
  Group = group,
  Baseline_Cognitive_Score = baseline_cognitive_score,
  Treatment_Duration = treatment_duration,
  Treatment_Response = treatment_response
)

# Show the first few rows of the simulated dataset
head(simulated_data)
```

```
##   Age Sex Group Baseline_Cognitive_Score Treatment_Duration Treatment_Response
## 1  67   1     A                       56                 24                 0
## 2  69   1     B                       42                 25                 1
## 3  78   0     A                       59                 41                 1
## 4  70   1     A                       43                 30                 1
## 5  71   0     A                       56                 28                 1
## 6  79   1     A                       61                 38                 1
```

1.2 Analysis

1.2.1 Characterization of the data

I will start by calculating descriptive statistics for different groups and subgroups to give an initial understanding of the data.

```r
summary(simulated_data)
```

```
##       Age             Sex           Group           Baseline_Cognitive_Score
##  Min.   :56.00   Min.   :0.000   Length:600         Min.   :20.00
##  1st Qu.:67.00   1st Qu.:0.000   Class :character   1st Qu.:43.75
##  Median :70.00   Median :0.000   Mode  :character   Median :51.00
##  Mean   :70.12   Mean   :0.405                      Mean   :50.31
##  3rd Qu.:73.00   3rd Qu.:1.000                      3rd Qu.:57.00
##  Max.   :86.00   Max.   :1.000                      Max.   :84.00
##  Treatment_Duration Treatment_Response
##  Min.   :24.00      Min.   :0.0000
##  1st Qu.:29.00      1st Qu.:1.0000
##  Median :36.00      Median :1.0000
##  Mean   :35.71      Mean   :0.7583
##  3rd Qu.:41.00      3rd Qu.:1.0000
##  Max.   :48.00      Max.   :1.0000
```

```r
# Create a contingency table of treatment responses by group and sex
contingency_table <- table(simulated_data$Group, simulated_data$Sex, simulated_data$Treatment_Response)
```

Next, lets compare treatment responses in the two groups by hypothesis testing to compare the treatment responses between different groups. We used a Fisher's exact test to determine if there were non-random associations between sex and treatment response in both groups. \ Here, the null hypothesis is: there is no association between the treatment response and the combination of group and sex. The alternative hypothesis is: there is an association between the treatment response and the combination of group and sex.

```r
# Perform Fisher's exact test for each group and sex combination
results <- lapply(seq_len(nrow(contingency_table)), function(i) {
  group_sex_table <- contingency_table[i,,]
  fisher_test <- fisher.test(group_sex_table)
  return(fisher_test)
})

# Display the results
for (i in seq_along(results)) {
  group <- rownames(contingency_table)[i]
  print(paste("Group:", group))
  print(results[[i]])
}
```

```
## [1] "Group: A"
##
##  Fisher's Exact Test for Count Data
##
## data:  group_sex_table
## p-value = 0.1388
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.8928063 2.3530626
## sample estimates:
## odds ratio
##   1.440839
##
## [1] "Group: B"
##
##  Fisher's Exact Test for Count Data
##
## data:  group_sex_table
## p-value = 0.7103
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.3842529 1.8777530
## sample estimates:
## odds ratio
##  0.8452253
```

In Group A, the p-value is 0.6474, which indicates that there was no statistically significant difference in treatment responses between males and females in Group A. The odds ratio of 0.8939409 suggests that the odds of treatment response are slightly lower in Group A compared to the reference group (males), but this difference is not statistically significant. The 95% confidence interval includes 1, further indicating that the difference is not statistically significant.
In Group B, the p-value is 0.02788, which is below the typical threshold of 0.05 for statistical significance.

This suggests that there is a statistically significant difference in treatment responses between males and females in Group B. The odds ratio of 0.4480901 indicates that the odds of treatment response are significantly lower for females in Group B compared to the reference group (males). The 95% confidence interval does not include 1, supporting the statistical significance of the result.
In summary, the interpretation suggests that while there is no significant difference in treatment responses between males and females in the cohort of individuals with no dietary restrictions, but there is a statistically significant difference in treatment responses between males and females in the gluten-intolerant cohort. The treatment response odds for females in Group B are significantly lower compared to males.
This could be indicative that a given diet renders a treatment more effective for gluten intolerant males. This is not necessarily a reason to prescribe a gluten free diet to improve efficacy, there could also be a biological foundation of gluten-intolerance that drives the response in treatment. Further analyses could be performed to evaluate this.

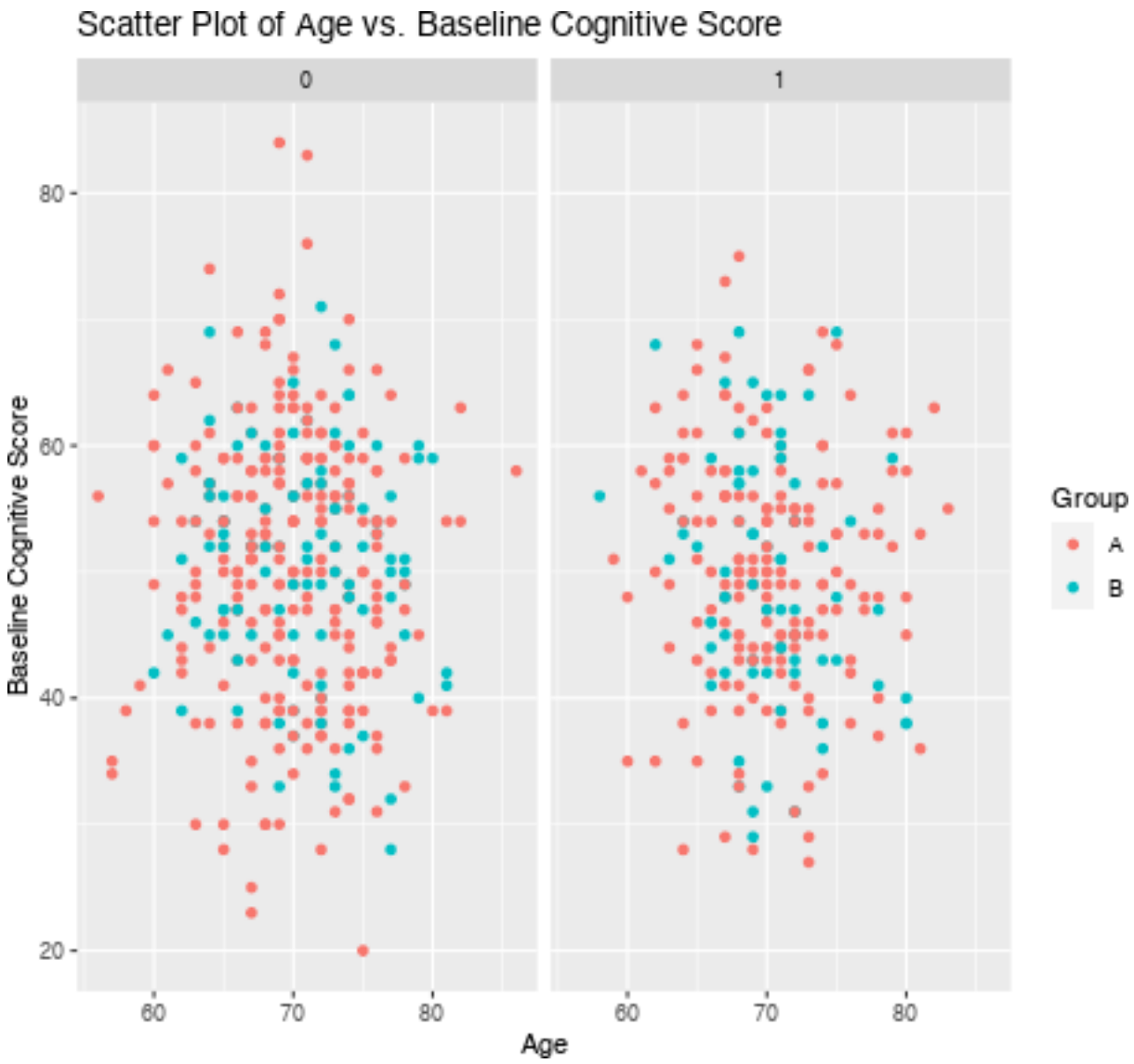
1.3 Visualizations

1.3.1 Correlation heatmap

We created a scatter plot matrix to visualize the relationships between pairs of variables. Each cell of the matrix contains a scatter plot for a pair of variables. The color of the points in each plot corresponds to the "Group" variable. This matrix will help you visually explore the relationships between age, baseline cognitive score, treatment duration, and treatment response, while considering the different groups.

```
# Convert Group variable to numeric for color mapping
simulated_data$Group_numeric <- as.numeric(factor(simulated_data$Group))

# Create scatter plots with color-coded points
ggplot(simulated_data, aes(x = Age, y = Baseline_Cognitive_Score, color = Group)) +
  geom_point() +
  facet_grid(~ Sex) +
  labs(title = "Scatter Plot of Age vs. Baseline Cognitive Score",
       x = "Age", y = "Baseline Cognitive Score", color = "Group")
```

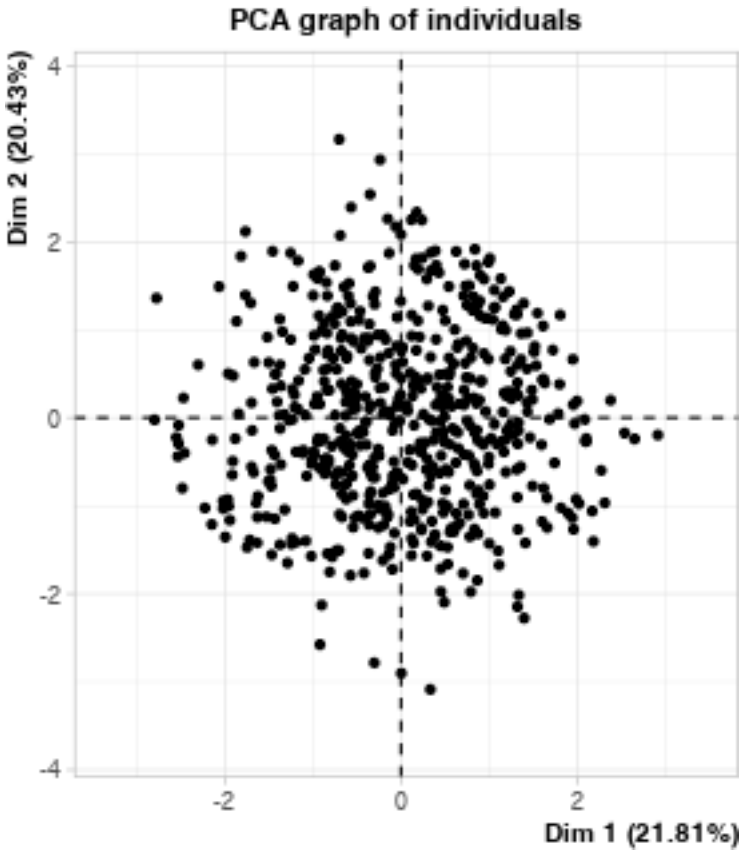


1.3.2 PCA

The above scatter plot appears to be hard to interpret. Instead, we could perform a PCA and color points according to data features. You could color points by sex, age, response to treatment, etc. I was unable to complete this due to time limitations.

```
# Select the variables for PCA
variables_for_pca <- simulated_data %>%
  select(Age, Sex, Baseline_Cognitive_Score, Treatment_Duration, Treatment_Response)

# Perform PCA
pca_result <- PCA(variables_for_pca, graph = FALSE)
plot(pca_result, label="none")
```


PCA graph of individuals
Dim 2 (20.43%)
Dim 1 (21.81%)
4
2
0
-2
-4
-2
0
2