PS2_Submit

2025-01-24

Question A: Homogeneous Treatment Effects Model

1. Describe the Unit-Level Treatment Effect (ITE)

- **Mathematical Representation**:
  The individual treatment effect (ITE) is defined as:

  $$\text{ITE}_i = Y_i(1) - Y_i(0)$$

  where $Y_i(1)$ represents the outcome (spending) for customer i if they are a member, and $Y_i(0)$ represents the outcome if they are not.

- **Verbal Representation**:
  The ITE captures the difference in Walmart spending for an individual customer if they are a Sam's Club Plus member versus if they are not.

2. Prove $Y_i = \alpha + \textbf{ITE}_i D_i + U_i$

Using the Rubin Causal Model (RCM), the observed outcome can be expressed as:

$$Y_i = Y_i(0)(1 - D_i) + Y_i(1)D_i$$

Expanding this equation:

$$Y_i = Y_i(0) + (Y_i(1) - Y_i(0))D_i$$

Let:
- $\alpha = Y_i(0)$: The baseline outcome (spending) for non-members.
- $\text{ITE}_i = Y_i(1) - Y_i(0)$: The individual treatment effect.

Substituting these definitions gives:

$$Y_i = \alpha + \text{ITE}_i D_i + U_i$$

where U_i represents the error term, capturing unobserved factors influencing Y_i. Assuming exogeneity, $\mathbb{E}[U_i] = 0$.

Examples of U_i:
U_i could include factors such as: - Household size. - Income level. - Shopping preferences.

3. Compute the Bias of $\hat{\beta}_n$ for β

Definition of Bias:

$$\text{Bias}_\theta(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta$$

Using the independence assumption $\{Y_i(1), Y_i(0)\} \perp D_i$:

$$\text{Bias} = [\mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0]] - [\mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]]$$

This simplifies to:

$$\text{Bias} = [\mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]] - [\mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]] = 0$$

Thus, under the independence assumption, the bias is zero.

4. Bias with Simplified Expressions

Given $\hat{\beta} = \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]$, the bias can be rewritten as:

$$\text{Bias} = [\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]] - [\mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]]$$

Under the overlap condition $0 < \Pr(D_i = 1 \mid X) < 1$, this expression ensures no bias because the treatment and control groups are comparable within strata of X.

Question B: Regression Model for ATU

1. Specify α and U_i

To model the Average Treatment Effect for the Untreated (ATU):

$$Y_i = \alpha + \text{ATU} \cdot D_i + U_i$$

Here:
- $\alpha = \mathbb{E}[Y_i(0)]$: The average outcome for non-treated individuals.
- U_i: Captures deviations from the average untreated outcome, including unobserved factors.

2. Exogeneity Condition

The exogeneity condition requires that $\mathbb{E}[U_i \mid D_i] = 0$, ensuring that the treatment assignment does not depend on unobserved confounders.

Question C: Conditional Random Assignment (CRA)

1. Correlation of Variables with Treatment

- **Number of Walmart visits (likely correlated)**: Customers who visit Walmart frequently may perceive higher benefits from membership.
- **Amazon spending (uncorrelated)**: Membership at Walmart likely does not influence spending on Amazon.
- **Number of children (likely correlated)**: Families with children may visit Walmart more often, leading to higher membership uptake.

2. Checking the CRA Assumption

To check CRA:
- Stratify the data by tenure levels.
- Perform t-tests on pre-determined variables (e.g., number of visits, number of children) to assess balance across treatment groups.

3. Identification of CATE

CATE is identified when:
1. Conditional independence: $\{Y(1), Y(0)\} \perp D \mid X$.
2. Overlap: $0 < \Pr(D = 1 \mid X = x) < 1$.

4. Lack of Overlap

If $\Pr(D = 1 \mid X = x) = 0$ or 1, treatment and control groups are not comparable, making CATE unidentifiable.

5. Estimating ATE

Code Example:

```
estimate_ate <- function(data) {
  tenure_groups <- unique(data$tenureD)

  cate_estimates <- sapply(tenure_groups, function(x) {
    group_data <- subset(data, tenureD == x)
    mean(group_data$spend[group_data$scp == 1]) -
    mean(group_data$spend[group_data$scp == 0])
  })

  emp_dist <- table(data$tenureD) / nrow(data)
  ate <- sum(cate_estimates * emp_dist)

  return(ate)
}
```

Question D

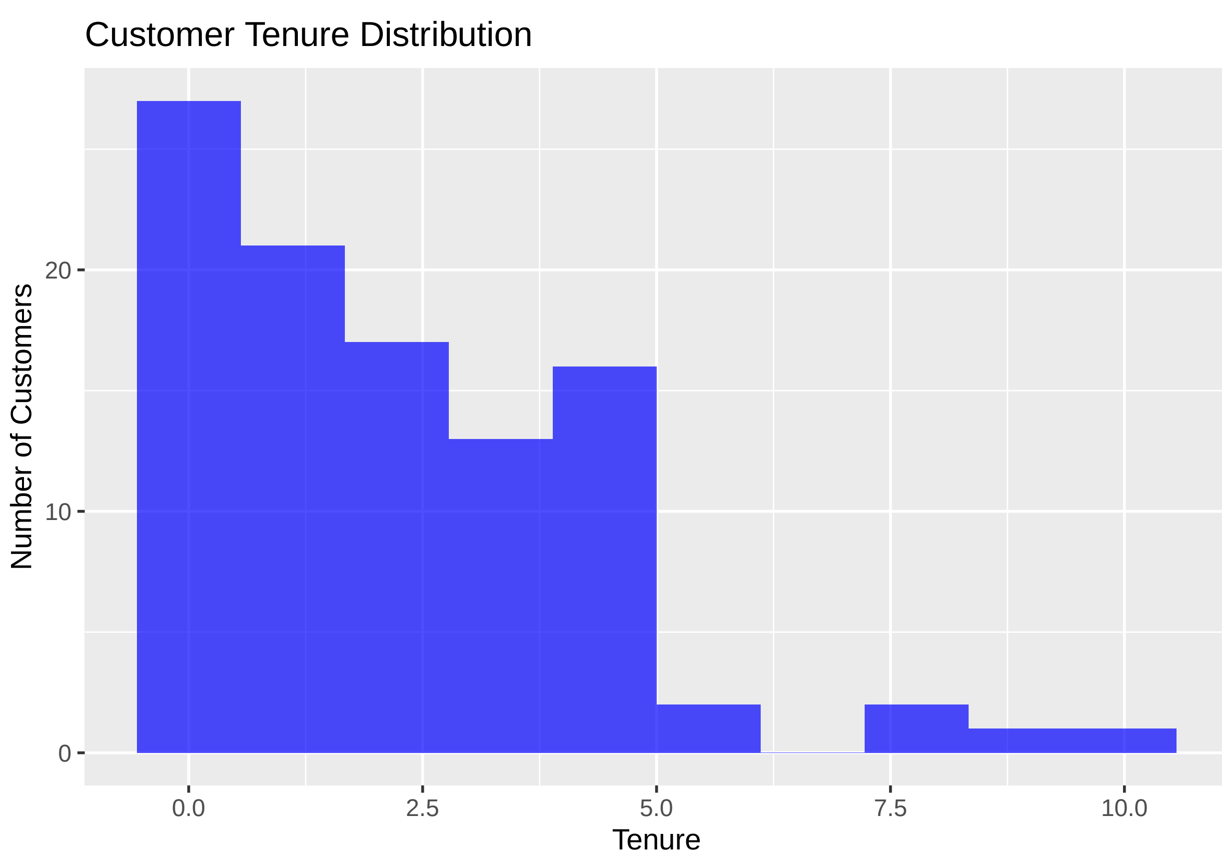
```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
walmart_data <- read.csv("SCP_data.csv", stringsAsFactors = FALSE)
# Part 1: tenure distirbtuion
tenure_plot <- ggplot(walmart_data, aes(x = tenure)) +
  geom_histogram(bins = 10, alpha = 0.7, fill = "blue") +
  ggtitle("Customer Tenure Distribution") +
  xlab("Tenure") +
  ylab("Number of Customers")
print(tenure_plot)
```



```r
# Given the distribution being skewed,
# we would not want to say that the indepdence assumption is gone
```

```r
# Part 2: discrete assignment
walmart_data <- walmart_data %>% mutate(
  tenureD = case_when(
    tenure <= 1 ~ 1,
    tenure > 5 ~ 3,
    TRUE ~ 2
  )
)
# Discrete size tenure should be independence,
# and we could create a more uniform distribution based on so
```

```r
# Part 3: Conditional Average Treatment Effect Estimation
cate <- lm(spend ~ scp + factor(tenureD), data = walmart_data)
cate
```

```
##
## Call:
## lm(formula = spend ~ scp + factor(tenureD), data = walmart_data)
##
## Coefficients:
##      (Intercept)              scp  factor(tenureD)2  factor(tenureD)3
##            17.69            36.73             36.58            101.76
```

```r
# Part 4: Estimator Comparison

# dsicrete partition CATE
d_partition <- walmart_data %>%
  group_by(tenureD) %>%
  summarize(
    mean_spend_scp_1 = mean(spend[scp == 1]),
    mean_spend_scp_0 = mean(spend[scp == 0]),
    partition_ate = mean_spend_scp_1 - mean_spend_scp_0
  )

d_partition
```

```
## # A tibble: 3 x 4
##   tenureD mean_spend_scp_1 mean_spend_scp_0 partition_ate
##     <dbl>            <dbl>            <dbl>         <dbl>
## 1       1             55.3             17.5          37.8
## 2       2             90.0             55.2          34.8
## 3       3            160.             112.           47.9
```

```r
# difference in expectant values (assumes homogeneity)
exp_diff <- mean(walmart_data$spend[walmart_data$scp == 1], na.rm = TRUE) -
  mean(walmart_data$spend[walmart_data$scp == 0], na.rm = TRUE)
exp_diff
```

```
## [1] 56.32509
```

```r
# The partition/ tenure D is more preferred because we assume heterogeneity effect;
# mean difference appraoch may overestiamte the effect of the memership in spending
```

```r
# Part 5: Comparison
true_ate <- mean(walmart_data$spend1 - walmart_data$spend0)

d_partition_value <- mean(d_partition$partition_ate) # Aggregate to a single value
comparison <- data.frame(
  method = c("Discrete CATE", "Difference in Expected Values", "Real ATE"),
  value_estimate = c(d_partition_value, exp_diff, true_ate)
)
print(comparison)
```

```
##                          method value_estimate
## 1                 Discrete CATE       40.17710
## 2 Difference in Expected Values       56.32509
## 3                      Real ATE       40.03792
```