Homework 4

ECON 470, Spring 2025

Sushmita Rajan

Here is a link to my repository: {https://github.com/sarajan03/econ470spring2025/tree/main/Homework4}
emove all SNPs, 800-series plans, and prescription drug only plans (i.e., plans that do not offer Part C benefits). Provide a box and whisker plot showing the distribution of plan counts by county over time. Do you think that the number of plans is sufficient, too few, or too many?

```
# Step 1: Filter the data
filtered_df <- final_data %>%
  filter(snp == "No",              # Remove SNPs
         eghp == "No",              # Remove eghp plans
         partd == "Yes")           # Keep only plans with Part C benefits

# Step 2: Group by 'county' and 'year' (since you have 'county' and 'year' columns)
plan_counts <- filtered_df %>%
  group_by(county, year) %>%  # Group by county and year
  summarise(plan_count = n(), .groups = 'drop')  # Count the number of plans in each group

# Step 3: Calculate summary statistics for the plan counts across all counties for each year
plan_stats <- plan_counts %>%
  group_by(year) %>%
  summarise(
    min = min(plan_count),
    q1 = quantile(plan_count, 0.25),
    median = median(plan_count),
    mean = mean(plan_count),
    q3 = quantile(plan_count, 0.75),
    max = max(plan_count)
  )

# Step 4: Reshape data for boxplot
# Pivot the summary statistics to a long format so it can be plotted
```

```r
plan_stats_long <- plan_stats %>%
  pivot_longer(cols = c(min, q1, median, mean, q3, max),
               names_to = "statistic",
               values_to = "value")

# Step 5: Create a box and whisker plot based on summary statistics
question1 <-ggplot(plan_stats_long, aes(x = factor(year), y = value, fill = statistic)) +
  geom_boxplot() +
  labs(title = "Box and Whisker Plot of Summary Statistics by Year",
       x = "Year",  # 'year' is on the x-axis
       y = "Summary Statistics Value") +
  theme_minimal() +
  theme(legend.title = element_blank()) # Optional: to remove the legend title
```

Provide bar graphs showing the distribution of star ratings in 2010, 2012, and 2015. How has this distribution changed over time?

```r
# Step 1: Filter data for the years 2010, 2012, and 2015
filtered_data <- final_data %>%
  filter(year %in% c(2010, 2012, 2015))  # Filter for specific years

# Step 2: Group by year and Star_Rating to count occurrences of each rating
rating_counts <- filtered_data %>%
  group_by(year, Star_Rating) %>%
  summarise(count = n(), .groups = 'drop')

# Step 3: Create the bar plot

question2 <-ggplot(rating_counts, aes(x = factor(Star_Rating), y = count, fill = factor(year)
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Distribution of Star Ratings Over Time (2010, 2012, 2015)",
       x = "Star Rating",
       y = "Count of Plans",
       fill = "Year") +
  theme_minimal() +
  scale_x_discrete(limits = sort(unique(rating_counts$Star_Rating)))  # Ensure ratings are o
```

```
Warning in scale_x_discrete(limits = sort(unique(rating_counts$Star_Rating))): Continuous li
i Did you mean `limits = factor(...)` or `scale_*_continuous()`?
```

Plot the average benchmark payment over time from 2010 through 2015. How much has the average benchmark payment risen over the years?

```r
# Step 1: Calculate the average benchmark payment using mean_risk for each year
avg_benchmark <- final_data %>%
  filter(year >= 2010 & year <= 2015) %>%
  group_by(year) %>%
  summarise(average_benchmark = mean(mean_risk, na.rm = TRUE))  # Using mean_risk as benchmar

# Step 2: Plot the average benchmark payment (mean_risk) over time (2010-2015)
question3 <-ggplot(avg_benchmark, aes(x = year, y = average_benchmark)) +
  geom_line(color = "blue", size = 1) +   # Line plot to show the trend
  geom_point(color = "red", size = 3) +   # Red points to highlight each year
  labs(title = "Average Benchmark Payment (Mean Risk) Over Time (2010-2015)",
       x = "Year",
       y = "Average Benchmark Payment (Mean Risk)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for bette
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

Plot the average share of Medicare Advantage (relative to all Medicare eligibles) over time from 2010 through 2015. Has Medicare Advantage increased or decreased in popularity? How does this share correlate with benchmark payments?

```r
# # Step 1: Calculate the Share of Medicare Advantage Enrollment relative to Total Medicare
# medicare_share <- final_data %>%
#   filter(year >= 2010 & year <= 2015) %>%
#   mutate(ma_share = avg_enrolled / avg_eligibles)  # Share of Medicare Advantage

# # Step 2: Calculate the Average Share per Year
# avg_share_per_year <- medicare_share %>%
#   group_by(year) %>%
#   summarise(average_ma_share = mean(ma_share, na.rm = TRUE))  # Average share for each yea

# # Step 3: Plot the Trend of Medicare Advantage Share Over Time (2010-2015)
# question4<-ggplot(avg_share_per_year, aes(x = year, y = average_ma_share)) +
#   geom_line(color = "green", size = 1) +    # Line plot to show the trend
#   geom_point(color = "red", size = 3) +     # Red points to highlight each year
#   labs(title = "Average Share of Medicare Advantage Over Time (2010-2015)",
#        x = "Year",
#        y = "Average Share of Medicare Advantage") +
#   theme_minimal() +
```

```
#   theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for rea

# # Step 4: Check the Correlation Between the Share of Medicare Advantage and Benchmark Payme
# correlation_data <- medicare_share %>%
#   group_by(year) %>_
```

Estimate ATEs For the rest of the assignment, we'll use a regression discontinuity design to estimate the average treatment effect from receiving a marginally higher rating. We'll focus only on 2010.

Calculate the running variable underlying the star rating. Provide a table showing the number of plans that are rounded up into a 3-star, 3.5-star, 4-star, 4.5-star, and 5-star rating.

```
# ma.data.2010 <- final_data %>%
#   filter(year==2010)

# star_rating_table <- ma.data.2010 %>%
#   filter(Star_Rating >= 3 & Star_Rating <= 5) %>%
#   group_by(Star_Rating) %>%
#   summarise(number_of_plans = n()) %>%
#   arrange(Star_Rating)

# # Print the table showing the number of plans for each star rating
# print(star_rating_table)
```

Using the RD estimator with a bandwidth of 0.125, provide an estimate of the effect of receiving a 3-star versus a 2.5 star rating on enrollments. Repeat the exercise to estimate the effects at 3.5 stars, and summarize your results in a table.

```
# # 1. Filter and preprocess the data
# ma.rd1 <- ma.data.2010 %>%
#   filter(
#     !is.na(Star_Rating),
#     !is.na(avg_enrollment),
#     !is.na(avg_eligibles)
#   ) %>%
#   mutate(
#     mkt_share = avg_enrollment / avg_eligibles
#   )

# # 2. Estimate RD at 3.0 star cutoff
# est_3.0 <- rdrobust(
```

```
#   y = ma.rd1$mkt_share,
#   x = ma.rd1$Star_Rating,
#   c = 3.0,
#   h = 0.125,
#   p = 1,
#   kernel = "uniform",
#   vce = "hc0",
#   masspoints = "off"
# )

# # 3. Estimate RD at 3.5 star cutoff
# est_3.5 <- rdrobust(
#   y = ma.rd1$mkt_share,
#   x = ma.rd1$Star_Rating,
#   c = 3.5,
#   h = 0.125,
#   p = 1,
#   kernel = "uniform",
#   vce = "hc0",
#   masspoints = "off"
# )

# # 4. Summarize results in a table
# results <- data.frame(
#   Cutoff = c(3.0, 3.5),
#   Estimate = c(est_3.0$coef[1], est_3.5$coef[1]),
#   Std_Error = c(est_3.0$se[1], est_3.5$se[1])
# )

# print(results)
```

Repeat your results for bandwidhts of 0.1, 0.12, 0.13, 0.14, and 0.15 (again for 3 and 3.5 stars). Show all of the results in a graph. How sensitive are your findings to the choice of bandwidth?

```
# # Load necessary libraries
# library(dplyr)
# library(rdrobust)
# library(ggplot2)

# # 1. Filter and preprocess the data
# ma.rd1 <- ma.data.2010 %>%
```

```
#   filter(
#     !is.na(Star_Rating),
#     !is.na(avg_enrollment),
#     !is.na(avg_eligibles)
#   ) %>%
#   mutate(
#     mkt_share = avg_enrollment / avg_eligibles
#   )

# # 2. Define bandwidths and the cutoffs
# bandwidths <- c(0.1, 0.12, 0.13, 0.14, 0.15)
# cutoffs <- c(3.0, 3.5)

# # 3. Initialize an empty data frame to store results
# results_df <- data.frame(
#   Bandwidth = numeric(),
#   Cutoff = character(),
#   Estimate = numeric(),
#   Std_Error = numeric()
# )

# # 4. Loop through bandwidths and cutoffs to run the RD analysis
# for (cutoff in cutoffs) {
#   for (bw in bandwidths) {
#     est <- rdrobust(
#       y = ma.rd1$mkt_share,
#       x = ma.rd1$Star_Rating,
#       c = cutoff,
#       h = bw,
#       p = 1,
#       kernel = "uniform",
#       vce = "hc0",
#       masspoints = "off"
#     )

#     # Store the results
#     results_df <- rbind(results_df, data.frame(
#       Bandwidth = bw,
#       Cutoff = as.character(cutoff),
#       Estimate = est$coef[1],
#       Std_Error = est$se[1]
#     ))
```

```
#   }
# }

# # 5. Plot the results
# ggplot(results_df, aes(x = Bandwidth, y = Estimate, color = Cutoff)) +
#   geom_line() +
#   geom_point() +
#   geom_errorbar(aes(ymin = Estimate - 1.96 * Std_Error, ymax = Estimate + 1.96 * Std_Error)
#   labs(
#     title = "RD Treatment Effect Across Different Bandwidths",
#     x = "Bandwidth",
#     y = "Estimated Treatment Effect",
#     color = "Cutoff"
#   ) +
#   theme_minimal()
```

Examine (graphically) whether contracts appear to manipulate the running variable. In other words, look at the distribution of the running variable before and after the relevent threshold values. What do you find?

```
# ggplot(ma.rd1, aes(x = Star_Rating)) +
#   geom_histogram(
#     bins = 50, fill = "skyblue", color = "black", alpha = 0.7
#   ) +
#   xlim(2.8, 3.2) +  # Focus on the range near 3.0
#   geom_vline(xintercept = 3.0, color = "red", linetype = "dashed", size = 1) +
#   ggtitle("Histogram of Star Ratings Around the 3.0 Threshold") +
#   xlab("Star Rating") +
#   ylab("Frequency") +
#   theme_minimal()

# # Similarly, for the second threshold (e.g., 3.5)
# ggplot(ma.rd1, aes(x = Star_Rating)) +
#   geom_histogram(
#     bins = 50, fill = "lightgreen", color = "black", alpha = 0.7
#   ) +
#   xlim(3.3, 3.7) +  # Focus on the range near 3.5
#   geom_vline(xintercept = 3.5, color = "blue", linetype = "dashed", size = 1) +
#   ggtitle("Histogram of Star Ratings Around the 3.5 Threshold") +
#   xlab("Star Rating") +
#   ylab("Frequency") +
#   theme_minimal()
```

Similar to question 4, examine whether plans just above the threshold values have different characteristics than contracts just below the threshold values. Use HMO and Part D status as your plan characteristics. Summarize your findings from 5-9. What is the effect of increasing a star rating on enrollments? Briefly explain your results.

I'm currently facing some issues with my analysis. Specifically, not all of my plots are rendering properly, and I'm encountering an error when trying to run the regression discontinuity (RD) model. However, I do believe that plans that meet or exceed certain threshold values, like the 3.0 or 3.5-star rating thresholds, likely exhibit different characteristics—such as enrollment levels or plan types—compared to those just below these thresholds. I suspect that there is a significant effect on enrollments when a plan's rating crosses these thresholds, potentially due to better visibility, more favorable perceptions, or eligibility for certain benefits associated with higher ratings. The RD analysis should help shed light on this by estimating the treatment effect of crossing the threshold in terms of enrollments, but I need to resolve the issues with the plots and the RD model error to move forward.