Problem Set 2

Romain Fernex

2025-02-14

Contents

1 Problem 1

1.1 Load the data

```
load("/Users/rfernex/Documents/Education/SciencesPo/Courses/S2/Econometrics/TD/Psets/Pset2/Data/401k.Rda
```

1.2 Question 1

```
mean_participation <- mean(data$prate)
mean_match <- mean(data$mrate)
cat("average participation rate :", mean_participation)
```

```
average participation rate : 87.36291
```

```
cat("average match rate :", mean_match)
```

```
average match rate : 0.7315124
```

1.3 Question 2

```
model <- lm(prate ~ mrate, data = data)

stargazer(model, type = "text",
          title = "Regression Summary: prate vs mrate",
          dep.var.labels = "Participation rate",
          covariate.labels = c("Match rate", "Intercept"),
          omit.stat = c("f"))
```

```
Regression Summary: prate vs mrate
===============================================
                        Dependent variable:
                    ---------------------------
                        Participation rate
-----------------------------------------------
Match rate                   5.861***
                              (0.527)

Intercept                    83.075***
                              (0.563)

-----------------------------------------------
Observations                   1,534
R2                             0.075
Adjusted R2                    0.074
Residual Std. Error     16.085 (df = 1532)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

```r
num_obs <- nobs(model)
r_squared <- summary(model)$r.squared

cat("Sample Size :", num_obs)
```

```
Sample Size : 1534
```

```r
cat("R-squared :", sprintf("%.2f%%", r_squared*100))
```

```
R-squared : 7.47%
```

1.4 Question 3

```r
data$fitted_values <- fitted(model)
sum_of_residuals <- sum(residuals(model))
cat("Sum of Residuals:", sprintf("%.2f", sum_of_residuals), "\
")
```

```
Sum of Residuals: 0.00
```

```r
kable(head(data$fitted_values, 10), col.names = c("Fitted Values"))
```

Fitted Values
84.30628
91.39819
88.40904
85.53711
86.18183
93.74262
86.18183
85.06822
84.36489
86.59210

1.5 Question 4

```r
intercept <- coef(model)["(Intercept)"]
coefficient_mrate <- coef(model)["mrate"]

cat("Intercept :", sprintf("%.2f", intercept), "\
")
```

```
Intercept : 83.08
```

```r
cat("Coefficient for the match rate :", sprintf("%.2f", coefficient_mrate), "\
")
```

```
Coefficient for the match rate : 5.86
```

Based on the coefficient for the match rate, increasing the match rate by 1 percent raises the participation rate by 5.86 percent. The intercept is at 83.08%, this means that the predicted participation rate will always be superior or equal to 83%. This seems to indicate that participation rate of eligible workers tends to be pretty high regardless of their company's generosity.

1.6 Question 5

```r
source_mrate = 3.5 # adaptable
predicted_y <- intercept + coefficient_mrate * source_mrate
cat("Predicted participation rate for a match rate of", source_mrate, ":", predicted_y)
```

```
Predicted participation rate for a match rate of 3.5 : 103.5892
```

The predicted participation rate stands above a hundred which is unreasonable. **potential issues**

- The value used is outside the fitted data range which may lead to unreasonable predictions
- It could also be possible that the relation between the regressor and the dependent variable is in fact not linear. (misspecification)

1.7 Question 6

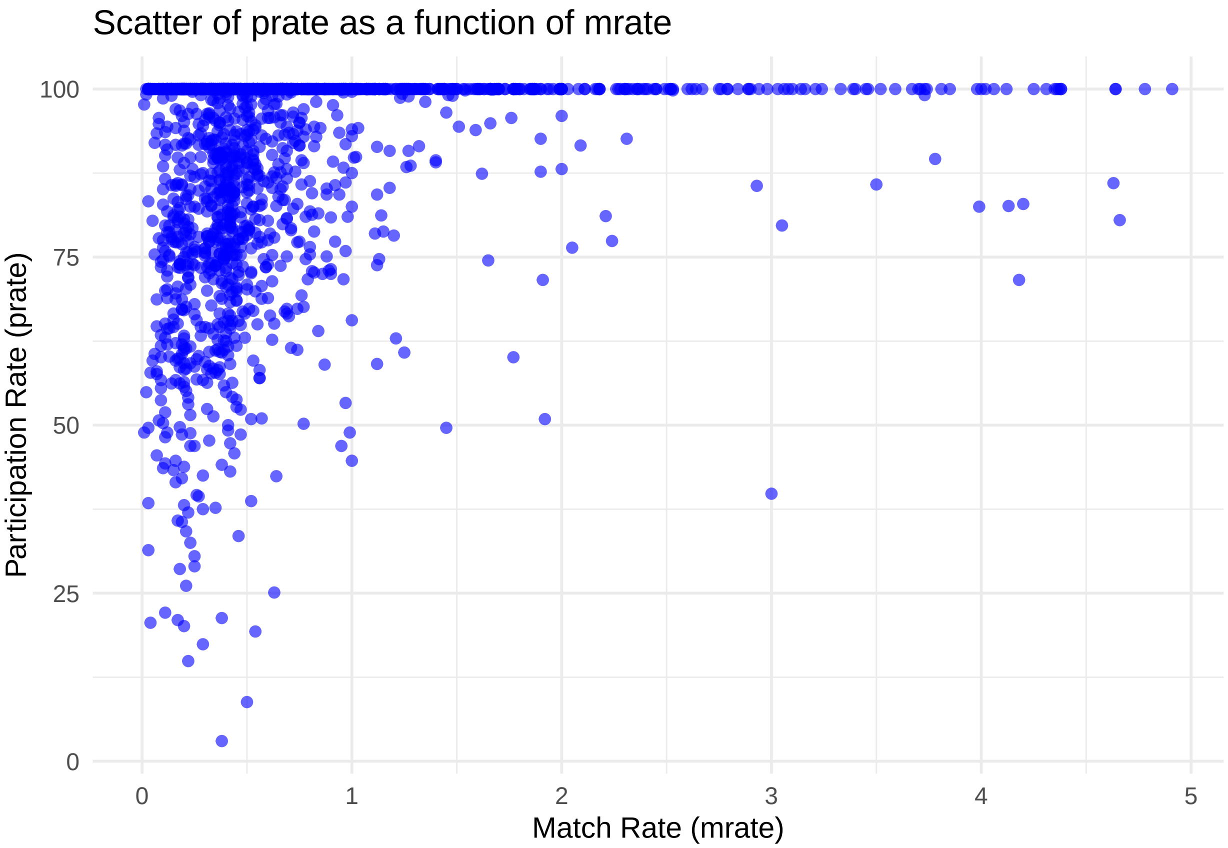
```r
cat("The match rate explains", sprintf("%.2f", r_squared*100),"% of the variation in
 ↪  participation rate")
```

```
The match rate explains 7.47 % of the variation in participation rate
```

This seems like a rather small number given the fact the the match rate indicates how much the firm is willing to contribute to the pension plan. This means more than 90% of the variation in prate is attributable to other factors. Nonetheless, it does not mean that the match rate is not significantly related to the participation rate.

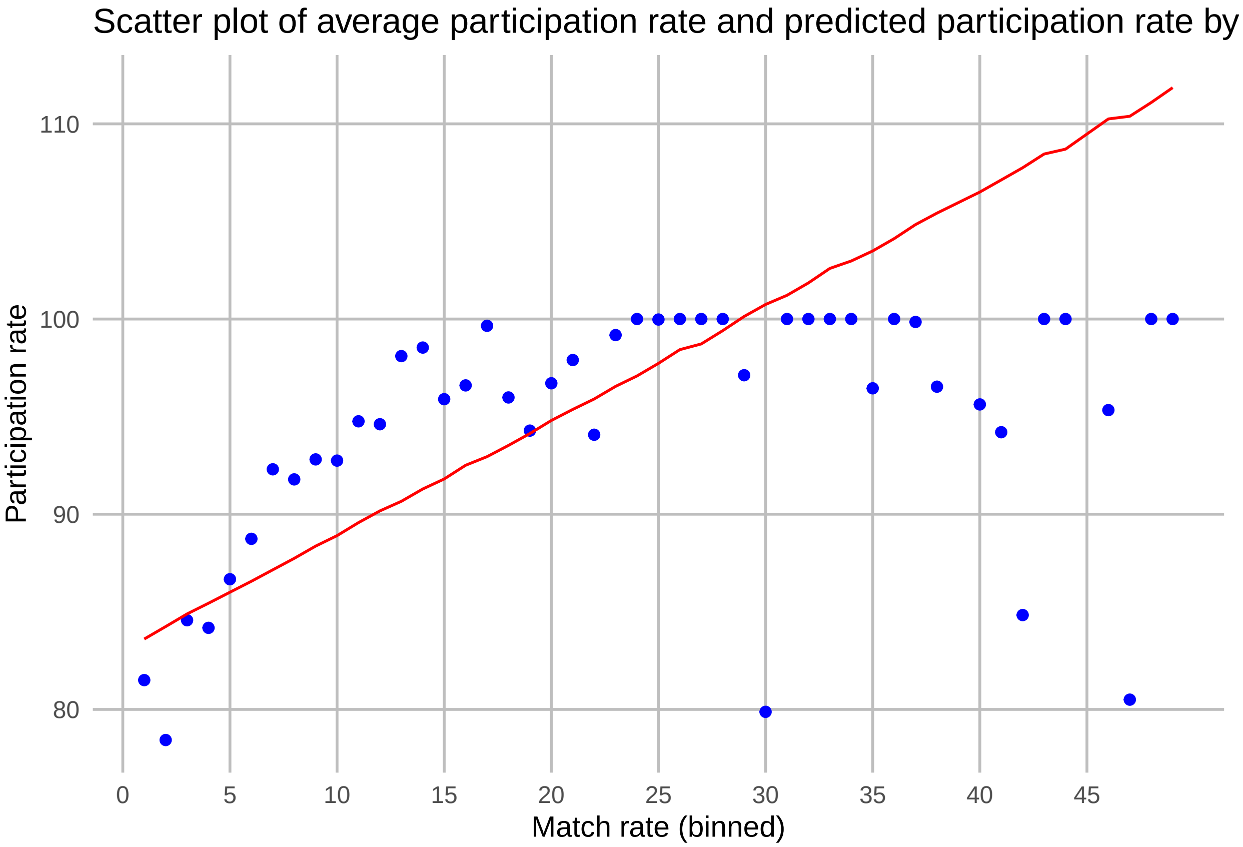
1.8 Question 7

```r
ggplot(data, aes(x = mrate, y = prate)) +
  geom_point(color = "blue", alpha = 0.6) +
  labs(
    title = "Scatter of prate as a function of mrate",
    x = "Match Rate (mrate)",
    y = "Participation Rate (prate)"
  ) +
  theme_minimal()
```



The Scatter plot seems to suggest a positive, non-linear relation between the participation rate and the match rate;

```
# Creates column with binned values
data <- data %>% mutate(rnd_mrate = round(mrate, digits = 1),
                        mrate_binned = cut(rnd_mrate,
                                           breaks = seq(min(rnd_mrate, na.rm = TRUE),
                                                        max(rnd_mrate, na.rm = TRUE),
                                                        by = 0.1),
                                           include.lowest = TRUE))

# Calculate the average prate by bin
Average_mrate_df <- data %>%
  group_by(mrate_binned) %>%
  summarise(avg_prate_by_bin = mean(prate, na.rm = TRUE))

# Add predicted values to the dataset
data <- data %>% mutate(predicted_values = predict(model))

# Calculate the average predicted prate by bin
Predicted_mrate_df <- data %>%
  group_by(mrate_binned) %>%
  summarise(predicted_prate_by_bin = mean(predicted_values, na.rm = TRUE))

# Merge the two dataframes for plotting
plot_data <- Average_mrate_df %>%
  left_join(Predicted_mrate_df, by = "mrate_binned")

# Create the scatterplot with average prate and predicted prate
ggplot() +
  # Scatterplot for average prate
  geom_point(data = Average_mrate_df, aes(x = as.numeric(mrate_binned), y =
  ↪  avg_prate_by_bin), color = "blue") +
  # Line for predicted prate
  geom_line(data = Predicted_mrate_df, aes(x = as.numeric(mrate_binned), y =
  ↪  predicted_prate_by_bin), color = "red") +
  theme_minimal() +
  labs(title = "Scatter plot of average participation rate and predicted participation
  ↪  rate by match rate",
       x = "Match rate (binned)",
       y = "Participation rate") +
  scale_x_continuous(breaks = seq(0, max(as.numeric(Average_mrate_df$mrate_binned), na.rm
  ↪  = TRUE), by = 5)) +
  theme(panel.grid.major = element_line(color = "gray", linewidth = 0.5),
        panel.grid.minor = element_blank())
```



We notice clearly the positive relation between match rate and participation rate. However, based on the scatterplot, this relation is not linear and there remains a few outliers. Thus we can establish that the problem that our model suffers from is a misspecification problem.