Problem Set 04
Matching Methods in Employee Performance Analysis

TAs: Aida Hatami, Sam Fathinejad, Fatemeh Salehi, Reza Sahour

March 23, 2025

This problem set uses the dataset **Employee_Performance.csv**, which provides information on employees, including demographics, work history, job performance, and job satisfaction. The objective is to estimate causal effects in an observational setting using matching methods. The methods applied in this analysis include **Propensity Score Matching (PSM)**, **Exact Matching**, **Nearest Neighbor Matching**, **Mahalanobis Matching**, and **Inverse Probability Weighting (IPW)**.

Section 1: Conceptual questions with a Case Study

Scenario: You are part of a research team at a consulting firm tasked with evaluating whether a new training program for mid-level employees improves their productivity and satisfaction. Since the firm did not randomly assign employees to the training program, your goal is to use matching methods to estimate the causal effect.

The dataset includes: age, gender, department, job level, number of projects, education level, training hours, satisfaction score, productivity score, team ID, and resignation status.

Instructions: Answer the following questions based on this scenario. Use formal concepts where possible and justify your reasoning.

(a) Describe the main threats to causal identification in this setting. Which ones can be addressed through matching?

(b) Explain the Conditional Independence Assumption (CIA). Provide two examples of covariates that might violate CIA if unobserved and explain why.

(c) What is the overlap (common support) assumption? How would you detect and visualize violations of this assumption in practice?

(d) Suppose you estimate propensity scores and observe a very high concentration of scores close to 0 or 1. What does this indicate? How would it impact your analysis?

(e) Define and contrast ATE, ATT, and ATU. Which estimand is most appropriate in this context if the firm is only interested in the effect of training on those who received it?

(f) You consider using exact matching on age, education, and department. Discuss the feasibility and statistical consequences of this approach in high-dimensional settings. Relate your answer to the curse of dimensionality.

(g) If some treated units fall outside the region of common support, what would be a principled approach to handle this problem? What trade-offs are involved?

(h) Discuss how you would evaluate covariate balance. Provide two quantitative metrics and one visual method.

(i) Suppose after matching you observe that treated and control groups still differ significantly in job level and prior performance score. What would you do to address this?

(j) Explain how poor overlap and poor balance can lead to biased treatment effect estimates even after matching. Use causal diagrams or conceptual logic if helpful.

(k) Compare Inverse Probability Weighting (IPW) with Stabilized Inverse Probability Weighting (SIPW). What problem does SIPW address in traditional IPW, and how is it implemented? In practice, when would you prefer SIPW over IPW? Provide an intuitive explanation using an example where propensity scores are close to 0 or 1.

Section 2: The Effect of Training on Salary

Binary Treatment and Propensity Score Matching

Background: In this section, we evaluate the effect of participating in training (represented by the variable `treatment`) on employee salary. This treatment indicator has already been constructed based on employee characteristics and approximates the likelihood of receiving training in a real-world observational setting.

(a) Examine the distribution of the variable `treatment`. What proportion of the employees are treated? Then, estimate the propensity score for this treatment assignment using logistic regression with the following covariates: `age`, `years_at_company`, `projects_handled`, and `remote_work_frequency`. Present summary statistics and plot the distribution of the estimated propensity scores separately for the treated and control groups.

(b) Apply Propensity Score Matching (PSM) using 1-to-1 nearest neighbor matching with a caliper of 0.2 and no replacement. Assess covariate balance before and after matching using:

- Standardized Mean Differences (SMD)
- Graphical diagnostics (PS histograms, bias plots)

Evaluate whether balance has improved and whether common support holds.

(c) Using the matched sample, estimate the Average Treatment Effect on the Treated (ATT) in two different ways:

(i) Compute the difference in average `monthly_salary` between treated and control employees in the matched sample.

(ii) Run an OLS regression of `monthly_salary` on the matching covariates (`age`, `years_at_company`, `projects_handled`, `remote_work_frequency`) *without* including the `treatment` variable.

Compare the results from (i) and (ii) and discuss: To what extent is the variation in salary explained by covariates alone? What does this imply about the strength and role of the treatment?

(d) On the full sample (without matching), estimate a regression of `monthly_salary` on `treatment` and the same covariates as in part (c). Compare the results with the matched sample. Comment on bias and efficiency.

(e) Estimate a regression using `training_hours` as a continuous variable instead of `treatment`. Interpret the coefficient. Does this suggest linear gains in salary with training intensity, or is there evidence of diminishing returns?

Section 3: Nearest Neighbor vs. Mahalanobis Matching

The Effect of Remote Work Frequency on Employee Outcomes

Background: This section investigates how different levels of remote work (`remote_work_frequency`) affect employee performance and retention. You will apply and compare two matching techniques: **Nearest Neighbor Matching (NNM)** and **Mahalanobis Distance Matching**.

(a) Define treatment as a binary variable: **Treated = employees in the top 25% of `remote_work_frequency`, Control = bottom 25%**. Drop the middle 50% of the sample. Estimate propensity scores using the covariates: `age`, `education_level`, `years_at_company`, and `overtime_hours`. Display summary statistics and histograms of the propensity scores for treated and control groups.

(b) Apply 1-to-1 **Nearest Neighbor Matching (NNM)** using a caliper of 0.2 and no replacement. Assess covariate balance before and after matching using `pstest` and graphical diagnostics. Comment on the success of matching.

(c) Apply **Mahalanobis Distance Matching** using fewer covariates to reduce runtime (e.g., `age`, `years_at_company`, `projects_handled`). Use the `teffects nnmatch` command with the `mahalanobis` option. If your code runs slowly, you may use a random sample . Evaluate covariate balance and compare matched sample size with part (b).

(d) For both matching methods, estimate the effect of high remote work on:

- `performance_score` (OLS regression)
- `resigned` (logistic regression)

Compare the size, sign, and significance of the treatment effect across the two methods.

(e) Discuss which matching method is more appropriate in this context and explain why. What are the key assumptions behind each method and how do they impact your interpretation?

Section 4: Exact Matching and Inverse Probability Weighting (IPW)

Effect of Multi-level Promotions on Employee Satisfaction

Background: Promotions in this dataset are not binary but rather have three distinct levels:

- `0`: No promotion
- `1`: Partial/intermediate promotion
- `2`: Full promotion

Our goal is to estimate the causal effect of different levels of promotion on employee satisfaction. Since standard IPW assumes binary treatment, we extend the methodology in two ways: **(i)** by conducting pairwise comparisons between levels, and **(ii)** using multinomial logistic regression to model the treatment assignment.

Inverse Probability Weighting (IPW): When comparing two levels of treatment (e.g., Level 2 vs. Level 0), we restrict the sample and apply binary IPW. The IPW estimator in this case is:

$$\hat{\tau}_{HT} = \frac{1}{n}\sum_{i=1}^{n}\frac{D_i Y_i}{\pi_i(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-D_i)Y_i}{1-\pi_i(X_i)}$$

where $D_i = 1$ if unit i received treatment (e.g., promotion level 2) and 0 otherwise. The propensity score $\pi_i(X_i)$ is estimated via logistic regression on covariates X_i.

Stabilized IPW (SIPW):

$$\hat{\tau}_{SIPW} = \frac{\frac{1}{n}\sum_i \frac{Y_i D_i}{\hat{\pi}(X_i)}}{\frac{1}{n}\sum_i \frac{D_i}{\hat{\pi}(X_i)}} - \frac{\frac{1}{n}\sum_i \frac{Y_i(1-D_i)}{1-\hat{\pi}(X_i)}}{\frac{1}{n}\sum_i \frac{(1-D_i)}{1-\hat{\pi}(X_i)}}$$

This version improves stability in finite samples.

(a) Exact Matching:
Using `age`, `edu_level`, and `years_at_company` as matching variables, perform exact matching across the three promotion levels. For each promoted employee at level 1 or 2, find a match in level 0 with the same covariate profile. Report the number of matched triplets or matched sets. Discuss how strictness in criteria affects the matched sample size.

(b) Extended Exact Matching:
Repeat the matching from part (a) but include `overtime_hours` and `projects_handled` as additional covariates. Compare the size of matched sets to part (a) and discuss trade-offs between match quality and available sample size.

(c) IPW and SIPW Estimation:
Estimate the effect of promotion level on `employee_satisfaction_score` using two approaches:

- **(i) Pairwise IPW/SIPW:** For each pair of promotion levels (0 vs 1, 0 vs 2, and 1 vs 2), estimate binary propensity scores via logistic regression using covariates: `age`, `edu_level`, `years_at_company`, `overtime_hours`, and `projects_handled`. Compute IPW and SIPW weights and estimate average treatment effects for each pair.
- **(ii) Multinomial IPW/SIPW:** Fit a multinomial logistic model for promotion with the same covariates. Use the estimated probabilities to compute weights and run a weighted OLS regression of satisfaction on dummy-coded promotion levels. Report coefficients for each level and compare to pairwise estimates.

Interpret the results and discuss how IPW/SIPW estimates compare with those from exact matching.

Section 5: The Effect of Team Training Spillovers

Performance Gains for Untrained Employees

Background: In a collaborative work environment, employees may improve not only from formal training they receive, but also from working alongside trained teammates. Such *peer effects* or *spillovers* may occur through informal learning, shared routines, or enhanced team processes. In this section, we assess whether employees who themselves did not receive training perform better when embedded in highly trained teams.

Setup: We focus on employees with `training_hours` = 0. The treatment variable `team_treatment` is a binary indicator equal to 1 if the employee belongs to a team with above-median training intensity. The outcome is `performance_score`.

(a) **Conceptual Motivation:** Explain how untrained employees might benefit from trained teammates. Identify at least two plausible channels through which peer effects may arise (e.g., peer learning, workflow improvements, shared standards).

(b) **Empirical Strategy:** Restrict the sample to individuals with `training_hours` = 0. Estimate the effect of `team_treatment` on `performance_score` using Inverse Probability Weighting (IPW). Proceed as follows:

  - Recode `job_title` into four categories: Analyst, Technical (Developer/Engineer/Technician), Consultant, and Specialist/Manager.
  - Estimate a logistic model for `team_treatment` using the covariates: `age`, `edu_level`, `job_title_category`, `years_at_company`, and `overtime_hours`.
  - Use the estimated propensity scores to compute IPW weights.
  - Run a weighted OLS regression of `performance_score` on `team_treatment`, using the computed weights.

  Report and interpret the estimated coefficient. Is there evidence of performance gains from indirect training exposure?

(c) **Implications for HR Policy:** If peer spillovers are statistically or practically significant, how should training investments be allocated?

  - Should firms train a few key team members or aim for broader coverage?

- How does the composition of job roles within a team affect knowledge diffusion?
- What organizational structures (e.g., team size, leadership, collaboration culture) might strengthen or weaken peer effects?