Matching

一、核心理论与概念

1. 条件独立假设（CIA）

- **定义**：$(Y^1, Y^0) \perp D|X$，即给定观测变量 X，潜在结果 Y^1（处理组）和 Y^0（控制组）与处理变量 D 独立。

  Definition: $(Y^1, Y^0) \perp D|X$, which means that given the observed variables X, the potential outcomes Y^1 (treatment group) and Y^0 (control group) are independent of the treatment variable D.

- **含义**：在 X 的每个取值下，处理组和控制组的潜在结果均值相等，即：

$$E[Y^1|D=1,X] = E[Y^1|D=0,X], \quad E[Y^0|D=1,X] = E[Y^0|D=0,X]$$

  Implication: At each value of X, the mean of potential outcomes for the treatment group and the control group is equal.

- **作用**：满足 CIA 意味着通过控制 X 可消除选择偏差，满足 “后门准则”，是推断因果关系的关键假设。

  Function: Satisfying the CIA implies that by controlling for X, selection bias can be eliminated, meeting the "back - door criterion", which is a key assumption for inferring causal relationships.

- **案例应用**：吸烟与肺癌研究中，若不控制年龄、收入等变量，吸烟者与非吸烟者的肺癌死亡率差异可能受混淆因素影响；控制 X 后，可认为吸烟与肺癌的关联更接近因果关系。

  Case Application: In a study of smoking and lung cancer, if variables such as age and income are not controlled, the difference in lung cancer mortality rates between smokers and non - smokers may be affected by confounding factors. After controlling for X, the association between smoking and lung cancer can be considered closer to a causal relationship.

2. 匹配方法的核心逻辑

- **目标**：通过构造处理组与控制组的 “可交换性”（即 covariates 平衡），估计处理效应（如平均处理效应 ATE、处理组平均处理效应 ATT）。

  Objective: By constructing the "exchangeability" (i.e., covariate balance) between the treatment group and the control group, estimate the treatment effects (such as the Average Treatment Effect (ATE) and the Average Treatment Effect on the Treated (ATT)).

- **关键问题**：观测数据中处理组和控制组的 covariates 不平衡（如吸烟者更年长），导致简单均值差有偏。

  Key Problem: In the observational data, the covariates of the treatment group and the control group are unbalanced (e.g., smokers are older), which leads to a biased simple mean difference.

- **解决思路**：通过分层、加权或匹配，使处理组和控制组在 X 上分布一致。

  Solution: Through stratification, weighting, or matching, make the distributions of the treatment group and the control group on X consistent.

二、主要方法与案例

1. 子分类（Subclassification）

- **步骤**：

1. 将 X 分层（如按年龄分为 20-40 岁、41-70 岁等）。
2. 计算各层内处理组和控制组的结果均值。
3. 按控制组的层权重加权处理组均值，得到调整后的死亡率（如年龄调整后的吸烟死亡率）。

- **案例**：Cochran（1968）研究吸烟与死亡率，原始数据显示雪茄 / 烟斗吸烟者死亡率最高，但调整年龄后，香烟吸烟者死亡率反而是最高的，因雪茄吸烟者平均年龄更大。
- **公式**：年龄调整后死亡率 = $\sum$（各年龄层死亡率 × 控制组对应年龄层比例）。
- **Steps**:

1. Stratify X (e.g., divide by age into 20 - 40 years old, 41 - 70 years old, etc.).
2. Calculate the mean outcomes of the treatment group and the control group within each stratum.
3. Weight the mean of the treatment group by the stratum weights of the control group to obtain the adjusted mortality rate (e.g., age - adjusted smoking mortality rate).

- **Case**: Cochran (1968) studied smoking and mortality. The original data showed that cigar/piped smokers had the highest mortality rate, but after age adjustment, cigarette smokers had the highest mortality rate because cigar smokers were on average older.
- **Formula**: Age - adjusted mortality rate = $\sum$ (Mortality rate in each age stratum × Proportion of the corresponding age stratum in the control group).

2. 精确匹配（Exact Matching）

- **定义**：为处理组每个单位找到控制组中 X 完全相同的单位，用其结果作为反事实。
- **案例**：职业培训项目中，处理组某学员年龄 18 岁，在控制组中找年龄 18 岁的非学员，用其收入作为该学员若未参加培训的反事实收入。
- **局限性**：当 X 维度高时，难以找到精确匹配（维度灾难），如同时匹配年龄、收入、教育程度时，样本可能稀疏。
- **Definition**: For each unit in the treatment group, find a unit in the control group with exactly the same X, and use its outcome as the counterfactual.
- **Case**: In a vocational training program, if a trainee in the treatment group is 18 years old, find a non - trainee of the same age in the control group, and use their income as the counterfactual income of the trainee if they had not participated in the training.
- **Limitation**: When the dimension of X is high, it is difficult to find an exact match (the curse of dimensionality). For example, when matching age, income, and education level simultaneously, the sample may be sparse.

3. 近似匹配（Approximate Matching）

- **距离度量**：
  - **欧氏距离**Euclidean Distance：$\|X_i - X_j\| = \sqrt{\sum (X_{ni} - X_{nj})^2}$。
  - **标准化欧氏距离**Standardized Euclidean Distance：考虑变量方差，避免量纲影响。
  - **马氏距离**Mahalanobis Distance：考虑变量协方差，如 $\|X_i - X_j\| = \sqrt{(X_i - X_j)' \hat{\Sigma}_X^{-1} (X_i - X_j)}$
- **案例**：泰坦尼克号生存分析中，按年龄和性别匹配，计算匹配后的 ATE，从原始 35.4% 降至 16.1%，因控制了 “女性和儿童优先获救” 的混淆因素。
- **Case**: In the survival analysis of the Titanic, matching by age and gender and calculating the matched ATE. The original ATE was 35.4%, which decreased to 16.1% after matching because the confounding factor of "women and children first" was controlled.

4. 倾向得分方法（Propensity Score）

- **定义**：倾向得分 $p(X) = Pr(D = 1|X)$，即给定 X 时接受处理的概率，可通过 logit/probit 模型估计。
- **核心定理**：若满足 CIA，则 $(Y^1, Y^0) \perp D|p(X)$，即只需控制 $p(X)$ 即可平衡所有 X。
- **应用步骤**：

1. 估计 $p(X)$（如用 logit 模型）。
2. 检查共同支撑（$0 < p(X) < 1$，确保各 $p(X)$ 区间均有处理组和控制组）。
3. 用倾向得分匹配或加权估计处理效应。

- **案例**：NSW 职业培训项目中，非实验控制组与处理组 covariates 差异大（如处理组更年轻、教育程度更低），用倾向得分匹配后，估计的培训效应从负转正，接近实验真实值（$1794）。
- **Definition**: The propensity score $p(X) = Pr(D = 1|X)$, which is the probability of receiving treatment given X, and can be estimated by a logit/probit model.
- **Core Theorem**: If the CIA is satisfied, then $(Y^1, Y^0) \perp D|p(X)$, that is, only $p(X)$ needs to be controlled to balance all X.
- **Application Steps**:

1. Estimate $p(X)$ (e.g., using a logit model).
2. Check the common support ($0 < p(X) < 1$, ensuring that there are both treatment and control group units in each $p(X)$ interval).
3. Use propensity score matching or weighting to estimate the treatment effect.

- **Case**: In the NSW vocational training program, there were large differences in covariates between the non - experimental control group and the treatment group (e.g., the treatment group was younger and had a lower education level). After propensity score matching, the estimated training effect changed from negative to positive, approaching the experimental true value ($1794).

三、关键假设与常见问题

1. 共同支撑（Common Support）

- **定义**：$0 < Pr(D=1|X) < 1$，即对每个 X，同时存在处理组和控制组单位。
- **作用**：若无共同支撑，无法计算有效权重（如某 $p(X)$ 区间只有处理组，无控制组，无法匹配）。
- **案例**：泰坦尼克号数据中，12 岁男性头等舱乘客无匹配对象，因控制组中无该年龄层，需转而估计 ATT（处理组平均效应）。
- **Definition**: $0 < Pr(D=1|X) < 1$, which means that for each X, there are both treatment and control group units.
- **Function**: Without common support, effective weights cannot be calculated (e.g., if there is only a treatment group and no control group in a certain $p(X)$ interval, matching cannot be done).
- **Case**: In the Titanic data, there was no matching object for 12 - year - old male first - class passengers because there was no such age group in the control group. In this case, the ATT (average treatment effect on the treated) needs to be estimated instead.

2. 维度灾难（Curse of Dimensionality）

- **问题**：当匹配变量 X 维度高时，样本稀疏，难以找到匹配对象，导致估计偏差。
- **解决**：用倾向得分将多维度 X 压缩为单维度 $p(X)$，降低维度。
- **Problem**: When the dimension of the matching variable X is high, the sample is sparse, making it difficult to find matching objects and leading to estimation bias.
- **Solution**: Use the propensity score to compress the multi - dimensional X into a single - dimensional $p(X)$ to reduce the dimension.

3. 偏差校正（Bias Correction）

- **原因**：近似匹配中 X_i 与 X_j 不完全相同，导致 $\mu^0(X_i) - \mu^0(X_j)$ 非零，引入偏差。
- **方法**：用回归估计 $\hat{\mu}^0(X)$，调整匹配差异，如：
  $\hat{\delta}^{BC}_{ATT} = \frac{1}{N_T} \sum[(Y_i - Y_j) - (\hat{\mu}^0(X_i) - \hat{\mu}^0(X_j))]$
- **Reason**: In approximate matching, X_i and X_j are not exactly the same, resulting in $\mu^0(X_i) - \mu^0(X_j)$ being non - zero and introducing bias.
- **Method**: Use regression to estimate $\hat{\mu}^0(X)$ and adjust the matching difference.

四、考试重点与答题思路

1. 概念辨析（可能出简答题）

- 对比 CIA 与独立性假设：CIA 是条件独立（给定 X），独立性假设是无条件独立，CIA 更现实。
- 解释 “平衡”：处理组和控制组的 covariates 均值相等，如匹配后年龄均值相同。
- Compare the CIA with the independence assumption: The CIA is conditional independence (given X), while the independence assumption is unconditional independence. The CIA is more realistic.
- Explain "balance": The means of covariates in the treatment group and the control group are equal, such as the same mean age after matching.

2. 案例分析（可能出应用题）

- **例题**：某研究想考察 “大学教育是否提高收入”，但发现大学生更可能来自高收入家庭。如何设计计量模型？
- **答题思路**：

1. 指出混淆因素：家庭收入（X）同时影响教育选择（D）和收入（Y）。
2. 应用 CIA：假设 $(Y^1, Y^0) \perp D|X$，即控制家庭收入后，教育与潜在收入独立。
3. 方法选择：用倾向得分匹配（估计 $p(X) = Pr(D = 1|X)$），或子分类（按家庭收入分层）。
4. 数据处理：若家庭收入是连续变量，用 logit 估计倾向得分，再匹配；若离散，直接分层。
5. 结果解释：匹配后比较大学生与非大学生的收入差，即为教育的因果效应。

- **Example Question**: A study wants to examine whether a college education increases income, but finds that college students are more likely to come from high - income families. How to design an econometric model?
- **Answering Ideas**:

1. Identify the confounding factor: Family income (X) affects both the education choice (D) and income (Y).
2. Apply the CIA: Assume that $(Y^1, Y^0) \perp D|X$, that is, after controlling for family income, education is independent of potential income.
3. Method selection: Use propensity score matching (estimate $p(X) = Pr(D = 1|X)$) or subclassification (stratify by family income).
4. Data processing: If family income is a continuous variable, use logit to estimate the propensity score and then match; if it is discrete, directly stratify.
5. Result interpretation: Compare the income difference between college students and non - college students after matching, which is the causal effect of education.