HW2: Introduction to Causal Inference

Problem 1 (30'):

Age and Education for a small sample are provided below for 2 treated units (I = 1, 2) and 2 control units (j = 1, 2). Both covariates are predictive of the outcome of Income (in $10k).

	Age (in years)	Edu (=1 if Master+; 0 otherwise)	Income (=$10k)
Treated i=1	25	1	15
Treated i=2	30	1	22
Control j=1	30	0	10
Control j=2	40	1	15

The covariance matrix between age and education is given $\sum_{AGE,EDU} = \begin{pmatrix} 10 & 0.2 \\ 0.2 & 1 \end{pmatrix}$.

a) (10') Conduct **optimal** matching to find the matched unit j to the treated unit i, denoted by *j(i),* using Mahalanobis distance. Fill in the last column in the table below. (Hint, in R the function "*solve(matrix(c(a,b,c,d),2,2))*" to solve for $\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1}$; For matrix multiplication in R, please use "*%*%*"). Provide computation details.

Matching Pair	Treated i	Control j(i)
1	*i* =1	
2	*i* =2	

b) (5') Estimate ACE using the matched pair.

Propensity of being treated e(x) for each unit is estimated and provided in the table below.

c) (10') Construct propensity score weights and fill in the last two columns in the table below.

	e(x)	Income (=$10k)	PS weight (w)	Income*w
Treated i=1	0.25	15		
Treated i=2	0.4	22		
Control j=1	0.33	10		
Control j=2	0.5	15		

d) (5') Estimate the average causal effect using the measure of risk difference (RD).

Problem 2 (35')

Apply propensity score methods to assess the causal effect of a treatment (New_Medication) on an outcome (Heart_Disease_Incident) using both propensity score matching and inverse probability weighting (IPW) methods. This problem reinforces concepts like confounding adjustment, balance assessment, and interpretation of causal effects. A sample of an observational study on cardiovascular disease for 500 patients, including several covariates (Age, Sex, BMI, Blood Pressure (BP), and Diabetes), along with a binary treatment variable (New_Medication) and outcome variable (Heart_Disease_Incident), is collected and saved under File on CANVAS. Ensure you have these R packages installed and loaded, including tableone, Matching, survey, ipw.

a) (5') Descriptive Statistics and Covariate Balance

   1) Create a Table 1 for all covariates by New_Medication status, report both unadjusted means and standardized mean differences (SMDs).
   2) Identify covariates with SMD > 0.2 (indicating imbalance) and comment on how covariate imbalances could bias the treatment effect estimate.

b) (15') Propensity Score Matching

   1) Run logistic regression to estimate the propensity score (probability of receiving New_Medication given all five covariates).
   2) Match each treated unit to a control unit with a similar propensity score using 1:1 nearest-neighbor matching without replacement.
   3) Generate a new Table 1 after matching to examine covariate balance.
   4) Compare Heart_Disease_Incident between matched treatment and control groups using a paired t-test. Interpret the results.

c) (15') Inverse Probability Weighting (IPW)

   1) Construct IPW for each unit based on the estimated propensity scores obtained in Part b1) (Hint: Weight treated units by 1/PS and control units by 1/(1 - PS)).
   2) Using the survey package, assess covariate balance in the weighted dataset by creating a Table 1 with SMDs.
   3) Compare Heart_Disease_Incident between weighted treatment and weighted control groups by using a weighted regression model to estimate the effect of New_Medication. Interpret the results.

Please answer questions in each part, including SMD tables if requested. Provide R code for each part in an appendix.

Problem 3 (35')

The crude RD, RR and OR can be expressed in terms of the parameters of the following linear, log linear and linear logistic models for the observed outcome Y:

$$pr(Y = 1|A = a) = \psi'_0 + \psi'_1 a$$

$$\log pr(Y = 1|A = a) = \theta'_0 + \theta'_1 a$$

$$\operatorname{logit} pr(Y = 1|A = a) = \beta'_0 + \beta'_1 a$$

The crude *associational* RD equals ψ_1', RR equals $e^{\theta_1'}$, and OR equals $e^{\beta_1'}$. Assuming treatment is unconfounded, these same estimates will also be unbiased for the corresponding *causal* parameters of the following MSMs:

$$pr(Y^a = 1) = \psi_0 + \psi_1 a$$

$$\log pr(Y^a = 1) = \theta_0 + \theta_1 a$$

$$\operatorname{logit} pr(Y^a = 1) = \beta_0 + \beta_1 a$$

Using the data below

	$L_0 = 1$		$L_0 = 0$	
	$A_0 = 1$	$A_0 = 0$	$A_0 = 1$	$A_0 = 0$
$Y = 1$	108	24	20	40
$Y = 0$	252	16	30	10
Total	360	40	50	50

a) (10') Estimate the causal RD, RR and OR by *standardization* method.
b) (10') MSM estimation involves creation of weights. Create weights and stabilized weights using the data above.
c) (15') Estimate the causal RD, RR and OR by R (or SAS and STATA if you preferred) using MSM method with the model "Y ~ A" where the outcome Y is specified as a binary variable. To estimate ψ_1', one would specify the *identity* link; to estimate θ_1', the *log* link; and for β_1', the *logit* link.