PS C236A / Stat C239A
Problem Set 3 - Solutions

1) The ATT is defined as:

$$E(r_{i1} - r_{i0}|Z_i = 1).$$

We claim that if we assume

$$\begin{aligned} (r_{i0})_{i=1}^N &\perp\!\!\!\perp Z|X_{Z=1} \\ P(Z=1|X_{Z=1}) &< 1 \end{aligned}$$

then conditioning by the propensity score will make estimates of the ATT unbiased.

To clear up notation:

$$\begin{aligned} Z &\equiv (Z_i)_{i=1}^N \\ \{Z=1\} &\equiv \{i : Z_i = 1\} \\ X_{Z=1} &\equiv (X_i)_{i\in\{Z=1\}} \\ \mathcal{I}_p^1 &\equiv \{i : e(X_i) = p, Z_i = 1\} \\ X_{Z=1}\{p\} &\equiv (X_i)_{i\in\mathcal{I}_p^1} \end{aligned}$$

If our population is large, with some of the population assigned to treatment, and the rest assigned to control, we can estimate the ATT by

$$E_{e(X_{Z=1})}[E[r_{i1}|Z_i = 1, X_{Z=1}\{e(X_{Z=1})\}] - E[r_{i0}|Z_i = 0, X_{Z=1}\{e(X_{Z=1})\}]|e(X_{Z=1})].$$

In words, we first fix a propensity score p. For both treatment and control responses, we compute the expectation of all responses such that the propensity score of their covariates is p. We then take the expectation of the difference of these expectations over the distribution of $e(X_{Z=1})$.
Note that the expectation

$$E[r_{i0}|Z_i = 0, X_{Z=1}\{e(X_{Z=1})\}]$$

is well defined under the assumption that

$$P(Z=1|X_{Z=1}) < 1.$$

Simplifying the ATT expression, we obtain:

$$\begin{aligned} &E_{e(X_{Z=1})}[E[r_{i1}|Z_i = 1, X_{Z=1}\{e(X_{Z=1})\}] - E[r_{i0}|Z_i = 0, X_{Z=1}\{e(X_{Z=1})\}]|e(X_{Z=1})] \\ =&E_{e(X_{Z=1})}[E[r_{i1}|Z_i = 1, X_{Z=1}\{e(X_{Z=1})\}, e(X_{Z=1})]] \\ &- E_{e(X_{Z=1})}[E[r_{i0}|Z_i = 0, X_{Z=1}\{e(X_{Z=1})\}, e(X_{Z=1})]] \\ =&E_{e(X_{Z=1})}[E[r_{i1}|Z_i = 1, X_{Z=1}\{e(X_{Z=1})\}]] - E_{e(X_{Z=1})}[E[r_{i0}|Z_i = 0, X_{Z=1}\{e(X_{Z=1})\}]] \\ =&E_{X_{Z=1}}[E[r_{i1}|Z_i = 1, X_{Z=1}]] - E_{X_{Z=1}}[E[r_{i0}|Z_i = 0, X_{Z=1}]] \end{aligned}$$

The outer expectation in the last expression is now over the distribution $X_{Z=1}$.
The third equality follows from $e(X_{Z=1})$ being a function of $X_{Z=1}$, and the last equality follows from

$$X_{Z=1} = (X_{Z=1}\{e(X_{Z=1})\})_{0\leq e(X_{Z=1})\leq 1}.$$

Now, since

$$(r_{i0})_{i=1}^{N} \perp\!\!\!\perp Z | X_{Z=1},$$

we obtain

$$\begin{aligned}
&E_{X_{Z=1}}[E[r_{i1}|Z_i=1, X_{Z=1}]] - E_{X_{Z=1}}[E[r_{i0}|Z_i=0, X_{Z=1}]] \\
=&E_{X_{Z=1}}[E[r_{i1}|Z_i=1, X_{Z=1}]] - E_{X_{Z=1}}[E[r_{i0}|Z_i=1, X_{Z=1}]] \\
=&E[r_{i1}|Z_i=1] - E[r_{i0}|Z_i=1] \\
=&E[r_{i1}-r_{i0}|Z_i=1] = \text{ATT}.
\end{aligned}$$

Thus, when conditioning on the propensity score, the estimate of the ATT is unbiased (asymptotically).
To obtain an unbiased ATE, we will need to modify the original assumptions to

$$\begin{aligned}
(r_{i0})_{i=1}^{N}, (r_{i1},)_{i=1}^{N} \perp\!\!\!\perp Z | X_{Z=1} \\
0 < P(Z=1|X_{Z=1}) < 1
\end{aligned}$$

This is the strongly ignorable assumption in Rosenbaum and Rubin (1983).

2) a) Note that these p_i we want to estimate are propensity scores. From Rosenbaum and Rubin, the probability of observing a given treatment assignment is

$$\prod_{i=1}^{10000} e(X_i)^{T_i}(1-e(X_i))^{T_i}$$

where $e(X_i)$ is the propensity score given observed covariates X_i. From our problem statement, we know that $e(X_i)$ is a function of sex, exercising more than 30 minutes a day, and watching TV for more than an hour a day. Let S_i, E_i, and V_i denote indicator variables for these covariates. It follows that the probability of our treatment assignment is:

$$\begin{aligned}
&\prod_{i=1}^{10000} e(X_i)^{T_i}(1-e(X_i))^{1-T_i} = \prod_{i=1}^{10000} e(S_i,E_i,V_i)^{T_i}(1-e(S_i,E_i,V_i))^{1-T_i} \\
=& \prod_{(s,e,v)\in\{0,1\}^3} e(s,e,v)^{\#(s,e,v,t)}(1-e(s,e,v))^{\#(s,e,v)-\#(s,e,v,t)} \qquad (1)
\end{aligned}$$

Here, s, e, and v, all take values equal to 0 or 1; $\#(s,e,v)$ denotes the number of units that have covariate indicators $S_i = s$, $E_i = e$, and $V_i = v$; and $\#(s,e,v,t)$ denotes the number of treated units that have those values of indicator variables.
We take the log of (1) to obtain:

$$\sum_{(s,e,v)\in\{0,1\}^3} \#(s,e,v,t)\log(e(s,e,v)) + (\#(s,e,v) - \#(s,e,v,t))\log(1-e(s,e,v))$$

We choose one of the eight possible choices of (s,e,v), and denote this choice as (s^*,e^*,v^*) Taking the derivative with respect to $e(s^*,e^*,v^*)$ setting this equal to zero, and solving for $e(s^*,e^*,v^*)$, we obtain:

$$\begin{aligned}
&\frac{\#(s^*,e^*,v^*,t)}{e(s^*,e^*,v^*)} - \frac{\#(s^*,e^*,v^*) - \#(s^*,e^*,v^*,t)}{1-e(s^*,e^*,v^*)} = 0 \\
\implies\quad &\#(s^*,e^*,v^*,t)(1-e(s^*,e^*,v^*)) = (\#(s^*,e^*,v^*) - \#(s^*,e^*,v^*,t))e(s^*,e^*,v^*)) \\
\implies\quad &e(s^*,e^*,v^*) = \frac{\#(s^*,e^*,v^*,t)}{\#(s^*,e^*,v^*)}
\end{aligned}$$

Our estimate of the propensity score $e(s^*, e^*, v^*)$ is

$$\widehat{e(s^*, e^*, v^*)} = \frac{\#(s^*, e^*, v^*, t)}{\#(s^*, e^*, v^*)},$$

which is the number of treated units having covariates (s^*, e^*, v^*) divided by the number of units with covariates (s^*, e^*, v^*).

This method of obtaining these estimates is called *maximum likelihood estimation*.

b) By the law of large numbers, the sample proportion of treated units with covariates (s, e, v) will converge to the true proportion of treated units with covariates (s, e, v): which is the propensity score $e(s, e, v)$. If, for example, the propensity score does not depend on V_i, we will expect our estimate of $e(s, e, V_i = 1)$ to be the same (asymptotically) as $e(s, e, V_i = 0)$; this convergence is not affected if only a subset of these covariates affect the propensity score.

(Note: Under the assumption that the propensity score only depends on a subset of the three covariates, this method of finding a balance score is finer than the true propensity score. Conditioning on this balance score will give you the same estimates (asymptotically) as conditioning on the propensity score.)

c) Let $C_i(s, e, v) = 1$ if unit i has $S_i = s$, $E_i = e$ and $V_i = v$. The standard estimate for the ATE is

$$\frac{1}{N} \sum_{(s,e,v)\in\{0,1\}^3} \sum_{i=1}^{10000} C_i(s, e, v) \left(\frac{Y_i(1)T_i}{e(s, e, v)} - \frac{Y_i(0)(1 - T_i)}{1 - e(s, e, v)} \right).$$

A possible estimate for the ATT is

$$\sum_{(s,e,v)\in\{0,1\}^3} \sum_{i=1}^{10000} \frac{\#(s, e, v, t)}{\sum T_i} \frac{C_i(s, e, v)}{\#(s, e, v)} \left(\frac{Y_i(1)T_i}{e(s, e, v)} - \frac{Y_i(0)(1 - T_i)}{1 - e(s, e, v)} \right).$$

We see that the estimate of the ATE is unbiased:

$$\begin{aligned}
& \mathbb{E}\left(\frac{1}{N} \sum_{(s,e,v)\in\{0,1\}^3} \sum_{i=1}^{10000} C_i(s, e, v) \left(\frac{Y_i(1)T_i}{e(s, e, v)} - \frac{Y_i(0)(1 - T_i)}{1 - e(s, e, v)} \right) \right) \\
= & \frac{1}{N} \sum_{(s,e,v)\in\{0,1\}^3} \sum_{i=1}^{10000} C_i(s, e, v) \left(\frac{Y_i(1)\mathbb{E}(T_i)}{e(s, e, v)} - \frac{Y_i(0)(1 - \mathbb{E}(T_i))}{1 - e(s, e, v)} \right) \\
= & \frac{1}{N} \sum_{(s,e,v)\in\{0,1\}^3} \sum_{i=1}^{10000} C_i(s, e, v) \left(\frac{Y_i(1)e(s, e, v)}{e(s, e, v)} - \frac{Y_i(0)(1 - e(s, e, v))}{1 - e(s, e, v)} \right) \\
= & \frac{1}{N} \sum_{(s,e,v)\in\{0,1\}^3} \sum_{i=1}^{10000} C_i(s, e, v) \left(Y_i(1) - Y_i(0) \right) \\
= & \frac{1}{N} \sum_{i=1}^{10000} Y_i(1) - Y_i(0) = ATE
\end{aligned}$$

The estimate of the ATT is unbiased as well, though it takes a bit more work to show:

$$\begin{aligned}
& \mathbb{E}\left(\sum_{(s,e,v)\in\{0,1\}^3}\sum_{i=1}^{10000}\frac{\#(s,e,v,t)}{\sum T_i}\frac{C_i(s,e,v)}{\#(s,e,v)}\left(\frac{Y_i(1)T_i}{e(s,e,v)}-\frac{Y_i(0)(1-T_i)}{1-e(s,e,v)}\right)\right) \\
= & \sum_{(s,e,v)\in\{0,1\}^3}\sum_{i=1}^{10000}\frac{\#(s,e,v,t)}{\sum T_i}\frac{C_i(s,e,v)}{\#(s,e,v)}\left(\frac{Y_i(1)\mathbb{E}(T_i)}{e(s,e,v)}-\frac{Y_i(0)(1-\mathbb{E}(T_i))}{1-e(s,e,v)}\right) \\
= & \sum_{(s,e,v)\in\{0,1\}^3}\sum_{i=1}^{10000}\frac{\#(s,e,v,t)}{\sum T_i}C_i(s,e,v)\left(\frac{Y_i(1)}{\#(s,e,v)}-\frac{Y_i(0)}{\#(s,e,v)}\right) \\
= & \sum_{(s,e,v)\in\{0,1\}^3}\frac{\#(s,e,v,t)}{\sum T_i}\mathbb{E}\left(Y_i(1)-Y_i(0)|(s,e,v)\right) \\
= & E_{X_{T=1}}[E(Y_i(1)|(s,e,v))]-E_{X_{T=1}}[E(Y_i(0)|(s,e,v))]
\end{aligned}$$

Now, under the assumption $Y_i \perp\!\!\!\perp T_i|X_i$,

$$\begin{aligned}
& E_{X_{T=1}}[E(Y_i(1)|(s,e,v))]-E_{X_{T=1}}[E(Y_i(0)|(s,e,v))] \\
= & E_{X_{T=1}}[E(Y_i(1)|T_i=1,(s,e,v))]-E_{X_{T=1}}[E(Y_i(0)|T_i=1,(s,e,v))] \\
= & E(Y_i(1)-Y_i(0)|T_i=1)
\end{aligned}$$

If the propensity score is unknown, we may estimate it as in part a).

d) The ATE cannot be estimated by conditioning on the propensity score: the common support assumption is violated (though an ATE for people not weighing 500 pounds could be estimated). Specifically, people that weigh 500 pounds will never be treated. However, the assumptions necessary to estimate the ATT do hold.

The ATT could be estimated a few different ways. Probably the most common method to estimate is to match treated people to control people using the propensity score in some way (for example, by matching to the nearest neighbor with replacement), and then to take the average of the difference between the treated units and the matched control units.

e) We are assuming the model:

$$T_i = \alpha + \beta\text{weight}_i + \epsilon_i$$

where ϵ_i has the distribution

$$\epsilon_i = \begin{cases} 1-\alpha-\beta\text{weight}_i) & \text{with probability } = \alpha+\beta\text{weight}_i, \\ -\alpha-\beta\text{weight}_i & \text{with probability } = 1-\alpha-\beta\text{weight}_i. \end{cases}$$

Note that:

$$\begin{aligned}
\mathbb{E}(\epsilon_i) &= (1-\alpha-\beta\text{weight}_i))(\alpha+\beta\text{weight}_i)+(-\alpha-\beta\text{weight}_i)(1-\alpha-\beta\text{weight}_i) \\
&= (1-\alpha-\beta\text{weight}_i))(\alpha+\beta\text{weight}_i)-(\alpha+\beta\text{weight}_i)(1-\alpha-\beta\text{weight}_i)=0
\end{aligned}$$

Thus, letting

$$X = \begin{pmatrix} 1 & \text{weight}_1 \\ 1 & \text{weight}_2 \\ \vdots & \vdots \\ 1 & \text{weight}_N \end{pmatrix} \text{ and } \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}$$

the expectation of the OLS estimate is

$$\begin{aligned}
\mathbb{E}[(X'X)^{-1}X'Y] &= \mathbb{E}[(X'X)^{-1}X'(X(\alpha,\beta)')+\epsilon)] \\
&= \mathbb{E}[(X'X)^{-1}X'(X(\alpha,\beta)')]+\mathbb{E}[(X'X)^{-1}X'\epsilon] \\
&= (\alpha,\beta)'+(0,0)'=(\alpha,\beta)'
\end{aligned}$$

Thus, even though the ϵ are not identically distributed, because they all have $\mathbb{E}(\epsilon_i) = 0$, OLS will still estimate the α and β parameters unbiasedly.

3) a) This question is better worded as: "Show that, when estimating a regression within a bounded region, the estimates at the boundaries are more variable than anywhere else."

Suppose the regression is fitted through N points in total. Assume that errors for the regression are i.i.d with mean 0 and variance σ^2. It can be shown (for example, in Mathematical Statistics and Data Analysis by John Rice) that the variance for the regression estimate at x_i is

$$\sigma^2 \left(\frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{N} (x_j - \bar{x})^2} \right)$$

This variance increases as the distance between x_i and $\bar{x}$ grows; estimates are most variable at the boundary.

Estimates of the LATE can compare the regression estimates at the cut points, which can be quite variable if the cut point is far away from the mean of the regressed covariate.

b,c) In b) and c), I meant to write that the coin flip only reduced scores for students scoring between c and $c + 10$ points. In which case, the smoothness of covariates at the cut point is violated: Below the cut point contains people who scored between $c - 5$ and c, as well as those that scored $c + 5$ and $c + 10$ points on the exam. Above the cut point contains future income only include people that scored between c and $c + 5$ points. Asymptotically, you get imbalance in the Test Score Before Coin Toss covariate at the cut point.

However, when regressing on data to the left of the cut point, the regression estimate for the future outcome for at the cut point will be higher when this coin flip mechanism is in place than when it is not in place: the estimate of the LATE (the difference of the regression estimates at the cutpoint) will be smaller with the coin flip than without the coin flip. Without the coin flip, this estimate be unbiased for the LATE; so with the coin flip, this estimate will be a lower bound (asymtotically) to the true LATE.

Instead, suppose additionally that the distribution of scores on the interval $(c, c + 5)$ is the same of the distribution of the scores on the interval $(c + 10, c + 15)$, which is closer to the interpretation of the problem as written. (By distribution of the scores is the same, I mean that: if you graph a histogram of the test scores within $(c, c+5)$, and another histogram of the test scores within $(c+10, c+15)$, the histograms would look the same.) We assume the model written in part c).

We introduce some notation: Let x_i denote the pre-coin-flip test score of person i. Let z_i be defined as

$$z_i = \begin{cases} y_i, & x_i < c, \\ y_i - 10\beta, & x_i \geq c. \end{cases}$$

Suppose that, of all people with test scores after the coin flip within $(c - 5, c)$, N^- of those people had test scores before the coin flip within $(c - 5, c)$, and N^+ had test scores within $(c + 5, c + 15)$. Similarly, suppose that, of all people with post-coin-flip test scores within $(c, c + 5)$, M^- people had pre-coin-flip test scores between $(c, c + 5)$ and M^+ had test scores between $(c + 10, c + 15)$. Suppose that people with post-coin-flip scores between $(c - 5, c)$ are ordered so that the first N^- people correspond to those having pre-test-scores within $(c, c + 5)$, and the last N^+ people have scores within $(c + 10, c + 15)$.

By the equal in distribution assumption, after the coin flip, the mean and the variance of the test scores within $(c, c + 5)$ will be the same as before the coin flip. The mean future incomes will now have an average (asymptotically) of

$$\frac{N^-(\bar{z}) + N^+(\bar{z} + 10\beta)}{N^- + N^+} = \bar{z} + \frac{N^+}{N^- + N^+} 10\beta.$$

Also by the equal distribution assumption, asymptotically we have

$$\begin{aligned}
\sum_{i=1}^{N^-}(x_i-\bar{x})^2 &= \frac{N^-}{N^-+N^+}\sum_{i=1}^{N^-+N^+}(x_i-\bar{x})^2 \\
\sum_{i=1}^{N^-}(x_i-\bar{x})(z_i-\bar{z}) &= \frac{N^-}{N^-+N^+}\sum_{i=1}^{N^-+N^+}(x_i-\bar{x})(z_i-\bar{z}) \\
\sum_{i=N^++1}^{N^-+N^+}(x_i-\bar{x}) &= \frac{N^+}{N^-+N^+}\sum_{i=1}^{N^-+N^+}(x_i-\bar{x})
\end{aligned}$$

And we have

$$\begin{aligned}
&\sum_{i=1}^{N^-}(x_i-\bar{x})(z_i-\bar{z}-\frac{N^+}{N^-+N^+}10\beta)+\sum_{i=N^-+1}^{N^-+N^+}(x_i-\bar{x})(z_i+10\beta-\bar{z}-\frac{N^+}{N^-+N^+}10\beta) \\
= &\sum_{i=1}^{N^-+N^+}(x_i-\bar{x})(z_i-\bar{z})-\frac{N^+}{N^-+N^+}10\beta\sum_{i=1}^{N^-}(x_i-\bar{x}) \\
&+10\beta\sum_{i=N^-+1}^{N^-+N^+}(x_i-\bar{x})-\frac{N^+}{N^-+N^+}10\beta\sum_{i=N^-+1}^{N^-+N^+}(x_i-\bar{x}) \\
= &\sum_{i=1}^{N^-+N^+}(x_i-\bar{x})(z_i-\bar{z})-\frac{N^+}{N^-+N^+}10\beta\sum_{i=1}^{N^-+N^+}(x_i-\bar{x})+10\beta\sum_{i=N^-+1}^{N^-+N^+}(x_i-\bar{x}) \\
\approx &\sum_{i=1}^{N^-+N^+}(x_i-\bar{x})(z_i-\bar{z})-10\beta\sum_{i=N^-+1}^{N^-+N^+}(x_i-\bar{x})+10\beta\sum_{i=N^-+1}^{N^-+N^+}(x_i-\bar{x}) \\
= &\sum_{i=1}^{N^-+N^+}(x_i-\bar{x})(z_i-\bar{z}) \qquad (2)
\end{aligned}$$

Consider fitting a regression to people with test scores within $(c-5,c)$. The slope of the regression line pre-coin-flip will be

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{i=1}^{N^-}(x_i-\bar{x})(z_i-\bar{z})}{\sum_{i=1}^{N^-}(x_i-\bar{x})^2} \\
&\approx \frac{\frac{N^-}{N^-+N^+}\sum_{i=1}^{N^-+N^+}(x_i-\bar{x})(z_i-\bar{z})}{\frac{N^-}{N^-+N^+}\sum_{i=1}^{N^-+N^+}(x_i-\bar{x})^2} \\
&= \frac{\sum_{i=1}^{N^-+N^+}(x_i-\bar{x})(z_i-\bar{z})}{\sum_{i=1}^{N^-+N^+}(x_i-\bar{x})^2}
\end{aligned}$$

And, by (2), the regression line after the coin flip will be:

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{i=1}^{N^-}(x_i-\bar{x})(z_i-\bar{z}-\frac{N^+}{N^-+N^+}10\beta)+\sum_{i=N^-+1}^{N^-+N^+}(x_i-\bar{x})(z_i+10\beta-\bar{z}-\frac{N^+}{N^-+N^+}10\beta}{\sum_{i=1}^{N^-+N^+}(x_i-\bar{x})^2} \\
&\approx \frac{\sum_{i=1}^{N^-+N^+}(x_i-\bar{x})(z_i-\bar{z})}{\sum_{i=1}^{N^-+N^+}(x_i-\bar{x})^2}
\end{aligned}$$

Thus, asymptotically, the slope of the fitted regression lines before and after the coin flip will be the same. Additionally, since the new estimated intercept of the regression line is

$$\bar{y} - \hat{\beta}\bar{x} = \bar{z} + \frac{N^+}{N^- + N^+}10\beta - \hat{\beta}\bar{x}$$

it follows that the post-coin-flip estimate at the cutpoint is

$$\hat{y}_{post} = \hat{y}_{pre} + \frac{N^+}{N^- + N^+}10\beta$$

Following the same process, we find that a similar result holds for the regression fitted to people with post-coin-flip test scores in $(c, c+5)$:

$$\hat{y}_{post} = \hat{y}_{pre} + \frac{M^+}{M^- + M^+}10\beta$$

It follows that, if $\frac{N^+}{N^-+N^+} = \frac{M^+}{M^-+M^+}$ the estimate of the LATE will be unbiased when regressing on the post-coin-flip test scores.

4) See `HW3_Answers.R` for solutions