1 Model

Latent variables:

$$
\begin{aligned}
Y_i(0) &= \alpha_0 + \beta_0' X_i + \varepsilon_i^0 && (1)\\
Y_i(1) &= \alpha_1 + \beta_1' X_i + \varepsilon_i^1 && (2)\\
T_i^* &= \alpha_T + \beta_T' X_i - u_i && (3)
\end{aligned}
$$

where we asssume that the pair $(\varepsilon_i^0, \varepsilon_i^1)$ is mean zero and independent of X_i, that u_i is mean zero and independent of X_i, and further that the pair $(\varepsilon_i^0, \varepsilon_i^1)$ is independent of u_i. We never observe any of $Y_i(0)$, $Y_i(1)$, or T_i^*. Rather, we see (X_i, T_i, Y_i), where

$$
\begin{aligned}
T_i &= \mathbf{1}(T_i^* > 0) && (4)\\
Y_i &= T_i Y_i(1) + (1 - T_i) Y_i(0) && (5)
\end{aligned}
$$

Parameters:

$$
\begin{aligned}
ATE &= \mathbb{E}[Y_i(1) - Y_i(0)] && (6)\\
TOT &= \mathbb{E}[Y_i(1) - Y_i(0) | T_i = 1] && (7)
\end{aligned}
$$

2 Regression

Plug in (1) and (2) into (5):

$$
\begin{aligned}
Y_i &= Y_i(0) + T_i \left(Y_i(1) - Y_i(0) \right) && (8)\\
&= \alpha_0 + \beta_0' X_i + (\alpha_1 - \alpha_0) T_i + (\beta_1 - \beta_0)' T_i X_i + \varepsilon_i && (9)
\end{aligned}
$$

where $\varepsilon_i = \varepsilon_i^0 + T_i \left(\varepsilon_i^1 - \varepsilon_i^0 \right)$ is a composite (heteroskedastic) error term. This motivates a regression of Y_i on X_i, T_i, and their interactions. Then note that

$$
\begin{aligned}
ATE &= \mathbb{E}[Y_i(1) - Y_i(0)] = \alpha_1 - \alpha_0 + (\beta_1 - \beta_0)' \mathbb{E}[X_i] && (10)\\
TOT &= \mathbb{E}[Y_i(1) - Y_i(0)] = \alpha_1 - \alpha_0 + (\beta_1 - \beta_0)' \mathbb{E}[X_i | T_i = 1] && (11)
\end{aligned}
$$

so that a natural way to estimate these parameters is to estimate the regression in (9) and then to use the sample mean to compute $\mathbb{E}[X_i]$ or $\mathbb{E}[X_i | T_i = 1]$.

3 Reweighting

The reason why we have to control for X_i in the regression approach is that treatment is associated with X_i. An alternative approach is to use Bayes' Rule to reweight observations so that the covariates are similar between treatment and control. Note that for any function $g(\cdot)$, we have

$$
\begin{aligned}
\mathbb{E}\left[g(X_i) | T_i = 1 \right] &= \mathbb{E}\left[g(X_i) \frac{p(X_i)}{1 - p(X_i)} \frac{1-q}{q} \middle| T_i = 0 \right] && (12)\\
\mathbb{E}\left[g(X_i) \frac{q}{p(X_i)} | T_i = 1 \right] &= \mathbb{E}\left[g(X_i) \frac{1-q}{1 - p(X_i)} \middle| T_i = 0 \right] = \mathbb{E}\left[g(X_i) \right] && (13)
\end{aligned}
$$

where $p(X_i) = P(T_i = 1 | X_i)$ is the propensity score, or the conditional probability of treatment given covariates. Equation (12) means that we can reweight the sample so that the distribution of X_i among control units is the same as the distribution of X_i among treated units. Equation (13) means that we can reweight the sample so that the distribution of X_i among control units is the same as the distribution of X_i in the population, and likewise for treated units. Both of these equations are easy to prove using iterated expectations. For example, we have

$$
\begin{aligned}
\mathbb{E}\left[g(X_i) \frac{q}{p(X_i)} \middle| T_i = 1 \right] &= \frac{1}{q} \mathbb{E}\left[T_i g(X_i) \frac{q}{p(X_i)} \right] = \mathbb{E}\left[\mathbb{E}\left[T_i g(X_i) \frac{1}{p(X_i)} \middle| X_i \right] \right] && (14)\\
&= \mathbb{E}\left[\mathbb{E}\left[T_i | X_i \right] g(X_i) \frac{1}{p(X_i)} \right] = \mathbb{E}\left[g(X_i) \frac{p(X_i)}{p(X_i)} \right] = \mathbb{E}\left[g(X_i) \right] && (15)
\end{aligned}
$$

The other results follow from these kinds of calculations, and you can check them yourself. That suggests the following estimators for TOT (the case of ATE is analogous):

$$\widehat{\theta} = \frac{\sum_{i=1}^{n} T_i Y_i}{\sum_{i=1}^{n} T_i} - \frac{\sum_{i=1}^{n} (1-T_i) \frac{p(X_i)}{1-p(X_i)} \frac{1-q}{q} Y_i}{\sum_{i=1}^{n} (1-T_i)} \tag{16}$$

where we assume $p(X_i)$ and q are known, which is almost always wrong. More practically, people implement this idea as

$$\widehat{\theta}_N = \frac{\sum_{i=1}^{n} T_i Y_i}{\sum_{i=1}^{n} T_i} - \frac{\sum_{i=1}^{n} (1-T_i) \frac{\widehat{p}(X_i)}{1-\widehat{p}(X_i)} Y_i}{\sum_{i=1}^{n} (1-T_i) \frac{\widehat{p}(X_i)}{1-\widehat{p}(X_i)}} \tag{17}$$

where the weights are additionally forced to sum to one. This is a good idea. Here is the standard algorithm for estimating a reweighting estimator for TOT:

```
1. logit T X1 X2 X3 X4
2. predict double phat
3. gen double W=phat/(1-phat)
4. reg Y T [aw=W]
```

The reweighting estimate of TOT is the coefficient on `T` in this regression. Usually people take the standard error on treatment as the standard error. If $n > 300$ or so, this works quite well. You can prove that to yourself using the techniques from the last problem set.

4 Matching

Keep the focus on TOT, as before. Here, the idea is to use various notions of distance to "match" observations. Let $W(i,j)$ denote the proximity of unit i to unit j. The definition of $W(i,j)$ depends on the matching approach in question. These estimators can be written as

$$\widetilde{\theta} = \frac{\sum_{i=1}^{n} T_i \left\{ Y_i - \widehat{Y}_i(0) \right\}}{\sum_{i=1}^{n} T_i} \tag{18}$$

where

$$\widehat{Y}_i(0) = \frac{\sum_{j=1}^{n} (1-T_j) W(i,j) Y_j}{\sum_{j=1}^{n} (1-T_j) W(i,j)} \tag{19}$$

is the imputed counterfactual outcome for unit i.

Programming matching estimators is a pain, because you have to loop over observations, which is slow. You also typically need to choice tuning parameters, such as a bandwidth. So you often end up resorting to cross-validation to choose them, which means recomputing the matching estimator, or an analogue of it, again and again. In other words, if looping over observations is slow, then cross-validating an estimator that loops over observations is really slow. (But computers are fast, so maybe this isn't such a big deal.)