REGRESSION WITH COVARIATES

In the previous lesson, we reformulated model based inference as a linear regression problem. In this lesson, we add covariates $\underline{X}_i$ that are often collected in completely randomized experiments and observational studies.

In the completely randomized study, we have seen that $\bar{Y}_1 - \bar{Y}_0$ is unbiased for the ATE $E(Y(1) - Y(0))$

Investigators sometimes nevertheless consider regressing the outcomes on treatment assignment Z_i and covariates $\underline{X}_i$:

$$Y_i(Z_i) = \alpha^* + \tau^* Z_i + \underline{\beta}^{*\prime} \underline{X}_i + v_i$$

REGRESSION WITH COVARIATES

Recall that the ordinary least squares (OLS) estimator makes the residuals and weighted residuals sum to 0:

$$\sum_{i=1}^{n} v_i = \sum_{i=1}^{n} Z_i v_i = \sum_{i=1}^{n} \underline{X}_i v_i = 0$$

It follows that $\bar{Y}_0 = \hat{\alpha}^* + \underline{\hat{\beta}}^{*\prime} \underline{\bar{X}}_0$ and $\bar{Y}_1 = \hat{\alpha}^* + \hat{\tau}^* + \underline{\hat{\beta}}^{*\prime} \underline{\bar{X}}_1$

$\underline{\bar{X}}_0$ is the vector of sample means in the group that does not receive treatment, and $\underline{\bar{X}}_1$ is the vector of sample means in the group receiving treatment.

REGRESSION WITH COVARIATES

From $\bar{Y}_0 = \hat{\alpha}^* + \underline{\hat{\beta}}^{*\prime}\underline{\bar{X}}_0$ and $\bar{Y}_1 = \hat{\alpha}^* + \hat{\tau}^* + \underline{\hat{\beta}}^{*\prime}\underline{\bar{X}}_1$ we have

$$\bar{Y}_1 - \bar{Y}_0 = \hat{\tau}^* + \underline{\hat{\beta}}^{*\prime}(\underline{\bar{X}}_1 - \underline{\bar{X}}_0)$$

Since in a completely randomized experiment

$$Y_i(0),\ Y_i(1),\ \underline{X}_i \perp\!\!\!\perp Z_i$$

$E(\underline{\bar{X}}_0) = E(\underline{\bar{X}}_1)$ so that $\hat{\tau}^* \neq \bar{Y}_1 - \bar{Y}_0$ (unless it just so happens that $\underline{\bar{X}}_1 = \underline{\bar{X}}_0$ for the sample in hand) is an alternative unbiased estimator of the ATE.

REGRESSION WITH COVARIATES

It would appear by assuming

$$Y_i(Z_i) = \alpha^* + \tau^* Z_i + \underline{\beta}^{*\prime} \underline{X}_i + v_i$$

we impose the restriction that the ATE is constant for all levels of the covariates.

Actually that is not the case. We only assume that the errors v_i are uncorrelated with the regressors, not the stronger condition $\mathrm{E}(v_i \mid Z_i, \underline{X}_i) = 0$. i.e. we did not assume $\alpha^* + \tau^* Z_i + \underline{\beta}^{*\prime} \underline{X}_i$ is the conditional expectation function $E(Y_i(Z_i) \mid Z_i, \underline{X}_i)$.

It can be shown the asymptotic variance of $\hat{\tau}^*$ is smaller than that of $\bar{Y}_1 - \bar{Y}_0$ (see ch. 7 of Imbens-Rubin for proof), hence $\hat{\tau}^*$ is preferred.

REGRESSION WITH COVARIATES

Investigators often use linear regression as well with covariates $\underline{X}_i$ in observational studies and treat $\hat{\tau}^*$ as an estimate of the ATE.

In this case, $\hat{\tau}^*$ is a consistent estimator of

$$\begin{aligned} &[E(Y(1) \mid Z = 1) - \underline{\beta}^{*\prime} E(\underline{X} \mid Z = 1)] \\ -&[E(Y(0) \mid Z = 0) - \underline{\beta}^{*\prime} E(\underline{X} \mid Z = 0)] \end{aligned}$$

But it is not necessarily the case, as with the completely randomized experiment, that $E(Y(z) \mid Z = z) = E(Y(z))$ and $E(\underline{\bar{X}}_0) = E(\underline{\bar{X}}_1)$.

REGRESSION WITH COVARIATES

Assume treatment assignmen is unconfounded given covariates $\underline{X}$. i.e. $Y(0), Y(1) \perp\!\!\!\perp Z \mid \underline{X}$. Then the conditional expectation function for the observed data can be written $E(Y \mid Z = z, \underline{X} = \underline{x}) = E(Y(z) \mid \underline{X} = \underline{x})$.

Also, assume the true model is $Y_i(z) = g(z, \underline{X}_i) + \varepsilon_i(z)$ where $\mathrm{E}(\varepsilon_i(z) \mid X_i) = 0$

In this case, ATE($\underline{X}$) = g(1,$\underline{X}$) – g(0,$\underline{X}$) and ATE = E(ATE($\underline{X}$)). Thus,

$$[E(Y(1) \mid Z = 1) - \underline{\beta}^{*\prime} E(\underline{X} \mid Z = 1)]$$
$$-[E(Y(0) \mid Z = 0) - \underline{\beta}^{*\prime} E(\underline{X} \mid Z = 0)]$$

reduces to

$$E(g(1, \underline{X}) - \underline{\beta}^{*\prime} \underline{X}) \mid Z = 1) - E(g(0, \underline{X}) - \underline{\beta}^{*\prime} \underline{X}) \mid Z = 0)$$

REGRESSION WITH COVARIATES

For the special case of $Y_i(z) = g(z, \underline{X}_i) + \varepsilon_i(z)$ where $\mathrm{E}(\varepsilon_i(z) \mid X_i) = 0$, $g(1, \underline{X}) = g(0, \underline{X}) + \tau$

i.e. $\mathrm{ATE}(\underline{X}) = \mathrm{ATE}$ for all $\underline{x}$ and,

$$E(g(1, \underline{X}) - \underline{\beta}^{*\prime}\underline{X}) \mid Z = 1) - E(g(0, \underline{X}) - \underline{\beta}^{*\prime}\underline{X}) \mid Z = 0)$$

reduces further to

$$\tau + [E(g(0, \underline{X}) - \underline{\beta}^{*\prime}\underline{X} \mid Z = 1) - E(g(0, \underline{X}) - \underline{\beta}^{*\prime}\underline{X} \mid Z = 0)]$$

where the bracketed term is the bias.

REGRESSION WITH COVARIATES

We see that

1. if the distribution of the covariates is the same in the treatment and control groups, or
2. $g(0, \underline{X}) = \underline{\beta}^{*\prime}\underline{X}$

the bias vanishes.

However, even in this additive case, where the value of the average treatment effect is constant across levels of the covariates, the bias incurred by using the linear regression model can be substantial if the linear specification is substantially off the mark and/or if the distribution of the covariates is very different in the treatment and control groups.

REGRESSION WITH COVARIATES

In observational studies, where the covariate distributions are often quite different in the treatment and control groups, and the investigator does not have enough knowledge to correctly specify the regression function, this can lead to very poor estimates of treatment effects.

This concern has motivated the development of other methods for the estimation of treatment effects in observational studies.