1 Introduction

1.1 Conditional Expectation

Definition 1 *Suppose we know a random variable, X, has taken a particular numerical value, say x. Then, one can calculate the* expected value *of another random variable, Y, given that $X = x$. This expected value is called the* Conditional Expectation *of Y given $X = x$ and is defined as:*

- *When Y is a* discrete *random variable taking on finite values $\{y_1, \ldots, y_J\}$, then*

$$E[Y|X = x] = E[Y|x] = \sum_{j=1}^{J} y_j \Pr[Y = y_j|X = x].$$

- *when Y is a* continuous *random variable with conditional distribution $F_{Y|X}$, then*

$$E[Y|X = x] = E[Y|x] = \int_{-\infty}^{\infty} y dF_{Y|X}.$$

Properties of Conditional Expectations:

1. $E[c(X)|X] = c(X)$, for any function $c(\cdot)$.
2. For any two functions $a(\cdot)$ and $b(\cdot)$, we have

$$E[a(X)Y + b(X)|X] = a(X)E[Y|X] + b(X).$$

3. If two random variables Y and X are statistically independent, then $E[Y|X] = E[Y]$.[1]
4. $E[E[Y|X]] = E[Y]$.
5. $E[Y|X] = E[E[Y|X, Z]|X]$.
6. If $E[Y|X] = E[Y]$, then $\text{cov}(Y, X) = 0$. In fact, *all* functions of X are uncorrelated with Y.[2]
7. If $E[Y^2] < \infty$ and $E[g(X)^2] < \infty$ for a function $g(\cdot)$, then

$$E\{[Y - E[Y|X]]^2|X\} \leq E\{[Y - g(X)]^2|X\}, \text{ y}$$

$$E\{[Y - E[Y|X]]^2\} \leq E\{[Y - g(X)]^2\}.$$

This last property is used a lot in the context of forecasting. The first inequality says that, if we were to measure the forecasting error as the *Mean Squared Forecasting Error*, conditioned on X, then the conditional expected values is better than *any* other function of X when predicting Y.

- For all two random variables $(Y, X^\top)^\top$ one can write: $Y = E[Y|X = x] + \varepsilon$, where $E[\varepsilon|X = x] = 0$.
- A lineal regression model implies $E[Y|X = x] = x^\top\beta$. A Probit or Logit model implies that $E[Y|X = x] = F_{\varepsilon|x}(x'\beta)$ with $F_{\varepsilon|x}$ equal to the CDF of a standard normal for a Probit and the CDF of a logistic for a Logit. Similarly, a Poisson model implies that $E[Y|X = x] = \exp(x'\beta)$.

[1] Recall that Y and X are statistically independent *iff* $F_{YX}(y, x) = F_Y(y) \times F_X(x)$.

[2] The *covariance* between two random variables Y and X is defined as $\text{cov}(Y, X) = E[YX] - E[Y]E[X]$.

2 Program Evaluation Techniques

Another area where sample selection plays a crucial role is in the estimation of *treatment effects*. Suppose that if observation i with characteristics $\mathbf{x}_i$ receives certain treatment (such as a government sponsored training programme) then an outcome y_{1i} is observed, if this observation does not receive the treatment one only observes y_{0i}. Let y_i denote the observed outcome (such as wage earnings) and d_i an indicator of whether individual i received treatment or not. Then the observed outcome can be written in terms of a *Roy's model*, i.e.

$$y_i = d_i y_{1i} + (1 - d_i) y_{0i}. \tag{1}$$

The literature has concentrated on the following two parameters of interest:[3]

$$\begin{aligned} &\text{Average Treatment Effect:} && \tau_{\text{ATE}} := \mathbb{E}[y_{1i} - y_{0i}], \\ &\text{Average Treatment on the Treated Effect:} && \tau_{\text{ATT}} := \mathbb{E}[y_{1i} - y_{0i} | d_i = 1]. \end{aligned}$$

The τ_{ATE} describes the expected effect of treatment for an *arbitrary* observation i chosen at random from the population, while the τ_{ATT} is the mean effect for those that actually participate in the programme, i.e. the Average Treatment Effect in the treated subpopulation.[4]

Remark 2 *Since for each observation i one only observes either y_{1i} or y_{0i}, but not both (missing data problem), the joint distribution $F_{10}(y_1, y_0)$ is not identified. One can only identified the marginals $F_1(y_1)$ and $F_0(y_0)$. It turns out that even when one can not identify $F_{10}(\cdot, \cdot)$, certain features of it such as τ_{ATE} can be identified under less restrictive conditions that independence between y_1 and y_0.*

Relationship between ATE & ATT:
Let $\mathbb{E}[y_{1i}] =: \mu_1$ and $\mathbb{E}[y_{0i}] =: \mu_0$, then by construction $y_{1i} = \mu_1 + v_{1i}$, $y_{0i} = \mu_0 + v_{0i}$, such that $\mathbb{E}[v_{1i}] = \mathbb{E}[v_{0i}] = 0$ and

$$\begin{aligned} y_{1i} - y_{0i} &= \mu_1 - \mu_0 + [v_{1i} - v_{0i}] \\ &= \tau_{\text{ATE}} + [v_{1i} - v_{0i}], \text{ and} \\ \mathbb{E}[y_{1i} - y_{0i} | d_i = 1] &= \tau_{\text{ATE}} + \mathbb{E}[v_{1i} - v_{0i} | d_i = 1], \\ \tau_{\text{ATT}} &= \tau_{\text{ATE}} + \mathbb{E}[v_{1i} - v_{0i} | d_i = 1]. \end{aligned}$$

Therefore τ_{ATT} differs from τ_{ATE} by the expected person-specific gain for those who participated.

3 Unconfoundedness based Methods

If subject i is randomly assigned to the treatment group or the control group, then Assumption **??** holds by construction and Assumption **??** is also bound to hold.

[3] One can also define the *conditional* Average Treatment Effect as $\tau_{\text{CATE}}(\mathbf{x}_i) := \mathbb{E}[y_{1i} - y_{0i} | \mathbf{x}_i]$. Notice that by the definitions of τ_{ATE} and τ_{ATT}, we have $\tau_{\text{ATE}} = \mathbb{E}[\tau_{\text{CATE}}(\mathbf{x}_i)]$ and $\tau_{\text{ATT}} = \mathbb{E}[\tau_{\text{CATE}}(\mathbf{x}_i) | d_i = 1]$.

[4] Imbens & Angrist (1994) defined a third parameter of interest they called the *Local Average Treatment Effect*, i.e. τ_{LATE}, which measures the effect of treatment upon observations at the margin of being treated.

Suppose we observe a sample $\{y_i, d_i, \mathbf{x}_i'\}_{i=1}^n$ from the joint distribution of $(y, d, \mathbf{x}')' \in \mathcal{Y} \times [0,1] \times \mathcal{X}$ such that:

1. (Ignorability) (y_1, y_0) and d are independent conditional of $\mathbf{x}$.

2. (Overlap) For all $\mathbf{x} \in \mathcal{X}$, $0 < \Pr\{d = 1|\mathbf{x}\} := p(\mathbf{x}) < 1$.

Assumption **??** is known in the statistical literature as: *ignorability of treatment*, *unconfoundedness* or simply *conditional independence*. The idea is the following: If we were to observe enough information (contained in $\mathbf{x}$) that determines treatment, then (y_1, y_0) might be independent of d (conditional on $\mathbf{x}$) - even though (y_1, y_0) and d might be correlated, they become *independent* once we partial out $\mathbf{x}$.[5] Because of the latter, Assumption **??** is also known in econometrics as *selection on observables*, i.e. once $\mathbf{x}$ is observed and conditioned on, there is *no* unobserved factor that influence both outcomes (y_1, y_0) and treatment d simultaneously.[6]

Assumption **??** guarantees that one observes individuals with the same characteristics $\mathbf{x}$ in both the *control* $(d = 0)$ and *treatment* $(d = 1)$ groups. $p(\mathbf{x})$ is known as the *propensity score*.

3.1 Regression based Methods

Define $\mu_1(\mathbf{x}) := \mathbb{E}[y_1|\mathbf{x}]$ and $\mu_0(\mathbf{x}) := \mathbb{E}[y_0|\mathbf{x}]$.[7] Notice that $\mu_1(\mathbf{x})$ and $\mu_0(\mathbf{x})$ are generally unknown, but Assumptions **??** and **??** provide a way to identify them as follows: Notice that one can rewrite Roy's model in (**??**) as

$$
\begin{aligned}
y &= y_0 + d(y_1 - y_0), \text{ so} \\
\mathbb{E}[y|\mathbf{x}, d] &= \mathbb{E}[y_0|\mathbf{x}, d] + d\{\mathbb{E}[y_1|\mathbf{x}, d] - \mathbb{E}[y_0|\mathbf{x}, d]\} \\
&= \mathbb{E}[y_0|\mathbf{x}] + d\{\mathbb{E}[y_1|\mathbf{x}] - \mathbb{E}[y_0|\mathbf{x}]\} \text{ by Assumption } \mathbf{??} \\
&= \mu_0(\mathbf{x}) \ + d\{\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})\}.
\end{aligned} \tag{2}
$$

[5] Good candidates for inclusion in $\mathbf{x}$ are *pre-treatment* variables which values do not change during the time treatment takes effect.

[6] *Selection on Observables* is inherently a non-testable restriction.

[7] So by construction $y_1 = \mu_1(\mathbf{x}) + v_1$, $y_0 = \mu_0(\mathbf{x}) + v_0$, such that $\mathbb{E}[v_1|\mathbf{x}] = \mathbb{E}[v_0|\mathbf{x}] = 0$ and

$$
\begin{aligned}
&\mathbb{E}[y_1 - y_0|\mathbf{x}, d = 1] \\
&= \tau_{\text{CATE}}(\mathbf{x}) + \mathbb{E}[v_1 - v_0|\mathbf{x}, d = 1], \\
&= \tau_{\text{CATE}}(\mathbf{x}) + \mathbb{E}[v_1 - v_0|\mathbf{x}, d = 1], \\
&= \tau_{\text{CATE}}(\mathbf{x}) + \mathbb{E}[v_1 - v_0|\mathbf{x}] \text{ by Assumption } \mathbf{??} \\
&= \tau_{\text{CATE}}(\mathbf{x}).
\end{aligned}
$$

Therefore[8]

$$\mu_0(\mathbf{x}) \equiv \mathbb{E}[y|\mathbf{x}, d=0] =: m_0(\mathbf{x}),$$
$$\mu_1(\mathbf{x}) \equiv \mathbb{E}[y|\mathbf{x}, d=1] =: m_1(\mathbf{x}).$$

Based on representation (**??**) and under Assumptions **??** and **??**, one can write

$$\begin{aligned}\mathbb{E}[y|\mathbf{x}, d=1] - \mathbb{E}[y|\mathbf{x}, d=0] &= \mathbb{E}[y_0|\mathbf{x}, d=1] + \mathbb{E}[y_1 - y_0|\mathbf{x}, d=1] - \mathbb{E}[y_0|\mathbf{x}, d=0], \\ &= \{\mathbb{E}[y_0|\mathbf{x}, d=1] - \mathbb{E}[y_0|\mathbf{x}, d=0]\} + \tau_{\text{CATE}}(\mathbf{x}), \\ m_1(\mathbf{x}) - m_0(\mathbf{x}) &= \tau_{\text{CATE}}(\mathbf{x}) \text{ by Assumption } \mathbf{??}.\end{aligned}$$

In conclusion

$$\begin{aligned}\tau_{\text{ATE}} &:= \mathbb{E}[\tau_{\text{CATE}}(\mathbf{x})] & \tau_{\text{ATT}} &:= \mathbb{E}[\tau_{\text{CATE}}(\mathbf{x})|d=1], \\ &= \mathbb{E}[m_1(\mathbf{x}) - m_0(\mathbf{x})], & &= \mathbb{E}[m_1(\mathbf{x}) - m_0(\mathbf{x})|d=1].\end{aligned}$$

Estimation:

If $\widehat{m}_1(\mathbf{x}_i)$ and $\widehat{m}_0(\mathbf{x}_i)$ represent consistent estimators of $m_1(\mathbf{x})$ and $m_0(\mathbf{x})$, Let $m_1(\mathbf{x}, \delta_1)$ and $m_0(\mathbf{x}, \delta_0)$ be parametric known functions, and let $\widehat{\delta}_1$ be estimated using those observations for which $d_i = 1$ and let $\widehat{\delta}_0$ be estimated using those observations for which $d_i = 0$, then

$$\widehat{\tau}_{\text{ATE}} = n^{-1} \sum_{i=1}^{n} [m_1(\mathbf{x}_i, \widehat{\delta}_1) - m_0(\mathbf{x}_i, \widehat{\delta}_0)],$$
$$\widehat{\tau}_{\text{ATT}} = (\textstyle\sum_{j=1}^{n} d_j)^{-1} \displaystyle\sum_{i=1}^{n} d_i [m_1(\mathbf{x}_i, \widehat{\delta}_1) - m_0(\mathbf{x}_i, \widehat{\delta}_0)],$$

and we have already discussed the way to obtain expressions for the asymptotic variances. See wooldridge then using the entire random sample of size n one has by the analogy principle

$$\widehat{\tau}_{\text{ATE}} = n^{-1} \sum_{i=1}^{n} [\widehat{m}_1(\mathbf{x}_i) - \widehat{m}_0(\mathbf{x}_i)],$$
$$\widehat{\tau}_{\text{ATT}} = (\textstyle\sum_{j=1}^{n} d_j)^{-1} \displaystyle\sum_{i=1}^{n} d_i [\widehat{m}_1(\mathbf{x}_i) - \widehat{m}_0(\mathbf{x}_i)].$$

3.2 Propensity Score based Methods

Another way to establish identification is to use *inverse propensity score weighting* as follows:

$$\begin{aligned}\mathbb{E}\left[\frac{dy}{p(\mathbf{x})}\middle|\mathbf{x}\right] &= \mathbb{E}\left[\frac{dy_1}{p(\mathbf{x})}\middle|\mathbf{x}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{dy_1}{p(\mathbf{x})}\middle|\mathbf{x}, d\right]\middle|\mathbf{x}\right] \\ &= \mathbb{E}\left[\frac{d\mathbb{E}[y_1|\mathbf{x}, d]}{p(\mathbf{x})}\middle|\mathbf{x}\right] = \mathbb{E}\left[\frac{d\mathbb{E}[y_1|\mathbf{x}]}{p(\mathbf{x})}\middle|\mathbf{x}\right] = \mu_1(\mathbf{x})\mathbb{E}\left[\frac{d}{p(\mathbf{x})}\middle|\mathbf{x}\right] = \mu_1(\mathbf{x})\frac{\mathbb{E}[d|\mathbf{x}]}{p(\mathbf{x})} \\ &= \mu_1(\mathbf{x}).\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\frac{(1-d)y}{1-p(\mathbf{x})}\middle|\mathbf{x}\right] &= \mathbb{E}\left[\frac{(1-d)y_0}{1-p(\mathbf{x})}\middle|\mathbf{x}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{(1-d)y_0}{1-p(\mathbf{x})}\middle|\mathbf{x}, d\right]\middle|\mathbf{x}\right] = \mathbb{E}\left[\frac{(1-d)\mathbb{E}[y_0|\mathbf{x}, d]}{1-p(\mathbf{x})}\middle|\mathbf{x}\right] \\ &= \mathbb{E}\left[\frac{(1-d)\mathbb{E}[y_0|\mathbf{x}]}{1-p(\mathbf{x})}\middle|\mathbf{x}\right] = \mu_0(\mathbf{x})\mathbb{E}\left[\frac{1-d}{1-p(\mathbf{x})}\middle|\mathbf{x}\right] = \mu_0(\mathbf{x})\frac{1-\mathbb{E}[d|\mathbf{x}]}{1-p(\mathbf{x})} \\ &= \mu_0(\mathbf{x}).\end{aligned}$$

[8] Since a sample from the joint distribution of $(y, \mathbf{x}', d)$ is observed, then $m_1(\cdot)$ and $m_0(\cdot)$ are said to be *nonparametrically identified*.

Recall

$$\tau_{\text{CATE}}(\mathbf{x}) := \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) = \mathbb{E}\left[\frac{dy}{p(\mathbf{x})} - \frac{(1-d)y}{1-p(\mathbf{x})}\middle|\mathbf{x}\right] = \mathbb{E}\left[\frac{(d-p(\mathbf{x}))y}{p(\mathbf{x})(1-p(\mathbf{x}))}\middle|\mathbf{x}\right], \text{ so}$$

$$\tau_{\text{ATE}} := \mathbb{E}[\tau_{\text{CATE}}(\mathbf{x})] = \mathbb{E}\left[\frac{(d-p(\mathbf{x}))y}{p(\mathbf{x})(1-p(\mathbf{x}))}\right]. \tag{3}$$

Now, notice that[9]

$$\frac{(d-p(\mathbf{x}))y}{1-p(\mathbf{x})} = \frac{(d-p(\mathbf{x}))y_0}{1-p(\mathbf{x})} + d(y_1 - y_0).$$

We now show that $\mathbb{E}[(d-p(\mathbf{x}))y_0/(1-p(\mathbf{x}))|\mathbf{x}] = 0$ by showing that

$$\begin{aligned}\mathbb{E}\left[(d-p(\mathbf{x}))y_0\middle|\mathbf{x}\right] &= \mathbb{E}\left[\mathbb{E}\left[(d-p(\mathbf{x}))y_0\middle|\mathbf{x},d\right]\middle|\mathbf{x}\right] = \mathbb{E}\left[(d-p(\mathbf{x}))\mathbb{E}[y_0|\mathbf{x},d]\middle|\mathbf{x}\right] \\ &= \mathbb{E}\left[(d-p(\mathbf{x}))\mathbb{E}[y_0|\mathbf{x}]\middle|\mathbf{x}\right] = \{\mathbb{E}[d|\mathbf{x}] - p(\mathbf{x})\}\mathbb{E}[y_0|\mathbf{x}] = 0.\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}\left[\frac{(d-p(\mathbf{x}))y}{1-p(\mathbf{x})}\middle|\mathbf{x}\right] &= \mathbb{E}\left[d(y_1-y_0)\middle|\mathbf{x}\right], \text{ and} \\ \mathbb{E}\left[\frac{(d-p(\mathbf{x}))y}{1-p(\mathbf{x})}\right] &= \mathbb{E}\left[d(y_1-y_0)\right].\end{aligned} \tag{4}$$

Now

$$\begin{aligned}\mathbb{E}\left[d(y_1-y_0)\right] &= \mathbb{E}[\mathbb{E}\left[d(y_1-y_0)\middle|d\right]] \\ &= \mathbb{E}\left[d(y_1-y_0)\middle|d=1\right]\Pr\{d=1\} + \mathbb{E}\left[d(y_1-y_0)\middle|d=0\right]\Pr\{d=0\} \\ &= \mathbb{E}\left[(y_1-y_0)\middle|d=1\right]\Pr\{d=1\},\end{aligned}$$

so replacing this on the right-hand side of (**??**) one obtains

$$\begin{aligned}\mathbb{E}\left[\frac{(d-p(\mathbf{x}))y}{1-p(\mathbf{x})}\right] &= \mathbb{E}\left[(y_1-y_0)\middle|d=1\right]\Pr\{d=1\} \\ \tau_{\text{ATT}} &= \mathbb{E}\left[\frac{(d-p(\mathbf{x}))y}{\rho[1-p(\mathbf{x})]}\right], \text{ where } \rho := \Pr\{d=1\}.\end{aligned} \tag{5}$$

Estimation:

Let $p(\mathbf{x},\delta)$ be a Logit or Probit model estimated via (Conditional) maximum likelihood, i.e. $\widehat{\delta}$, then

$$\begin{aligned}\widehat{\tau}_{\text{ATE}} &= n^{-1}\sum_{i=1}^{n}\left[\frac{(d_i - p(\mathbf{x}_i,\widehat{\delta}))y_i}{p(\mathbf{x}_i,\widehat{\delta})(1-p(\mathbf{x}_i,\widehat{\delta}))}\right], \\ \widehat{\tau}_{\text{ATT}} &= n^{-1}\sum_{i=1}^{n}\left[\frac{(d_i - p(\mathbf{x}_i,\widehat{\delta}))y_i}{\widehat{\rho}[1-p(\mathbf{x}_i,\widehat{\delta})]}\right].\end{aligned}$$

and we have already discussed the way to obtain expressions for the asymptotic variances.

[9] Write

$$\begin{aligned}(d-p(\mathbf{x}))y &= [d-p(\mathbf{x})][y_0 + d(y_1-y_0)] \\ &= [d-p(\mathbf{x})]y_0 + d[d-p(\mathbf{x})](y_1-y_0) \\ &= [d-p(\mathbf{x})]y_0 + d[1-p(\mathbf{x})](y_1-y_0).\end{aligned}$$

If $\widehat{p}(\mathbf{x}_i)$ represents a consistent estimator of $p(\mathbf{x})$, and we estimate ρ by $\widehat{\rho} = n^{-1}\sum_{j=1}^{n} d_j$, then using the entire random sample of size n one has by the analogy principle

$$\widehat{\tau}_{\mathrm{ATE}} = n^{-1}\sum_{i=1}^{n}\left[\frac{(d_i - \widehat{p}(\mathbf{x}_i))y_i}{\widehat{p}(\mathbf{x}_i)(1-\widehat{p}(\mathbf{x}_i))}\right],$$

$$\widehat{\tau}_{\mathrm{ATT}} = n^{-1}\sum_{i=1}^{n}\left[\frac{(d_i - \widehat{p}(\mathbf{x}_i))y_i}{\widehat{\rho}[1-\widehat{p}(\mathbf{x}_i)]}\right].$$