Contents

2
Classical linear models

Linear models are ubiquitous due to their utility (even for addressing elements of nonlinear processes). This chapter briefly addresses foundational ideas including projections, conditional expectation functions, analysis of variance (*ANOVA*), analysis of covariance (*ANCOVA*), linear regression, and omitted correlated variables.

2.1 A basic example

Consider a simple example. Suppose we're looking for a solution to

$$Y = \alpha$$

where α is a constant, Y takes the values $\{Y_1 = 4, Y_2 = 6, Y_3 = 5\}$, and order is exchangeable. Clearly, there is no exact solution for α. How do we proceed? One approach is to consider what is unobserved or unknown in the response or outcome variable Y to be error $\{\varepsilon_1, \varepsilon_2, \varepsilon_3\}$ and to guess the parameters of interest (in this case, α) in a manner that extracts all that we know (say, summarized by X)[1] and leaves nothing known in the error. In other words, we're looking for the conditional expectation function (*CEF*)

[1] X represents what we know. In the above equation X is implicitly a vector of ones.

$E[Y \mid X]$. Extraction of all information in X, implies error cancellation or $E[\varepsilon \mid X] = 0$.[2]

2.1.1 Data generating process (DGP)

If we believe the errors have common variability, say $Var[\varepsilon_i] = \sigma^2$,[3] we envision the data generating process (DGP) is

$$\begin{aligned} Y_1 &= X_1\alpha + \varepsilon_1 = 1\alpha + \varepsilon_1 \\ Y_2 &= X_2\alpha + \varepsilon_2 = 1\alpha + \varepsilon_2 \\ Y_3 &= X_3\alpha + \varepsilon_3 = 1\alpha + \varepsilon_3 \end{aligned}$$

or in compact matrix form

$$Y = X\alpha + \varepsilon$$

where

$$\begin{aligned} Y &= \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix}, \; X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \\ \varepsilon &= \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix} \sim N\left(0, \sigma^2 I\right), \text{ and } E[\varepsilon \mid X] = 0 \end{aligned}$$

I is an $n \times n$ identity matrix, n is the sample size or number of observations, and $N(\cdot)$ refers to the normal distribution with first term equal to the mean vector and the second term is the variance-covariance matrix.[4] Notice,

$$Var[\varepsilon] = \sigma^2 I$$

is a very compact form and implies

$$\begin{aligned} Var[\varepsilon] &= E\left[(\varepsilon - E[\varepsilon])(\varepsilon - E[\varepsilon])^T\right] \\ &= \begin{bmatrix} Var[\varepsilon_1] & Cov[\varepsilon_1, \varepsilon_2] & Cov[\varepsilon_1, \varepsilon_3] \\ Cov[\varepsilon_1, \varepsilon_2] & Var[\varepsilon_2] & Cov[\varepsilon_2, \varepsilon_2] \\ Cov[\varepsilon_1, \varepsilon_3] & Cov[\varepsilon_2, \varepsilon_3] & Var[\varepsilon_3] \end{bmatrix} \\ &= \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix} \end{aligned}$$

---

[2] A complete statement of the result, the *CEF* decomposition theorem, and its proof can be found in the appendix to chapter 3 of *Accounting and Causal Effects: Econometric Challenges*.

[3] Knowledge of the variance leads to Gaussian or normal probability assignment by the maximum entropy principle (*MEP*). For details, see the discussion in chapter 13 of *Accounting and Causal Effects: Econometric Challenges*, or Jaynes [2003].

[4] See the appendix for a discussion of linear algebra basics.

where $Var[\cdot]$ is variance and $Cov[\cdot]$ is covariance.

2.2 Estimation

We estimate

$$Y = Xa + e$$

where a is an estimate of the unknown α and $e = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}$ estimates the unknowns ε. Since we're searching for a good approximation to the *CEF*, e is constructed to be unrelated to X, or as we say, orthogonal, $X^T e = 0$. That is, every column of X is constructed to be orthogonal or perpendicular to the residuals e. Since $e = Y - Xa$, we have

$$X^T e = X^T(Y - Xa) = 0$$

this orthogonality condition leads naturally to the normal equations

$$X^T Xa = X^T Y$$

and multiplication of both sides by the inverse gives the estimator for α

$$\begin{aligned} \left(X^T X\right)^{-1} X^T Xa &= \left(X^T X\right)^{-1} X^T Y \\ a &= \left(X^T X\right)^{-1} X^T Y \end{aligned}$$

For our example above, we have a sample size $n = 3$, $\left(X^T X\right)^{-1} = \frac{1}{n} = \frac{1}{3}$, and $X^T Y = \sum_{i=1}^{n} Y_i$. Hence, $a = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}$, the sample average as intuition suggests. For the present example, then $a = 5$ and $Var[a] = \frac{\sigma^2}{3}$.[5]

Further, (i) $E[a \mid X] = E[a] = \alpha$ (estimation is unbiased) and (ii) variation in the estimator is smallest amongst unbiased estimators with $Var[a \mid X] = \sigma^2 \left(X^T X\right)^{-1}$. To see this, (i)

$$\begin{aligned} E[a \mid X] &= E\left[\left(X^T X\right)^{-1} X^T Y \mid X\right] \\ &= E\left[\left(X^T X\right)^{-1} X^T (X\alpha + \varepsilon) \mid X\right] \\ &= E\left[\left(X^T X\right)^{-1} X^T X\alpha + \left(X^T X\right)^{-1} X^T \varepsilon \mid X\right] \\ &= \alpha + \left(X^T X\right)^{-1} X^T E[\varepsilon \mid X] \\ &= \alpha + 0 = \alpha \end{aligned}$$

[5] Variance of the estimator is discussed below.

By iterated expectations,[6] the unconditional expectation of the estimator, a, also equals the unknown parameter of interest, α

$$\begin{aligned} E_X\left[E\left[a \mid X\right]\right] &= E\left[a\right] \\ E_X\left[E\left[a \mid X\right]\right] &= E_X\left[\alpha\right] = \alpha \\ E\left[a\right] &= \alpha \end{aligned}$$

(ii)

$$\begin{aligned} Var\left[a \mid X\right] &= E\left[\left(a - E\left[a \mid X\right]\right)\left(a - E\left[a \mid X\right]\right)^T \mid X\right] \\ &= E\left[\left(a-\alpha\right)\left(a-\alpha\right)^T \mid X\right] \\ &= E\left[\left(\left(X^TX\right)^{-1}X^TY - \alpha\right)\left(\left(X^TX\right)^{-1}X^TY - \alpha\right)^T \mid X\right] \\ &= E\left[\left(\alpha + \left(X^TX\right)^{-1}X^T\varepsilon - \alpha\right)\left(\alpha + \left(X^TX\right)^{-1}X^T\varepsilon - \alpha\right)^T \mid X\right] \\ &= E\left[\left(\left(X^TX\right)^{-1}X^T\varepsilon\right)\left(\left(X^TX\right)^{-1}X^T\varepsilon\right)^T \mid X\right] \\ &= E\left[\left(X^TX\right)^{-1}X^T\varepsilon\varepsilon^TX\left(X^TX\right)^{-1} \mid X\right] \\ &= \left(X^TX\right)^{-1}X^TE\left[\varepsilon\varepsilon^T \mid X\right]X\left(X^TX\right)^{-1} \end{aligned}$$

Since $E\left[\varepsilon\varepsilon^T \mid X\right] = \sigma^2 I$, the above simplifies to yield the result as claimed above.[7]

$$\begin{aligned} Var\left[a \mid X\right] &= \sigma^2\left(X^TX\right)^{-1}X^TX\left(X^TX\right)^{-1} \\ &= \sigma^2\left(X^TX\right)^{-1} \end{aligned}$$

2.3 Projection matrix

The conditional expectation function is estimated as

$$\begin{aligned} \widehat{Y} &= Xa \\ &= X\left(X^TX\right)^{-1}X^TY \end{aligned}$$

The leading matrix, $X\left(X^TX\right)^{-1}X^T$, is so important it warrants special designation. It is the projection matrix, $P_X = X\left(X^TX\right)^{-1}X^T$. Notice,

[6] A proof of the law of iterated expectations is presented in the appendix.

[7] A complete demonstration of the minimum variance property can be found in the discussion of the Gauss-Markov theorem in chapter 3 of *Accounting and Causal Effects: Econometric Challenges*.

the matrix is *symmetric*, $P_X = (P_X)^T$ and it is *idempotent*. That is, multiplication by itself leaves it unchanged.

$$\begin{aligned} P_X P_X &= P_X \\ X\left(X^TX\right)^{-1}X^TX\left(X^TX\right)^{-1}X^T &= X\left(X^TX\right)^{-1}X^T \end{aligned}$$

This property says if a vector resides in the columnspace of X, it is a linear combination of the columns of X, then projecting the vector onto the columns of X leaves it unchanged — which matches our intuition. Further, the residuals are orthogonal to the columns of X, $e = Y - P_XY = (I - P_X)Y = M_XY$, and $M_XP_X = (I - P_X)P_X = P_X - P_X = 0$. Therefore, the residuals reside in the orthogonal subspace to the columnspace; this subspace is called the left nullspace.[8]

2.4 Different means (ANOVA)

We've explored estimation of an unknown mean in the example above and discovered that the best guess, in a minimum mean squared error or least squares sense, for the conditional expectation function is the sample average. Now, suppose we have a bit more information. We know that outcome is treated or not treated. Denote this by $D = 1$ for treatment and $D = 0$ for not treated. This suggests we're interested in $\alpha_1 = E[Y \mid D = 1]$ and $\alpha_0 = E[Y \mid D = 0]$ or we're interested in $\beta = E[Y \mid D = 1] - E[Y \mid D = 0]$. In other words, we're interested in two means and, intuitively, we estimate these via two sample averages or their difference. This setting is often referred to as analysis of variance or *ANOVA*, for short; this is the simplest case — a single factor, two factor-level *ANOVA*.

In the former (two mean) case, it's simplest and most direct to envision the following *DGP*

$$\begin{aligned} Y &= D_0\alpha_0 + D\alpha_1 + \varepsilon \\ &= X_1\alpha + \varepsilon \end{aligned}$$

where $X_1 = \begin{bmatrix} D_0 & D \end{bmatrix}$ (an $n \times 2$ matrix), $\alpha = \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix}$ (a two element parameter vector), and $D_0 = 1 - D$. While in the latter (mean difference)

[8] The fundamental theorem of linear algebra has two parts. The first part says that every $m \times n$ (rows by columns) matrix has the same number of linearly independent rows and columns, call this number r. The second part says the dimension (number of linearly independent vectors) of the rowspace, r, plus the dimension of its orthogonal subspace, the nullspace, $n - r$, spans all n length vectors. Analogously for the columnspace, the dimension of the columnspace, r, plus the dimension of the left nullspace, $m - r$, spans all m element vectors. See the appendix for more extensive discussion.

case, it's simplest and most direct to envision the DGP as

$$\begin{aligned} Y &= \alpha_0 + D\beta + \varepsilon \\ &= X_2\gamma + \varepsilon \end{aligned}$$

where $X_2 = \begin{bmatrix} \iota & D \end{bmatrix}$ (an $n \times 2$ matrix), $\gamma = \begin{bmatrix} \alpha_0 \\ \beta \end{bmatrix}$ (a two element parameter vector), and ι is a vector of n ones.

Of course, we can work with either one and derive all results. For example, $\beta = \begin{bmatrix} -1 & 1 \end{bmatrix} \alpha = \alpha_1 - \alpha_0$. Therefore, β is estimated via

$$b = \begin{bmatrix} -1 & 1 \end{bmatrix} a = a_1 - a_0$$

and

$$\begin{aligned} Var[b \mid X_1] &= \begin{bmatrix} -1 & 1 \end{bmatrix} Var[a \mid X_1] \begin{bmatrix} -1 \\ 1 \end{bmatrix} \\ &= Var[a_0 \mid X_1] + Var[a_1 \mid X_1] - 2Cov[a_0, a_1 \mid X_1] \end{aligned}$$

where $a = \begin{bmatrix} a_0 \\ a_1 \end{bmatrix}$, a_0 is the estimator for α_0, and a_1 is the estimator for α_1. Also, $\alpha_1 = \begin{bmatrix} 1 & 1 \end{bmatrix} \gamma = \alpha_0 + \beta = \alpha_0 + \alpha_1 - \alpha_0$. Hence, α_1 is estimated via

$$\begin{aligned} a_1 &= \begin{bmatrix} 1 & 1 \end{bmatrix} g \\ &= \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} a_0 \\ b \end{bmatrix} \\ &= a_0 + b \end{aligned}$$

and

$$\begin{aligned} Var[a_1 \mid X_2] &= \begin{bmatrix} 1 & 1 \end{bmatrix} Var[g \mid X_2] \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &= Var[a_0 \mid X_2] + Var[b \mid X_2] + 2Cov[a_0, b \mid X_2] \end{aligned}$$

The bigger point here is that estimation of the parameters to "best" approximate the conditional expectation function is achieved in the same manner as above (via orthogonalization of the residuals and what is known, X).

$$\begin{aligned} a &= \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} \\ &= \left(X_1^T X_1\right)^{-1} X_1^T Y \end{aligned}$$

and

$$\begin{aligned} Var[a \mid X_1] &= \begin{bmatrix} Var[a_0 \mid X_1] & Cov[a_0, a_1 \mid X_1] \\ Cov[a_0, a_1 \mid X_1] & Var[a_1 \mid X_1] \end{bmatrix} \\ &= \sigma^2 \left(X_1^T X_1\right)^{-1} \end{aligned}$$

Also,

$$\begin{aligned} g &= \begin{bmatrix} a_0 \\ b \end{bmatrix} \\ &= \left(X_2^T X_2\right)^{-1} X_2^T Y \end{aligned}$$

and

$$\begin{aligned} Var\left[g \mid X_2\right] &= \begin{bmatrix} Var\left[a_0 \mid X_2\right] & Cov\left[a_0, b \mid X_b\right] \\ Cov\left[a_0, b \mid X_2\right] & Var\left[b \mid X_2\right] \end{bmatrix} \\ &= \sigma^2 \left(X_2^T X_2\right)^{-1} \end{aligned}$$

2.4.1 ANOVA example 1

Suppose $(Y \mid D = 0)$ is the same as Y in the previous example, that is,

$$\{Y_1 = 4, Y_2 = 6, Y_3 = 5 \mid D = 0\}$$

and $(Y \mid D = 1)$ is

$$\{Y_4 = 11, Y_5 = 9, Y_6 = 10 \mid D = 1\}$$

with order exchangeable conditional on D. The estimated regression function is

$$E\left[Y \mid X_1\right] = 5D_0 + 10D$$

with

$$Var\left[a \mid X_1\right] = \sigma^2 \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}^{-1} = \frac{\sigma^2}{3} I$$

or

$$E\left[Y \mid X_2\right] = 5 + 5D$$

with

$$\begin{aligned} Var\left[g \mid X_2\right] &= \sigma^2 \begin{bmatrix} 6 & 3 \\ 3 & 3 \end{bmatrix}^{-1} \\ &= \frac{\sigma^2}{9} \begin{bmatrix} 3 & -3 \\ -3 & 6 \end{bmatrix} \\ &= \frac{\sigma^2}{3} \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \end{aligned}$$

From the first regression, the estimate of β is

$$b = \begin{bmatrix} -1 & 1 \end{bmatrix} a = -5 + 10 = 5$$

with

$$Var\left[b \mid X_1\right] = \begin{bmatrix} -1 & 1 \end{bmatrix} Var\left[a\right] \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \frac{2}{3}\sigma^2$$

which, of course, corresponds with the results from the second regression (the second element of g is $b = 5$, the coefficient on D, and $Var[b \mid X_2] = \frac{2}{3}\sigma^2$, the second row and second column element of $Var[g \mid X_2]$). Similarly, the estimate of α_1, from the first regression is $a_1 = 10$ with $Var[a_1 \mid X_1] = \frac{\sigma^2}{3}$, and, from the second regression

$$a_1 = \begin{bmatrix} 1 & 1 \end{bmatrix} g = 5 + 5 = 10$$

with

$$\begin{aligned} Var[a_1 \mid X_2] &= \begin{bmatrix} 1 & 1 \end{bmatrix} Var[g \mid X_2] \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &= \frac{\sigma^2}{3}(1 + 2 - 1 - 1) = \frac{\sigma^2}{3} \end{aligned}$$

Of course, the estimate of α_0 is directly available from either regression, $a_0 = 5$, with $Var[a_0 \mid X_1] = Var[a_0 \mid X_2] = \frac{\sigma^2}{3}$.

2.4.2 Multi-factor ANOVA and interactions

What if we know of other factors that may, in some way, be related to outcome? Then, the consistent approach is to include them in the analysis (to guard against omitted, correlated variables or Simpson's paradox). For simplicity, suppose we have another binary factor denoted $W = \{0, 1\}$. A saturated model includes W along with D and their product or interaction, $(D \times W)$. We envision the following *DGP.*

$$\begin{aligned} Y &= \alpha_0 + \beta D + \omega W + \delta(D \times W) + \varepsilon \\ &= X\gamma + \varepsilon \end{aligned}$$

where the regression or conditional expectation is

$$\begin{aligned} E[Y \mid X] &= \alpha_0 + \beta D + \omega W + \delta(D \times W) \\ &= X\gamma \end{aligned}$$

$\varepsilon \sim N\left(0, \sigma^2 I\right)$, $E[\varepsilon \mid X] = 0$, $X = \begin{bmatrix} \iota & D & W & (D \times W) \end{bmatrix}$ is an $n \times 4$ design matrix, and $\gamma = \begin{bmatrix} \alpha_0 \\ \beta \\ \omega \\ \delta \end{bmatrix}$.

Even though this is a richer *DGP*, estimation proceeds as before. That is, the minimum mean square error or least squares estimator for γ is $g = \left(X^T X\right)^{-1} X^T Y$, a four element vector, and its variability is summarized as $Var[g \mid X] = \sigma^2 \left(X^T X\right)^{-1}$, a 4×4 matrix, using the $(n \times 4)$ X matrix identified above.

ANOVA example 2

Suppose we continue the previous example by appending W.

Y	D	W
4	0	0
6	0	1
5	0	0
11	1	1
9	1	0
10	1	0

The estimated regression is

$$E[Y \mid X] = 4.5 + 5D + 1.5W + 0\,(D \times W)$$

An intuitive interpretation is D partitions Y into $\{4, 6, 5\}$ and $\{11, 9, 10\}$, as before, but W partitions $\{6, 11\}$ and $\{4, 5, 9, 10\}$, and $D \times W$ partitions $\{11\}$ and $\{4, 6, 5, 9, 10\}$. Hence, the coefficient on D is the mean difference between $\{9, 10\}$ and $\{4, 5\}$, that is, after conditioning on W,[9] leaving $9.5 - 4.5 = 5$. The coefficient on W, conditional on D, is the mean difference between $\{4, 5\}$ and $\{6\}$, or $6 - 4.5 = 1.5$, and $\{9, 10\}$ and $\{11\}$, or $11 - 9.5 = 1.5$. Since $(D \times W)$ separates $Y = 11$ from the rest but that difference is already explained by $(W \mid D)$, the coefficient on $(D \times W)$ is zero.

Perhaps, some elaboration is instructive.

$$E[Y \mid D = 0, W = 1] = 4.5 + 0 + 1.5 = 6$$

there is no residual associated with these conditions (this combination of D and W only occurs when $Y = 6$, in the sample). Similarly,

$$E[Y \mid D = 1, W = 1] = 4.5 + 5 + 1.5 = 11$$

there is no residual associated with these conditions (this combination of D and W only occurs when $Y = 11$, in the sample). On the other hand,

$$E[Y \mid D = 0, W = 0] = 4.5 + 0 + 0 = 4.5$$

there is some residual associated with these conditions (this combination of D and W occurs when $Y = 4$ or 5, and they occur with equal frequency in the sample). To complete the picture, we have

$$E[Y \mid D = 1, W = 0] = 4.5 + 5 + 0 = 9.5$$

there is some residual associated with these conditions (this combination of D and W occurs when $Y = 9$ or 10, and they occur with equal frequency in the sample).

---

[9] This is a key to understanding regression, each explanatory (RHS) variable contributes toward explaining response conditional on the other variables on the RHS.

ANOVA example 3

Now, suppose we perturb the above example slightly by altering W.

Y	D	W
4	0	0
6	0	0
5	0	1
11	1	1
9	1	0
10	1	0

The estimated regression is

$$E[Y \mid X] = 5 + 4.5D + 0W + 1.5(D \times W)$$

Similar arguments to those above provide some intuition.

$$E[Y \mid D = 0, W = 1] = 5 + 0 + 0 = 5$$

there is no residual associated with these conditions (this combination of D and W only occurs when $Y = 5$, in the sample). Similarly,

$$E[Y \mid D = 1, W = 1] = 5 + 4.5 + 1.5 = 11$$

there is no residual associated with these conditions (this combination of D and W only occurs when $Y = 11$, in the sample). On the other hand,

$$E[Y \mid D = 0, W = 0] = 5 + 0 + 0 = 5$$

there is some residual associated with these conditions (this combination of D and W occurs when $Y = 4$ or 6, and they occur with equal frequency in the sample). Finally, we have

$$E[Y \mid D = 1, W = 0] = 5 + 4.5 + 0 = 9.5$$

there is some residual associated with these conditions (this combination of D and W occurs when $Y = 9$ or 10, and they occur with equal frequency in the sample).

Notice, unlike the first two-factor example, if we don't include the interaction term we estimate

$$E[Y \mid X] = 4.75 + 5D + 0.75W$$

The estimated mean effects are different since the data is partitioned incompletely via the design matrix, $X = \begin{bmatrix} D & W \end{bmatrix}$, given what we know,

$$\begin{bmatrix} D & W & (D \times W) \end{bmatrix}$$

That is, this design matrix imposes a pooling restriction.[10] Consistency requires such pooling restrictions satisfy the equality of

$$E\left[Y \mid D=1, W=0\right]-E\left[Y \mid D=0, W=0\right]$$

and

$$E\left[Y \mid D=1, W=1\right]-E\left[Y \mid D=0, W=1\right]$$

as well as

$$E\left[Y \mid W=1, D=0\right]-E\left[Y \mid W=0, D=0\right]$$

and

$$E\left[Y \mid W=1, D=1\right]-E\left[Y \mid W=0, D=1\right]$$

Therefore, even though $E\left[Y \mid D=0, W=1\right]$ is uniquely associated with $Y=5$, the pooling restriction produces a residual

$$(e \mid D=0, W=1)=5-5.50=-0.50$$

Likewise, while $E\left[Y \mid D=1, W=1\right]$ is uniquely associated with $Y=11$, the pooling restriction improperly produces a residual

$$(e \mid D=1, W=1)=11-10.50=0.50$$

Also, $E\left[Y \mid D=0, W=0\right]$ is associated with $Y=\{4,6\}$, the pooling restriction produces residuals

$$(e \mid D=0, W=0)=4-4.75=-0.75$$

and

$$6-4.75=1.25$$

Finally, $E\left[Y \mid D=1, W=0\right]$ is associated with $Y=\{9,10\}$, the pooling restriction produces residuals

$$(e \mid D=1, W=0)=9-9.75=-0.75$$

and

$$10-9.75=0.25$$

In other words, inefficient error cancellation. Of course, by construction (orthogonality between the vector of ones for the intercept, the first column of X, and the residuals), the residuals sum to zero. Keeping in mind that *ANOVA* is a partitioning exercise crystallizes the implications of inappropriate pooling restrictions on the design matrix, X. Or equivalently, the implications of failing to fully utilize what we know,

$$\left[\begin{array}{ccc} D & W & (D \times W) \end{array}\right]$$

when estimating conditional expectations.

---

[10] Pooling restrictions are attractive as they allow, when appropriate, the data to be summarized with fewer parameters.

2.5 Omitted, correlated variables

The above example illustrates our greatest concern with conditional expectations or regression models. If we leave out a regressor (explanatory variable) it's effectively absorbed into the error term. While this increases residual uncertainty, which is unappealing, this is not the greatest concern. Recall the key condition for regression is $E[\varepsilon \mid X] = 0$. If this is violated, all inferences are at risk.

To illustrate the implications, return to the ANOVA examples. Let

$$\begin{aligned} X &= \begin{bmatrix} \iota & D & W & (D \times X) \end{bmatrix} \\ &= \begin{bmatrix} X_2 & x_3 \end{bmatrix} \end{aligned}$$

where $X_2 = \begin{bmatrix} \iota & D & W \end{bmatrix}$ and $x_3 = (D \times X)$. Suppose the DGP is

$$Y = X\beta + \varepsilon \quad \varepsilon \sim N\left(0, \sigma^2 I\right) \quad E[\varepsilon \mid X] = 0$$

or

$$Y = X_2\beta_2 + x_3\beta_3 + \varepsilon$$

When we estimate

$$Y = X_2 b_2 + residuals$$

by orthogonal construction,

$$\begin{aligned} b_2 &= \left(X_2^T X_2\right)^{-1} X_2^T Y \\ &= \left(X_2^T X_2\right)^{-1} X_2^T (X\beta + \varepsilon) \\ &= \left(X_2^T X_2\right)^{-1} X_2^T (X_2\beta_2 + x_3\beta_3 + \varepsilon) \\ &= \beta_2 + \left(X_2^T X_2\right)^{-1} X_2^T x_3\beta_3 + \left(X_2^T X_2\right)^{-1} X_2^T \varepsilon \end{aligned}$$

The last term is no problem as in large samples it tends to zero by $E[\varepsilon \mid X] = 0$. Our concern lies with the second term, $\left(X_2^T X_2\right)^{-1} X_2^T x_3\beta_3$. This term is innocuous if either $X_2^T x_3$ tends to zero in large samples (in other words, x_3 is uncorrelated with the other regressors), or $\beta_3 = 0$ (in other words, the third term was not a part of the DGP). Notice, this is extremely important, any correlation between the omitted regressor and the other regressors (for $\beta_3 \neq 0$) biases all of the estimates included in the model. The extent of the bias in b_2 is

$$bias(b_2) = \left(X_2^T X_2\right)^{-1} X_2^T x_3\beta_3$$

In ANOVA example 3, without x_3 we estimate

$$E[Y \mid X_2] = 4.75 + 5D + 0.75W$$

The bias in the parameter estimates is

$$\begin{aligned} bias\left(b_2\right) &= \left(X_2^T X_2\right)^{-1} X_2^T x_3 \beta_3 \\ &= \frac{1}{12}\begin{bmatrix} 5 & -4 & -3 \\ -4 & 8 & 0 \\ -3 & 0 & 9 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} 1.5 \\ &= \begin{bmatrix} -0.25 \\ 0.5 \\ 0.75 \end{bmatrix} \end{aligned}$$

Hence, to recover the parameters of interest (assuming our estimates are based on a representative sample of the population) subtract the bias from the above estimates and concatenate the missing parameter, β_3,

$$\begin{aligned} \beta_2 &= b_2 - bias\left(b_2\right) \\ &= \begin{bmatrix} 4.75 \\ 5 \\ 0.75 \end{bmatrix} - \begin{bmatrix} -0.25 \\ 0.5 \\ 0.75 \end{bmatrix} = \begin{bmatrix} 5 \\ 4.5 \\ 0 \end{bmatrix} \end{aligned}$$

And, with concatenation of β_3 we have

$$\beta = \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 4.5 \\ 0 \\ 1.5 \end{bmatrix}$$

Why doesn't this problem plague ANOVA example 2? Is it because $X_2^T x_3$ tends to zero? No, this is the same as ANOVA example 3. The reason is that the DGP is an unusual special case that excludes $x_3 = (D \times W)$ as $\beta_3 = 0$.

ANOVA example 4

Once more, suppose we perturb the above example by altering W.

Y	D	W
4	0	0
6	0	1
5	0	1
11	1	0
9	1	0
10	1	1

The estimated regression is

$$E\left[Y \mid X\right] = 4 + 6D + 1.5W - 1.5\left(D \times W\right)$$

Again, intuition follows from conditional expectations.

$$E[Y \mid D = 0, W = 1] = 4 + 0 + 1.5 = 5.5$$

this combination of D and W pools $Y = \{5, 6\}$, in the sample. While for

$$E[Y \mid D = 1, W = 1] = 4 + 6 + 1.5 - 1.5 = 10$$

there is no residual (this combination of D and W only occurs when $Y = 10$, in the sample). Also, for

$$E[Y \mid D = 0, W = 0] = 4 + 0 + 0 + 0 = 4$$

there is no residual (this combination of D and W occurs only when $Y = 4$, in the sample). Finally, we have

$$E[Y \mid D = 1, W = 0] = 4 + 6 + 0 + 0 = 10$$

there is some residual associated with these conditions as this combination of D and W pools $Y = 9$ or 11, and they occur with equal frequency in the sample.

Notice, if we don't include the interaction term we estimate

$$E[Y \mid X] = 4.5 + 5.25D + 0.75W$$

Again, the estimated mean effects are different since the design matrix, $X = \begin{bmatrix} D & W \end{bmatrix}$, incompletely partitions what we know,

$$\begin{bmatrix} D & W & (D \times W) \end{bmatrix}$$

and pooling restrictions require

$$E[Y \mid D = 1, W = 0] - E[Y \mid D = 0, W = 0]$$

and

$$E[Y \mid D = 1, W = 1] - E[Y \mid D = 0, W = 1]$$

to be equal as well as

$$E[Y \mid W = 1, D = 0] - E[Y \mid W = 0, D = 0]$$

and

$$E[Y \mid W = 1, D = 1] - E[Y \mid W = 0, D = 1]$$

to be equal.

Therefore, even though $E[Y \mid D = 1, W = 1]$ is uniquely associated with $Y = 10$, the pooling restriction inappropriately produces a residual

$$(e \mid D = 1, W = 1) = 10 - 10.50 = -0.50$$

Also, while $E[Y \mid D = 0, W = 1]$ is associated with $Y = \{5, 6\}$, the pooling restriction produces residuals

$$(e \mid D = 0, W = 1) = 5 - 5.25 = -0.25$$

and

$$6 - 5.25 = 0.75$$

Further, $E[Y \mid D = 0, W = 0]$ is uniquely associated with $Y = 4$, and the pooling restriction produces a residual

$$(e \mid D = 0, W = 0) = 4 - 4.5 = -0.5$$

Finally, $E[Y \mid D = 1, W = 0]$ is associated with $Y = \{9, 11\}$, and the pooling restriction produces residuals

$$(e \mid D = 1, W = 0) = 9 - 9.75 = -0.75$$

and

$$11 - 9.75 = 1.25$$

Again, by construction, the residuals sum to zero.

The *DGP* for ANOVA example 4 involves a different design matrix, X, than examples 2 and 3. Nonetheless the omitted, correlated variable bias stems from the analogous source. For ANOVA example 4 the bias is

$$\begin{aligned} bias(b_2) &= \left(X_2^T X_2\right)^{-1} X_2^T x_3 \beta_3 \\ &= \frac{1}{12} \begin{bmatrix} 8 & -6 & -6 \\ -6 & 9 & 3 \\ -6 & 3 & 9 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} (-1.5) \\ &= \begin{bmatrix} 0.5 \\ -0.75 \\ -0.75 \end{bmatrix} \end{aligned}$$

2.6 Linear regression

How do we proceed if we perceive outcome is related to explanatory variables and these variables are not binary but rather have continuous support? Let X denote an $n \times p$ matrix of explanatory variables, say $X_1, \ldots, X_{p-1}$, plus a vector of ones in the first column for the intercept. Now, we envision a *DGP* like $E[Y \mid X] = m(X) + \varepsilon$, where $m(X)$ is some function of X, $\varepsilon \sim N\left(0, \sigma^2 I\right)$, and $E[\varepsilon \mid X] = 0$. If the functional form of $m(X)$ is unknown (as is frequently the case), we often approximate $m(X)$ with a linear function, $X\beta$, where β is a p-elemen parameter vector. Further, the minimum mean squared error or least squares solution among linear functions (i.e., linear in the parameters) is the same as that above. That is, β

is estimated via $b = \left(X^T X\right)^{-1} X^T Y$ with $Var\left[b \mid X\right] = \sigma^2 \left(X^T X\right)^{-1}$, and the estimated regression or estimated conditional expectation function is $\widehat{Y} = Xb = P_X Y$.[11]

2.6.1 Example

It's time for an example. Continue with the running example except treatment, D, is initially unobserved.[12] Rather, we observe X along with outcome, Y. Suppose we have the following data.

Y	X
4	−1
6	1
5	0
11	1
9	−1
10	0

We envision the *DGP*

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where $\varepsilon \sim N\left(0, \sigma^2 I\right)$ and $E\left[\varepsilon \mid X\right] = 0$. The estimated regression is

$$E\left[Y \mid X\right] = 7.5 + 1.0X$$

where predicted and residual values are as follows.

$predicted\left(\widehat{Y}\right)$	$residuals\,(e)$
6.5	−2.5
8.5	−2.5
7.5	−2.5
8.5	2.5
6.5	2.5
7.5	2.5

Again, by construction, the sum of the residuals is zero and the average predicted value equals the sample average, $\overline{Y}$. Within each cluster (the first three and the last three observations), X perfectly explains the response, however there is no basis for the regression to distinguish the clusters. If treatment, D, is observed, then in combination with X we can perfectly explain observed outcome. Such a model is sometimes labelled analysis of covariance, or *ANCOVA*, for short.

---

[11] See the appendix to explore a more general case — generalized least squares (*GLS*).
[12] In this example, factor W is out of the picture.

2.6.2 Analysis of covariance

The *ANCOVA* label stems from combining the mean effects associated with *ANOVA* and covariates, X, which explain outcome. For the setting above, we envision the *DGP*

$$Y = \delta_0 + \delta_1 D + \delta_2 X + \varepsilon$$

or, in saturated form,

$$Y = \delta_0 + \delta_1 D + \delta_2 X + \delta_3 (D \times X) + \varepsilon$$

where $\varepsilon \sim N\left(0, \sigma^2 I\right)$ and $E[\varepsilon \mid X] = 0$. Suppose the data above is augmented by D, we have

Y	D	X
4	0	-1
6	0	1
5	0	0
11	1	1
9	1	-1
10	1	0

The estimated *ANCOVA* regression is

$$E[Y \mid D, X] = 5.0 + 5.0D + 1.0X + 0.0 (D \times X)$$

As observed outcome is perfectly predicted by D and X in the sample, the predicted values are equal to observed outcomes and the residuals are all zero. Further, as suggested above, the relation between outcome, Y, and the regressor, X, does not differ in the two treatment clusters; hence, the coefficient on the interaction term is zero. An interpretation of the regression is, on average, outcome differs between the two treatment clusters by 5 (the coefficient on D) with a baseline when $D = 0$ of 5 (the intercept), and within a cluster, outcome responds one-to-one (the coefficient on X is 1) with X. For instance, when $D = 0$ and the covariate is low, $X = -1$,

$$E[Y \mid D = 0, X = -1] = 5.0 + 5.0 (0) + 1.0 (-1) = 4$$

On the other hand, when $D = 1$ and the covariate is high, $X = 1$,

$$E[Y \mid D = 1, X = 1] = 5.0 + 5.0 (1) + 1.0 (1) = 11$$

and so on.

The omitted, correlated variable bias in the simple regression compared to ANCOVA is

$$\begin{aligned} bias(d_1) &= \left(X_1^T X_1\right)^{-1} X_1^T x_2 \delta_2 \\ &= \frac{1}{12} \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \end{bmatrix} 5 \\ &= \begin{bmatrix} 2.5 \\ 0 \end{bmatrix} \end{aligned}$$

where $X_1 = \begin{bmatrix} \iota & X \end{bmatrix}$ and $x_2 = D$. Omission of D causes no bias in the coefficient on X as D and X are uncorrelated; nonetheless, the intercept is biased.

Contents

3
Classical causal effects strategies

3.1 Causal effects and treatment effects

When evaluating accounting choices, we're deeply interested in their welfare effects. Does choice A make everyone better off or worse off compared with choice B? Or, does one choice make some better off and the other choice make others better off such that self-selection is a Pareto improvement? These are difficult questions and their resolution is invariably controversial. The root of the inference or modeling problem can be traced back to omitted, correlated regressor variables, as discussed in the above simpler settings.

Causal effects may involve choices with which we have experience in familiar environments or in new environments. Or, we may be interested in welfare effects associated with choices with which we have no experience in familiar or new environments. In the former case, where we have history on our side, we might pose treatment effect questions and employ historical data to help make an assessment. Treatment effects ask whether an individual's welfare is greater with treatment than without treatment. That is, other things are held constant and we attempt to explore the impact of treatment on welfare.

Treatment effects are less demanding than causal effects of unexplored choices in new environments. Nonetheless, treatment effect analysis poses serious challenges. The endogenous nature of choice often makes it difficult to hold other things constant. Observable outcome typically is an incomplete and ill-timed measure of welfare. While we're interested in the

individual's expected utility, we usually observe ex post outcomes. Ex post versus ex ante considerations may or may not be easily surmounted. Outcomes may represent gross gains rather than ex ante differences in utility. Gross gains may be related to net benefits if costs are well understood but individual specific features (for example, nonpecuniary considerations) may be particularly elusive. One of the most severe challenges is we typically observe data for an individual only with treatment or without treatment but not for both. This implies that we cannot directly assess an individual's treatment effect. However, homogeneity conditions may allow inference based on population-level treatment effect parameters (for example, mean or average treatment effects).

3.2 A simple treatment effect example

The above *ANCOVA* example illustrates a simple treatment effect analysis if, for instance, counterfactuals have the same probability distribution as those observed. Counterfactuals are conditions not observed. To fix ideas, let Y_1 denote outcome with treatment and Y_0 outcome without treatment. Then the treatment effect is $Y_1 - Y_0$. However, we observe $(Y_1 \mid D = 1)$ and $(Y_0 \mid D = 0)$ but don't observe the counterfactuals, $(Y_1 \mid D = 0)$ and $(Y_0 \mid D = 1)$. We would like to compare outcome with treatment to outcome without treatment for individuals who chose treatment (treatment effect on the treated — TT) and for individuals who chose no treatment (treatment effect on the untreated — TUT). Both treatment effects compare factual with counterfactual outcomes

$$\begin{aligned} TT &= (Y_1 \mid D = 1) - (Y_0 \mid D = 1) \\ &= (Y_1 - Y_0 \mid D = 1) \end{aligned}$$

and

$$\begin{aligned} TUT &= (Y_1 \mid D = 0) - (Y_0 \mid D = 0) \\ &= (Y_1 - Y_0 \mid D = 0) \end{aligned}$$

Suppose we have the following DGP (factual and counterfactual)

$$\begin{aligned} E\left[Y_1 \mid D = 1, X = -1\right] &= E\left[Y_1 \mid D = 0, X = -1\right] \\ &= E\left[Y_1 \mid X = -1\right] = 9 \end{aligned}$$

$$\begin{aligned} E\left[Y_1 \mid D = 1, X = 0\right] &= E\left[Y_1 \mid D = 0, X = 0\right] \\ &= E\left[Y_1 \mid X = 0\right] = 10 \end{aligned}$$

$$\begin{aligned} E\left[Y_1 \mid D = 1, X = 1\right] &= E\left[Y_1 \mid D = 0, X = 1\right] \\ &= E\left[Y_1 \mid X = 1\right] = 11 \end{aligned}$$

$$\begin{aligned} E[Y_0 \mid D=1, X=-1] &= E[Y_0 \mid D=0, X=-1] \\ &= E[Y_0 \mid X=-1] = 4 \end{aligned}$$

$$\begin{aligned} E[Y_0 \mid D=1, X=0] &= E[Y_0 \mid D=0, X=0] \\ &= E[Y_0 \mid X=0] = 5 \end{aligned}$$

$$\begin{aligned} E[Y_0 \mid D=1, X=1] &= E[Y_0 \mid D=0, X=1] \\ &= E[Y_0 \mid X=1] = 6 \end{aligned}$$

Treatment is said to be ignorable or selection is on observables as the regressors are sufficiently informative to make treatment, D, conditionally uninformative of outcome with treatment, Y_1, and outcome without treatment, Y_0. Then, the conditional average treatment effects on the treated ($ATT(X)$) and on the untreated ($ATUT(X)$) are

$$\begin{aligned} ATT(X=-1) &= E[Y_1 - Y_0 \mid D=1, X=-1] \\ &= 9-4=5 \end{aligned}$$

$$\begin{aligned} ATT(X=0) &= E[Y_1 - Y_0 \mid D=1, X=0] \\ &= 10-5=5 \end{aligned}$$

$$\begin{aligned} ATT(X=1) &= E[Y_1 - Y_0 \mid D=1, X=1] \\ &= 11-6=5 \end{aligned}$$

$$\begin{aligned} ATUT(X=-1) &= E[Y_1 - Y_0 \mid D=0, X=-1] \\ &= 9-4=5 \end{aligned}$$

$$\begin{aligned} ATUT(X=0) &= E[Y_1 - Y_0 \mid D=0, X=0] \\ &= 10-5=5 \end{aligned}$$

$$\begin{aligned} ATUT(X=1) &= E[Y_1 - Y_0 \mid D=0, X=1] \\ &= 11-6=5 \end{aligned}$$

If outcome represents net benefit, then, conditional on X, everyone is better off with treatment than without treatment. Since this is true for all levels of X, it is not surprising that, on applying iterated expectations, the unconditional average treatment effects on the treated (ATT) and on the

untreated ($ATUT$) indicate an average (over all common X) net benefit as well.[1]

$$\begin{aligned} ATT &= E_X\left[E\left[Y_1 - Y_0 \mid D = 1, X\right]\right] \\ &= E\left[Y_1 - Y_0 \mid D = 1\right] = 5 \end{aligned}$$

and

$$\begin{aligned} ATUT &= E_X\left[E\left[Y_1 - Y_0 \mid D = 0, X\right]\right] \\ &= E\left[Y_1 - Y_0 \mid D = 0\right] = 5 \end{aligned}$$

Of course, this degree of homogeneity implies the average treatment effect is

$$\begin{aligned} ATE &= \Pr\left(D = 1\right) ATT + \left(1 - \Pr\left(D = 1\right)\right) ATUT \\ &= \Pr\left(D = 1\right) E\left[Y_1 - Y_0 \mid D = 1\right] + \Pr\left(D = 0\right) E\left[Y_1 - Y_0 \mid D = 0\right] \\ &= E\left[Y_1 - Y_0\right] = 5 \end{aligned}$$

3.3 Treatment effects with limited common support

Unfortunately, the above *DGP*, where outcome reflects welfare, outcome is homogeneous, and common X support, is rarely encountered. Rather, it's typical to encounter some heterogeneity in outcome and limited common support.[2] To illustrate the implications of limited common support, suppose we have the following data (where relative population frequencies are reflected by their sample frequencies).

Y	Y_1	Y_0	D	X	X_1	X_0
4	13	4	0	0	−2	0
6	11	6	0	−1	−1	−1
5	11	5	0	−1	0	−1
4	12	4	0	0	−1	0
11	11	2	1	0	0	2
11	11	4	1	0	0	1
9	9	3	1	1	1	1
10	10	4	1	1	1	0

[1] Common support for X is important as our inferences stem from evidence we have rather than evidence we don't have in hand.

[2] Further, often outcome measures gross benefits (and perhaps incompletely) rather than net benefits so that welfare implications require knowledge of costs with and without treatment.

We don't observe the counterfactuals: $(Y_1, X_1 \mid D = 0)$ or $(Y_0, X_0 \mid D = 1)$, but the key to identifying any average treatment effect is

$$E[Y_1 \mid X_1 = x, D = 1] = E[Y_1 \mid X_1 = x, D = 0]$$

and

$$E[Y_0 \mid X_0 = x, D = 1] = E[Y_0 \mid X_0 = x, D = 0]$$

Therefore, the pivotal condition is outcome mean conditional independence of treatment, D. For the only commonly observed value, $x = 0$

$$E[Y_1 \mid X_1 = 0, D = 1] = E[Y_1 \mid X_1 = 0, D = 0] = 11$$

and

$$E[Y_0 \mid X_0 = 0, D = 1] = E[Y_0 \mid X_0 = 0, D = 0] = 4$$

conditional mean independence is satisfied. Hence, the only evidence-based assessment of the treatment effect is for $X_1 = X_0 = X = 0$, and

$$ATE(X = 0) = E[Y_1 - Y_0 \mid X = 0] = 11 - 4 = 7$$

Further, this conditional average treatment effect is homogeneous.

$$ATT(X = 0) = ATUT(X = 0) = ATE(X = 0) = 7$$

where

$$ATT(X = 0) = E[Y_1 - Y_0 \mid X = 0, D = 1] = 11 - 4 = 7$$

and

$$ATUT(X = 0) = E[Y_1 - Y_0 \mid X = 0, D = 0] = 11 - 4 = 7$$

While this conditional average treatment effect is, in principle, only non-parametrically identified, by good fortune, *ANCOVA* effectively estimates both the conditional (on $X = 0$) and unconditional average treatment effect via the coefficient on D.[3]

$$E[Y \mid D, X] = 4 + 7D - 1.5X$$

where the observables are

$$Y = DY_1 + (1 - D)Y_0$$

and

$$X = DX_1 + (1 - D)X_0$$

[3] More generally, we include an interaction term, $\left(D \times \left(X - \overline{X}\right)\right)$, but it's coefficient is zero for this *DGP*.

Further, the conditional average treatment effect also equals the unconditional average. Since the unconditional average treatment effect is unidentified by observable data, both of these results are merely fortuitous. That is, the only conclusion we can draw based on the evidence is for the average treatment effect conditional on $X = 0$. If there is a local interval of common X support, this is sometimes called a local average treatment effect.

To clarify this common support issue, suppose we perturb only the counterfactual outcomes with treatment as follows.

Y	Y_1	Y_0	D	X	X_1	X_0
4	3	4	0	0	−2	0
6	2	6	0	−1	−1	−1
5	11	5	0	−1	0	−1
4	1	4	0	0	−1	0
11	11	2	1	0	0	2
11	11	4	1	0	0	1
9	9	3	1	1	1	1
10	10	4	1	1	1	0

Now, the unconditional average treatment effect is $7.25 - 4 = 3.25$, the unconditional average treatment effect on the treated is unperturbed from above, $10.25 - 3.25 = 7$, and the unconditional average treatment effect on the untreated is $4.25 - 4.75 = -0.5$. Hence, outcome is heterogeneous, outcome supports self-selection,[4] and none of these unconditional average treatment effects are identified by the data. As above, the only treatment effect identifiable from the data is the conditional average treatment effect for $X = 0$, which continues to be $ATE\,(X = 0) = 11 - 4 = 7$. Attempting to extrapolate from the evidence to unconditional average treatment effects is not only a stab in the dark, it is misleading.

3.4 Local average treatment effects

The above example suggests the conditions for ignorable treatment may severely limit identification and estimation of treatment effects. A common complementary approach to expanding the set of regressors (ignorable treatment) is to employ instrumental variables. *Instrumental variables*, Z, are variables that are associated with treatment choice, D, but unrelated to the outcomes with and without treatment, Y_1 and Y_0. The idea is we

[4] There is evidence that an individual self-selects when their choice produces a better outcome than do the alternative choices. That is, those individuals choosing treatment are better off with treatment than without but those choosing no treatment are better off without treatment than with. As outcome may be an incomplete indicator of expected utility, expected utility maximizing behavior does not always produce evidence of self-selection.

can manipulate treatment choice with the instrument but leave outcomes unaffected. This permits extrapolation from observables to counterfactuals, $E[Y_1 \mid D = 0]$ and $E[Y_0 \mid D = 1]$.

If outcomes with treatment, Y_1, and outcomes without treatment, Y_0, are independent of a binary instrument, Z, then the discrete marginal treatment effect or local average treatment effect,

$$LATE = E[Y_1 - Y_0 \mid D_1 - D_0 = 1]$$

where $D_1 = (D \mid Z = 1)$ and $D_0 = (D \mid Z = 0)$ equals

$$\frac{E[Y \mid Z = 1] - E[Y \mid Z = 0]}{E[D \mid Z = 1] - E[D \mid Z = 0]}$$

This quantity (ratio) can be estimated from observables, therefore *LATE* is identified. In fact, this quantity (*estimand*) is estimated by standard two-stage instrumental variable estimation (*2SLS-IV*).

3.4.1 2SLS-IV estimation

Suppose we envision the regression in error form

$$Y = \alpha + \beta D + \varepsilon$$

but $E[\varepsilon \mid D] \neq 0$, then *OLS* provides inconsistent parameter estimates but instrumental variable estimation can rectify the problem. As the name suggests, *2SLS-IV* estimation involves two stages of projections. The first stage puts the explanatory variables of interest (here, treatment, D) in the columns of the instruments, Z. In other words, we construct[5]

$$\begin{aligned} \widehat{D} &= Z\left(Z^T Z\right)^{-1} Z^T \widetilde{D} \\ &= P_Z \widetilde{D} \end{aligned}$$

where $\widetilde{D} = D - \overline{D}$ is the estimated mean deviation. Then, we estimate

$$E[Y \mid D] = a + b\widehat{D}$$

where the estimate of β is

$$b = \left(X^T X\right)^{-1} X^T \widetilde{Y}$$

[5] To simplify matters, we work with a single variable by utilizing mean deviations of all variables. We discuss *2SLS-IV* estimation more generally in the appendix to this chapter.

and $X = \widehat{D}$ and $\widetilde{Y} = Y - \overline{Y}$. Since

$$\begin{aligned}
\left(X^T X\right)^{-1} X^T \widetilde{Y} &= \left(\widetilde{D}^T P_Z P_Z \widetilde{D}\right)^{-1} \widetilde{D}^T P_Z P_Z \widetilde{Y} \\
&= \left(P_Z \widetilde{D}\right)^{-1} \left(\widetilde{D}^T P_Z\right)^{-1} \widetilde{D}^T P_Z P_Z \widetilde{Y} \\
&= \left(P_Z \widetilde{D}\right)^{-1} P_Z \widetilde{Y} \\
&= \frac{\frac{1}{n} Z^T \widetilde{Y}}{\frac{1}{n} Z^T \widetilde{D}}
\end{aligned}$$

$\frac{\frac{1}{n} Z^T \widetilde{Y}}{\frac{1}{n} Z^T \widetilde{D}}$ estimates $\frac{E\left[\widetilde{Y}|Z=1\right] - E\left[\widetilde{Y}|Z=0\right]}{E\left[\widetilde{D}|Z=1\right] - E\left[\widetilde{D}|Z=0\right]} = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]} = LATE$. It's time for an example.

3.4.2 IV example 1

Suppose the *DGP* is

Y	D	Y_1	Y_0	Z
15	1	15	10	1
15	1	15	10	0
10	1	10	10	1
10	0	10	10	0
10	0	5	10	1
10	0	5	10	0

IV example 1: $LATE = 0$

If we estimate by *OLS* we find

$$E\left[Y \mid D\right] = 10 + 3\frac{1}{3}D$$

suggesting the average treatment effect is $3\frac{1}{3}$. As treatment is not ignorable, this is a false conclusion,

$$ATE = E\left[Y_1 - Y_0\right] = 10 - 10 = 0$$

Now, if we think of the first two rows as state 1 and successive pairs of rows similarly where treatment, D, is potentially manipulated via the instrument, Z, then we can estimate *LATE* via *2SLS-IV*. With this *DGP*, *LATE* is identified for only state 2 (rows 3 and 4) since $D_1 - D_0 = 1$ (the compliers — individuals induced to select treatment when $Z = 1$ but not when $Z = 0$). State 1 represents individuals who always select treatment and state 3 represents individuals who never select treatment. Clearly, $LATE = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]} = \frac{10-10}{1-0} = 0$ and *2SLS-IV* estimates $\frac{\frac{1}{n} Z^T \widetilde{Y}}{\frac{1}{n} Z^T \widetilde{D}} = 0$, in large samples. Hence, for this *DGP* and instrument,

$LATE = ATE$. but we should not expect this, in general, as the next examples illustrate.

3.4.3 IV example 2

Suppose the *DGP* is

Y	D	Y_1	Y_0	Z
15	1	15	10	1
10	0	15	10	0
10	1	10	10	1
10	1	10	10	0
10	0	5	10	1
10	0	5	10	0

IV example 2: $LATE = 5$

OLS estimates

$$E[Y \mid D] = 10 + 1\frac{2}{3}D$$

which again fails to identify the average treatment effect, $ATE = 0$. Now, the compliers are reflected by state 1 alone, and $LATE = 5$ while ATE continues to be zero. Also, *2SLS-IV* estimates $\frac{\frac{1}{n}Z^T\widetilde{Y}}{\frac{1}{n}Z^T\widetilde{D}} = 5$ in large samples.

3.4.4 IV example 3

The *DGP* along with the instrument identifies the particular marginal treatment effect. Consider another variation

Y	D	Y_1	Y_0	Z
15	1	15	10	1
15	1	15	10	0
10	0	10	10	1
10	0	10	10	0
5	1	5	10	1
10	0	5	10	0

IV example 3: $LATE = -5$

OLS again supplies an inconsistent estimate of *ATE*.

$$E[Y \mid D] = 10 + 1\frac{2}{3}D$$

As the compliers are individuals in state 3, $LATE = -5$ and *2SLS-IV* estimates $\frac{\frac{1}{n}Z^T\widetilde{Y}}{\frac{1}{n}Z^T\widetilde{D}} = -5$ in large samples.

3.4.5 IV example 4

Sometimes $LATE$ equals the average treatment effect on the treated. If no one adopts treatment when the instrument value equals zero, then $LATE = ATT$. Consider the DGP

Y	D	Y_1	Y_0	Z
15	1	15	10	1
10	0	15	10	0
20	0	20	20	1
20	0	20	20	0
10	1	10	10	1
10	0	10	10	0

IV example 4: $LATE = ATT$

OLS estimates

$$E[Y \mid D] = 15 - 2.5D$$

but $ATE = 1\frac{2}{3}$ (opposite directions, or a Simpson's paradox result) and $ATUT = 1.25$. $LATE = 2.5$ is defined by states 1 and 3 and since $\Pr(D = 1 \mid Z = 0) = 0$, $LATE = ATT = E[Y_1 - Y_0 \mid D = 1] = 2.5$. And, *2SLS-IV* estimates $\frac{\frac{1}{n}Z^T\widetilde{Y}}{\frac{1}{n}Z^T\widetilde{D}} = 2.5$ in large samples.

3.4.6 IV example 5

$LATE$ equals the average treatment effect on the untreated if everyone adopts treatment when the instrument equals unity. Consider the DGP

Y	D	Y_1	Y_0	Z
15	1	15	10	1
10	0	15	10	0
20	1	20	10	1
20	1	20	10	0
10	1	10	10	1
10	0	10	10	0

IV example 5: $LATE = ATUT$

OLS estimates

$$E[Y \mid D] = 10 + 6.25D$$

but $ATE = 5$ and $ATT = 6.25$. $LATE = 2.5$ is defined by states 1 and 3 and since $\Pr(D = 1 \mid Z = 1) = 1$, $LATE = ATUT = E[Y_1 - Y_0 \mid D = 1] = 2.5$. And, *2SLS-IV* estimates $\frac{\frac{1}{n}Z^T\widetilde{Y}}{\frac{1}{n}Z^T\widetilde{D}} = 2.5$ in large samples.

3.4.7 IV example 6

Unfortunately, if some individuals are induced to accept treatment when the instrument changes to one but others are induced to move away from treatment with the same instrumental variable manipulation, then extant instrumental variable strategies break down. Uniformity is a condition for *IV* identification of *LATE*, any defiers result in treatment effect identification failure. We illustrate the problem once again with a simple binary instrument. Consider the *DGP*

Y	D	Y_1	Y_0	Z
15	1	15	10	1
10	0	15	10	0
20	1	20	15	1
20	1	20	15	0
20	1	20	15	1
15	0	20	15	0
15	0	10	15	1
10	1	10	15	0

IV example 6: defiers

OLS estimates

$$E[Y \mid D] = 13\frac{1}{3} + 3\frac{2}{3}D$$

but $ATE = 2.5$, $ATT = 3$, and $ATUT = 1\frac{2}{3}$. $LATE = 5$ is defined by states 1 and 3 but state 4 violates uniformity. *2SLS-IV* estimates $\frac{\frac{1}{n}Z^T\tilde{Y}}{\frac{1}{n}Z^T\tilde{D}} = 15$ in large samples, a gross overstatement of the treatment effect.

This illustrates the trouble two-way flows cause in the identification of treatment effects. Extant *IV* strategies rely on uniformity either toward treatment or away from treatment by all individuals, not some individuals toward and others away from treatment in response to changes in the instrument. Further, this simple binary instrumental variable strategy identifies the local average treatment effect for an unidentified subpopulation of compliers. Nonetheless, binary *IV* identifies marginal treatment effects for this subpopulation, a parameter surely of some interest.

3.5 Treatment effects and control functions

Another approach that may be effective for identifying treatment effects utilizes control functions. That is, functions which directly control the source of selection bias. Consider a simple data generating process (to keep the discussion compact, there are no regressors).

$$Y_j = \mu_j + V_j, \quad j = 0, 1$$

where μ_j is the mean of outcome and V_j is the unobserved (not residual) portion of outcome for treatment j.

Y	D	Y_1	Y_0	V_1	V_0
15	1	15	9	3	-1
14	1	14	10	2	-2
13	1	13	11	1	-3
13	0	11	13	-1	3
14	0	10	14	-2	2
15	0	9	15	-3	1

If we attempt to estimate average treatment effects via an exogenous dummy variable regression[6]

$$E[Y \mid D] = \mu_0 + (\mu_1 - \mu_0) D$$

we find that OLS estimates

$$E[Y \mid D] = 14 + 0D$$

Suggesting all average treatment effects are zero. While it is the case, the unconditional average treatment effect is zero

$$E[Y_1 - Y_0] = 12 - 12 = 0$$

the means for outcome with treatment and with no treatment are not identified as OLS suggests the mean of each is 14 while the DGP clearly indicates the mean of each is 12. Further, we may have more interest in the average treatment effect on the treated and untreated but OLS does not identify either of these quantities. The fundamental problem is that the basic condition for a well-posed regression, $E[V_j \mid X] = 0$, is not satisfied. Rather,

$$E[V_1 \mid D = 1] = \frac{1}{3}(3 + 2 + 1) = 2$$

$$E[V_1 \mid D = 0] = \frac{1}{3}(-3 - 2 - 1) = -2$$

$$E[V_0 \mid D = 1] = \frac{1}{3}(-3 - 2 - 1) = -2$$

and

$$E[V_0 \mid D = 0] = \frac{1}{3}(3 + 2 + 1) = 2$$

Nonetheless, the average treatment effects on the treated (ATT) and untreated ($ATUT$) are well-defined.

$$\begin{aligned} ATT &= E[Y_1 \mid D = 1] - E[Y_0 \mid D = 1] \\ &= 12 + 2 - (12 - 2) \\ &= 4 \end{aligned}$$

[6] This is in the same spirit as a single factor $ANOVA$ with binary factor levels.

and

$$\begin{aligned} ATUT &= E[Y_1 \mid D=0] - E[Y_0 \mid D=0] \\ &= (12-2) - (12+2) \\ &= -4 \end{aligned}$$

Also, these quantities readily connect to the average treatment effect.

$$\begin{aligned} ATE &= \Pr(D=1)\, ATT + (1 - \Pr(D=1))\, ATUT \\ &= \frac{1}{2}(4) + \frac{1}{2}(-4) = 0 \end{aligned}$$

The key is to determine a path from observable data to these quantities. The control function approach attempts to include functions in the regression that control for the source of selection bias, $E[V_j \mid D]$. Then, the means can be properly identified and estimation from observable data is feasible.

The most popular control function approach was developed by Nobel laureate, James Heckman. Briefly, the idea is treatment selection by an individual reflects expected utility maximizing behavior. The data analyst (manager, social scientist, etc.) observes some factors influencing this choice but other factors are unobserved (by the analyst). These unobserved components lead to a stochastic process description of individual choice behavior. The key to this stochastic description is the probability assignment to the unobservable component. Heckman argues when the probability assignment is Gaussian or normal, then we can treat the problem as a truncated regression exercise. And, when common support conditions for the regressors are satisfied, in principle, average treatment effects on the treated, untreated, and the unconditional average are identified. Otherwise, when common support conditions are limited, local average treatment effects only are identified. We sketch the ideas below and relate them to the above example.[7]

3.5.1 Inverse Mills control function strategy

Consider the *DGP* where choice is represented by a latent variable characterizing the difference in expected utility associated with treatment or no treatment, observed choice, and outcome equations with treatment and without treatment.

latent choice equation:

$$EU = W\theta + V_D$$

observed choice:

$$D = \begin{array}{ll} 1 & \text{if } EU > 0 \quad V_D > -W\theta \\ 0 & \text{otherwise} \end{array}$$

[7] This subsection is heavily laden with notation — bear with us.

outcome equations:

$$\begin{aligned} Y_1 &= \mu_1 + X\beta_1 + V_1 \\ Y_0 &= \mu_0 + X\beta_0 + V_0 \end{aligned}$$

Heckman's two-stage estimation procedure is as follows. First, estimate θ via a probit regression of D on $W = \{\iota, X, Z\}$ and identify observations with common support (that is, observations for which the regressors, X, for the treated overlap with regressors for the untreated). Second, regress Y onto

$$\left\{\iota, D, X, D\left(X - E\left[X\right]\right), D\left(\frac{\phi}{\Phi}\right), (1-D)\frac{-\phi}{1-\Phi}\right\}$$

for the overlapping subsample. With full support, the coefficient on D is a consistent estimator of ATE; with less than full common support, we have a local average treatment effect.[8]

Wooldridge suggests identification of

$$ATE = \mu_1 - \mu_0 + E\left[X\right](\beta_1 - \beta_0)$$

via α in the regression

$$\begin{aligned} E\left[Y \mid X, Z\right] &= \mu_0 + \alpha D + X\beta_0 + D\left(X - E\left[X\right]\right)(\beta_1 - \beta_0) \\ &\quad + D\rho_{1V_D}\sigma_1\frac{\phi\left(W\theta\right)}{\Phi\left(W\theta\right)} - (1-D)\rho_{0V_D}\sigma_0\frac{\phi\left(W\theta\right)}{1-\Phi\left(W\theta\right)} \end{aligned}$$

This follows from the observable response

$$\begin{aligned} Y &= D\left(Y_1 \mid D=1\right) + (1-D)\left(Y_0 \mid D=0\right) \\ &= \left(Y_0 \mid D=0\right) + D\left[\left(Y_1 \mid D=1\right) - \left(Y_0 \mid D=0\right)\right] \end{aligned}$$

and applying conditional expectations

$$\begin{aligned} E\left[Y_1 \mid X, D=1\right] &= \mu_1 + X\beta_1 + \rho_{1V_D}\sigma_1\frac{\phi\left(W\theta\right)}{\Phi\left(W\theta\right)} \\ E\left[Y_0 \mid X, D=0\right] &= \mu_0 + X\beta_0 - \rho_{0V_D}\sigma_0\frac{\phi\left(W\theta\right)}{1-\Phi\left(W\theta\right)} \end{aligned}$$

[8] We should point out here that second stage OLS does not provide valid estimates of standard errors. As Heckman points out there are two additional concerns: the errors are heteroskedastic (so an adjustment such as White suggested is needed) and θ has to be estimated (so we must account for this added variation). Heckman identifies a valid variance estimator for this two-stage procedure.

Simplification produces Wooldridge's result.

$$\begin{aligned} E[Y \mid X, Z] &= E[(Y_0 \mid D=0) + D\{(Y_1 \mid D=1) - (Y_0 \mid D=0)\} \mid X, Z] \\ &= \mu_0 + X\beta_0 - \rho_{0V_D}\sigma_0 \frac{\phi(W\theta)}{1-\Phi(W\theta)} \\ &\quad + D\left(\mu_1 + X\beta_1 + \rho_{1V_D}\sigma_1 \frac{\phi(W\theta)}{\Phi(W\theta)}\right) \\ &\quad - D\left(\mu_0 + X\beta_0 - \rho_{0V_D}\sigma_0 \frac{\phi(W\theta)}{1-\Phi(W\theta)}\right) \end{aligned}$$

now rearrange terms

$$\begin{aligned} &\mu_0 + D\{\mu_1 - \mu_0 + E[X](\beta_1 - \beta_0)\} + X\beta_0 + D(X - E[X])(\beta_1 - \beta_0) \\ &+ D\rho_{1V_D}\sigma_1 \frac{\phi(W\theta)}{\Phi(W\theta)} - (1-D)\rho_{0V_D}\sigma_0 \frac{\phi(W\theta)}{1-\Phi(W\theta)} \end{aligned}$$

The coefficient on D, $\{\mu_1 - \mu_0 + E[X](\beta_1 - \beta_0)\}$, is ATE.

The key ideas behind treatment effect identification via control functions can be illustrated by reference to this case.

$$E[Y_j \mid X, D=j] = \mu_j + X\beta_j + E[V_j \mid D=j]$$

Given the conditions, $E[V_j \mid D=j] \neq 0$ unless $Corr(V_j, V_D) = \rho_{jV_D} = 0$. For $\rho_{jV_D} \neq 0$,

$$E[V_1 \mid D=1] = \rho_{1V_D}\sigma_1 E[V_D \mid V_D > -W\theta]$$

$$E[V_0 \mid D=1] = \rho_{0V_D}\sigma_0 E[V_D \mid V_D > -W\theta]$$

$$E[V_1 \mid D=0] = \rho_{1V_D}\sigma_1 E[V_D \mid V_D \leq -W\theta]$$

and

$$E[V_0 \mid D=0] = \rho_{0V_D}\sigma_0 E[V_D \mid V_D \leq -W\theta]$$

The final term in each expression is the expected value of a truncated standard normal random variate where

$$h_1 \equiv E[V_D \mid V_D > -W\theta] = \frac{\phi(-W\theta)}{1-\Phi(-W\theta)} = \frac{\phi(W\theta)}{\Phi(W\theta)}$$

and

$$h_0 \equiv E[V_D \mid V_D \leq -Z\theta] = -\frac{\phi(-W\theta)}{\Phi(-W\theta)} = -\frac{\phi(W\theta)}{1-\Phi(W\theta)}$$

Putting this together, we have

$$E[Y_1 \mid X, D=1] = \mu_1 + X\beta_1 + \rho_{1V_D}\sigma_1 \frac{\phi(W\theta)}{\Phi(W\theta)}$$

$$E\left[Y_0 \mid X, D=0\right] = \mu_0 + X\beta_0 - \rho_{0V_D}\sigma_0 \frac{\phi\left(W\theta\right)}{1-\Phi\left(W\theta\right)}$$

and counterfactuals

$$E\left[Y_0 \mid X, D=1\right] = \mu_0 + X\beta_0 + \rho_{0V_D}\sigma_0 \frac{\phi\left(W\theta\right)}{\Phi\left(W\theta\right)}$$

and

$$E\left[Y_1 \mid X, D=0\right] = \mu_1 + X\beta_1 - \rho_{1V_D}\sigma_1 \frac{\phi\left(W\theta\right)}{1-\Phi\left(W\theta\right)}$$

The appeal of Heckman's inverse Mills ratio strategy can be seen in its estimation simplicity and the ease with which treatment effects are then identified. Of course, this doesn't justify the identification conditions — only our understanding of the data can do that. The conditional average treatment effect on the treated is

$$ATT\left(X,Z\right) = \mu_1 - \mu_0 + X\left(\beta_1 - \beta_0\right) + \left(\rho_{1V_D}\sigma_1 - \rho_{0V_D}\sigma_0\right)\frac{\phi\left(W\theta\right)}{\Phi\left(W\theta\right)}$$

and by iterated expectations (with full support), we have the unconditional average treatment effect on the treated

$$ATT = \mu_1 - \mu_0 + E\left[X\right]\left(\beta_1 - \beta_0\right) + \left(\rho_{1V_D}\sigma_1 - \rho_{0V_D}\sigma_0\right)E\left[\frac{\phi\left(W\theta\right)}{\Phi\left(W\theta\right)}\right]$$

Also, the conditional average treatment effect on the untreated is

$$ATUT\left(X,Z\right) = \mu_1 - \mu_0 + X\left(\beta_1 - \beta_0\right) - \left(\rho_{1V_D}\sigma_1 - \rho_{0V_D}\sigma_0\right)\frac{\phi\left(W\theta\right)}{1-\Phi\left(W\theta\right)}$$

and by iterated expectations, we have the unconditional average treatment effect on the untreated

$$ATUT = \mu_1 - \mu_0 + E\left[X\right]\left(\beta_1 - \beta_0\right) - \left(\rho_{1V_D}\sigma_1 - \rho_{0V_D}\sigma_0\right)E\left[\frac{\phi\left(W\theta\right)}{1-\Phi\left(W\theta\right)}\right]$$

Since

$$\begin{aligned} ATE\left(X,Z\right) &= \Pr\left(D=1 \mid X,Z\right)ATT\left(X,Z\right) \\ &\quad + \Pr\left(D=0 \mid X,Z\right)ATUT\left(X,Z\right) \\ &= \Phi\left(W\theta\right)ATT\left(X,Z\right) + \left(1-\Phi\left(W\theta\right)\right)ATUT\left(X,Z\right) \end{aligned}$$

we have the conditional average treatment effect is

$$\begin{aligned} ATE\left(X,Z\right) &= \mu_1 - \mu_0 + X\left(\beta_1 - \beta_0\right) \\ &\quad + \left(\rho_{1V}\sigma_1 - \rho_{0V_D}\sigma_0\right)\phi\left(W\theta\right) - \left(\rho_{1V}\sigma_1 - \rho_{0V_D}\sigma_0\right)\phi\left(W\theta\right) \\ &= \mu_1 - \mu_0 + X\left(\beta_1 - \beta_0\right) \end{aligned}$$

and by iterated expectations, we have the unconditional average treatment effect is

$$ATE = \mu_1 - \mu_0 + E\left[X\right]\left(\beta_1 - \beta_0\right)$$

3.5.2 Back to the example

Now, we return to the example and illustrate this control function strategy. Suppose the first stage probit regression produces the following hazard rates (inverse Mills ratios) where $h = D*h_1+(1-D)\,h_0$, $h_1 = \rho_{D,1}\sigma_1\frac{\phi(W\theta)}{\Phi(W\theta)}$ $h_0 = -\rho_{D,0}\sigma_0\frac{\phi(W\theta)}{1-\Phi(W\theta)}$, and the standard deviations are $\sigma_1 = \sigma_0 = 2.16$.[9],[10]

Y	D	Y_1	Y_0	V_1	V_0	h
15	1	15	9	3	−3	3
14	1	14	10	2	−2	2
13	1	13	11	1	−1	1
13	0	11	13	−1	1	1
14	0	10	14	−2	2	2
15	0	9	15	−3	3	3

The large sample second stage regression is

$$E\left[Y \mid D, h\right] = 12 + 0D + 1.0\left(D \times h_1\right) - 1.0\left(\left(1-D\right) \times h_0\right)$$

Estimated average treatment effects consistently identify (again, a large sample result) the average treatment effects as follows. The average treatment effect is estimated via the coefficient on D

$$estATE = 0$$

---

[9] In other words, the sample is representative of the population. Hence,

$$\begin{aligned}\sigma_i &= \sqrt{\frac{1}{6}\left(3^2+2^2+1^2+(-3)^2+(-2)^2+(-1)^2\right)} \\ &= \sqrt{\frac{28}{6}} \approx 2.16\end{aligned}$$

[10] Clearly, we've omitted details associated with the first stage. Suffice to say we have regressors (instruments) related to selection, D, but that are uninformative about outcomes, Y_1 and Y_0 (otherwise we would include them in the output regressions). The instruments, $Z = \begin{bmatrix} Z_1 & Z_2 & Z_3 & Z_4 \end{bmatrix}$ (no intercept; tabulated below) employed are orthogonal to Y_1 and Y_0.

Z_1	Z_2	Z_3	Z_4
5	4	3	1
−6	−5	−4	−2
0	0	0	1
0	0	1	0
0	1	0	0
1	0	0	0

In fact, they form a basis for the nullspace to $\begin{bmatrix} Y_1 & Y_0 \end{bmatrix}$. When we return to this setting to explore Bayesian analysis, we'll be more explicit about this first stage estimation but we bypass this stage for now.

Other estimated averages of interest are

$$\begin{aligned} estE\left[Y_1 \mid D=1\right] &= 12+0+1.0\left(\frac{3+2+1}{3}\right) \\ &= 14 \\ estE\left[Y_1 \mid D=0\right] &= 12+1.0\left(\frac{-3-2-1}{3}\right) \\ &= 10 \\ estE\left[Y_0 \mid D=1\right] &= 12+0-1.0\left(\frac{3+2+1}{3}\right) \\ &= 10 \\ estE\left[Y_0 \mid D=0\right] &= 12-1.0\left(\frac{-3-2-1}{3}\right) \\ &= 14 \end{aligned}$$

Hence, the estimated average treatment effect on the treated is

$$\begin{aligned} estATT &= estE\left[Y_1 \mid D=1\right]-estE\left[Y_0 \mid D=1\right] \\ &= 14-10=4 \end{aligned}$$

and the estimated average treatment effect on the untreated is

$$\begin{aligned} estATUT &= estE\left[Y_1 \mid D=0\right]-estE\left[Y_0 \mid D=0\right] \\ &= 10-14=-4 \end{aligned}$$

We see the control function strategy has effectively addressed selection bias and allowed us to identify some average treatment effects of interest even though the DGP poses serious challenges.

3.6 Pursuit of higher explanatory power

A word of caution. Frequently, we utilize explanatory power to help gauge model adequacy. This is a poor strategy in the analysis of treatment effects. Higher explanatory power in either the selection equation or the outcome equations does not ensure identification of average treatment effects. We present two examples below in which higher explanatory power models completely undermine identification of treatment effects.

3.6.1 Outcomes model example

It might be tempting to employ the instrument $Z_5 = \left[\begin{array}{cccccc} 1 & 0 & -1 & -1 & 0 & 1 \end{array}\right]^T$ as a regressor as it perfectly explains observed outcome Y. Estimates are

$$E\left[Y \mid Z_1, D, h\right] = 14+1.0Z_5+0D+0\left(D\times h_1\right)+0\left(\left(1-D\right)\times h_0\right)$$

However, recall our objective is to estimate treatment effects and they draw from outcomes, Y_1 and Y_0, which are only partially observed and Z_5 is independent of these outcomes.[11] This regression produces severe selection bias, disguises endogeneity, suggests homogeneous outcome when it is heterogeneous, and masks self-selection. In other words, it could hardly be more misleading even though it has higher explanatory power.

3.6.2 Selection model example

Suppose we add the regressor,

$$X_1 = \begin{bmatrix} 1 & 0 & 1 & -1 & 0 & -1 \end{bmatrix}^T$$

to the instruments in the selection equation so that the regressors in the probit model are[12]

$$\begin{aligned} W &= \begin{bmatrix} X_1 & Z_1 & Z_2 & Z_3 & Z_4 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 5 & 4 & 3 & 1 \\ 0 & -6 & -5 & -4 & -2 \\ 1 & 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \end{bmatrix} \end{aligned}$$

Again, we suppress probit estimation details. The estimated outcomes model conditional on the "control functions" is

$$E\left[Y \mid Z_1, D, h\right] = 14 + 0D + 0\left(D \times h_1\right) + 0\left((1 - D) \times h_0\right)$$

As in the higher explanatory power outcomes model, this treatment effect identification strategy is a complete bust. Here, it is because the regressor, X_1, dominates the instruments in explaining treatment choice and it's the instruments that allow manipulation of choice without affecting outcome — the key to identifying properties of the counterfactuals. Hence, the regression is plagued by severe selection bias, disguises endogeneity and heterogeneity of outcomes, and hides self-selection inherent to the setting.

---

[11] Identification of instruments is extremely delicate because we don't observe a portion of the outcome distributions.

[12] Employment of a perfect predictor, say

$$X_2 = \begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 \end{bmatrix}^T$$

is well known to create estimation problems. In this case any positive weight on X_2 supplies an equally good fit and makes any other regressors superfluous in the selection equation. Results for the perfect regressor case parallel that presented, except with the perfect predictor, x_2, the coefficients on the control functions, $(D \times h_1)$ and $((1 - D) \times h_0)$, are actually indeterminant since they are a linear combination of the intercept and D.

To summarize, high explanatory power of either the selection equation or outcome equations does not indicate a well-specified model. As the above examples suggest, higher explanatory power can undermine our treatment effect identification strategy. When addressing counterfactuals and treatment effects, we have no choice but to rely on what we know prior to examining the evidence (namely, theory) in specifying the model.[13]

3.7 Bayesian analysis with control function principles

In spite of the apparent success of the classical strategy above, experience suggests Bayesian analysis employing control function principles is more robust than is the classical strategy. Perhaps, this reflects hazard rate (or inverse Mills ratio) sensitivity to estimation error. On the other hand, a Bayesian approach employs least squares estimation on augmented, "complete" data (pseudo-random draws from a truncated normal distribution). That is, instead of extrapolating into the tails via the hazard rate "correction," the Bayesian strategy utilizes data augmentation to "recover" missing counterfactual data.[14]

However, we suspect that it is at least as important that Bayesian analysis helps us or even forces to pay attention to what we know about the setting.[15] Also, Bayesian data augmentation allows the distribution of treatment effects as well as marginal treatment effects to be explored (our discussion above, limits inferences to treatment effect means).[16]

We next turn our attention to Bayesian analysis and consistent reasoning. First, we explore the importance of loss functions, maximum entropy probability assignment, conjugate families, and Bayesian analysis of some primitive data analytic problems. Then, we revisit treatment effects and discuss Bayesian analysis.

---

[13] We don't mean to imply that diagnostic checking based on the evidence is to be shunned. To the contrary, but we must exercise caution and bear in mind how we're exploiting observables to infer unobservables (e.g., counterfactuals).

[14] Bayesian analysis is data intensive. Its application to treatment effects is discussed in some detail in *Accounting and Causal Effects: Econometric Challenges*, ch. 12.

[15] Jaynes, 2003, *Probability Theory: The Logic of Science* gives a riveting account of these ideas.

[16] Heckman and others propose classical, factor analytic strategies to explore treatment effect distributions and marginal treatment effects.