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.[9] 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

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

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

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

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).

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

---

[9] The end of chapter section provides a more general demonstration of linear models equivalence.

and

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

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

$$b = \left[\begin{array}{cc} -1 & 1 \end{array} \right] a = -5 + 10 = 5$$

with

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

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

$$a_1 = \left[\begin{array}{cc} 1 & 1 \end{array} \right] g = 5 + 5 = 10$$

with

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

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

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*.

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

where the regression or conditional expectation is

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

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

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 $Var\left[g \mid X\right] = \sigma^2 \left(X^T X\right)^{-1}$, 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

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

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,[10] 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.

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

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

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

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,

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

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

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).

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.[11] Consistency requires such pooling restrictions satisfy the equality of

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

and

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

as well as

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

and

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

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

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

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

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.

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

$$\begin{array}{rcl} X & = & \left[\begin{array}{cccc} \iota & D & W & (D \times X) \end{array} \right] \\ & = & \left[\begin{array}{cc} X_2 & x_3 \end{array} \right] \end{array}$$

where $X_2 = \left[\begin{array}{ccc} \iota & D & W \end{array} \right]$ 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,

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

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

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

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,

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

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

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

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 $Var[b \mid X] = \sigma^2 \left(X^T X\right)^{-1}$, and the estimated regression or estimated conditional expectation function is $\widehat{Y} = Xb = P_X Y$.[12]

2.6.1 Example

It's time for an example. Continue with the running example except treatment, D, is initially unobserved.[13] 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 $E[\varepsilon \mid X] = 0$. The estimated regression is

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

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

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.

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 $E\left[\varepsilon \mid X\right] = 0$. 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

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

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.

2.7 Linear models equivalence

Earlier in this chapter, we demonstrated via example the equivalence of two linear models for identifying regressor effects on outcome. In this section, we'll generalize the claim and demonstrate equivalence via double residual regression (FWL).[14]

First, we can write the linear model (in error form) $Y = X\beta + \varepsilon$ as

$$Y = X_1\beta_1 + X_2\beta_2 + Z\omega + \varepsilon \tag{M1}$$

or

$$Y = (X_1 + X_2)\,\gamma_1 + X_2\gamma_2 + Z\omega + \varepsilon \tag{M2}$$

where Y is the outcome variable, X_1 and X_2 are regressors, Z are control variables, and ε, for simplicity, are spherical errors (independent with constant variance and $E[\varepsilon \mid X] = 0$).[15] The regressors may be continuous or, as discussed earlier in the chapter, may be indicator variables.[16] Clearly, X_1 and X_2 must have the same number of columns.

Next, we demonstrate two results regarding the estimators: $\widehat{\beta}_1 = \widehat{\gamma}_1$, the estimator for the effect of X_1 on Y conditional on X_2, Z, and the intercept (if one is included) and $\widehat{\gamma}_2 = \widehat{\beta}_2 - \widehat{\beta}_1$, the estimator for the difference between the effect of X_2 on Y and X_1 on Y each conditional on the other variables. Combining the two results implies a third result: $\widehat{\beta}_2 = \widehat{\gamma}_1 + \widehat{\gamma}_2$, the estimator for the effect of X_2 on Y conditional on X_1, Z, and the intercept (if one is included). For notational convenience, let $W = X_1 + X_2$. Now, (M2) can be rewritten as

$$Y = W\gamma_1 + X_2\gamma_2 + Z\omega + \varepsilon \tag{M2}$$

These claims imply (M1) and (M2) are informationally equivalent as the regressors occupy the same subspace (model errors and residuals are identical).

By double residual regression (FWL), we have the following least squares estimators.

$$\begin{aligned}
\widehat{\beta}_1 &= \left((M_2X_1)^T M_2X_1\right)^{-1} (M_2X_1)^T M_2Y \\
&= \left(X_1^T M_2X_1\right)^{-1} X_1^T M_2Y
\end{aligned}$$

---

[14] The equivalence of double residual regression and linear multiple regression is demonstrated in Schroeder's [2010, ch. 3] discussion of FWL and tests of restrictions.

[15] This also represents maximum entropy state of knowledge. If the model includes an intercept Z includes a column of ones along with the control variables.

[16] If the regressors, X_1 and X_2, are indicator variables that collectively sum to the number of observations, the model is constructed without an intercept to maintain linearly independent columns in the design matrix X.

$$\begin{aligned}\widehat{\beta}_2 &= \left((M_1X_2)^T M_1X_2\right)^{-1}(M_1X_2)^T M_1Y \\ &= \left(X_2^T M_1X_2\right)^{-1}X_2^T M_1Y\end{aligned}$$

$$\begin{aligned}\widehat{\gamma}_1 &= \left((M_2W)^T M_2W\right)^{-1}(M_2W)^T M_2Y \\ &= \left(W^T M_2W\right)^{-1}W^T M_2Y\end{aligned}$$

$$\begin{aligned}\widehat{\gamma}_2 &= \left((M_WX_2)^T M_WX_2\right)^{-1}(M_WX_2)^T M_WY \\ &= \left(X_2^T M_WX_2\right)^{-1}X_2^T M_WY\end{aligned}$$

where $M_j = I - P_j$ refers to the projection matrix orthogonal to the column subspace defined by j, P_j refers to the projection defined by the subspace j, and j refers to $\left[\begin{array}{cc} X_1 & Z \end{array}\right]$, $\left[\begin{array}{cc} X_2 & Z \end{array}\right]$, or $\left[\begin{array}{cc} W & Z \end{array}\right]$ for $j = 1, 2$, or W, respectively. For example,

$$\begin{aligned}P_W &= \left[\begin{array}{cc} W & Z \end{array}\right]\left(\left[\begin{array}{cc} W & Z \end{array}\right]^T\left[\begin{array}{cc} W & Z \end{array}\right]\right)^{-1}\left[\begin{array}{cc} W & Z \end{array}\right]^T \\ &= \left[\begin{array}{cc} W & Z \end{array}\right]\left[\begin{array}{cc} W^TW & W^TZ \\ Z^TW & Z^TZ \end{array}\right]^{-1}\left[\begin{array}{c} W^T \\ Z^T \end{array}\right]\end{aligned}$$

and

$$M_W = I - \left[\begin{array}{cc} W & Z \end{array}\right]\left[\begin{array}{cc} W^TW & W^TZ \\ Z^TW & Z^TZ \end{array}\right]^{-1}\left[\begin{array}{c} W^T \\ Z^T \end{array}\right]$$

so that P_W and M_W are orthogonal

$$\begin{aligned}P_WM_W &= \left[\begin{array}{cc} W & Z \end{array}\right]\left[\begin{array}{cc} W^TW & W^TZ \\ Z^TW & Z^TZ \end{array}\right]^{-1}\left[\begin{array}{c} W^T \\ Z^T \end{array}\right] \\ &\quad \times\left(I - \left[\begin{array}{cc} W & Z \end{array}\right]\left[\begin{array}{cc} W^TW & W^TZ \\ Z^TW & Z^TZ \end{array}\right]^{-1}\left[\begin{array}{c} W^T \\ Z^T \end{array}\right]\right) \\ &= \left[\begin{array}{cc} W & Z \end{array}\right]\left[\begin{array}{cc} W^TW & W^TZ \\ Z^TW & Z^TZ \end{array}\right]^{-1}\left[\begin{array}{c} W^T \\ Z^T \end{array}\right] \\ &\quad - \left[\begin{array}{cc} W & Z \end{array}\right]\left[\begin{array}{cc} W^TW & W^TZ \\ Z^TW & Z^TZ \end{array}\right]^{-1}\left[\begin{array}{c} W^T \\ Z^T \end{array}\right] \\ &\quad \times\left[\begin{array}{cc} W & Z \end{array}\right]\left[\begin{array}{cc} W^TW & W^TZ \\ Z^TW & Z^TZ \end{array}\right]^{-1}\left[\begin{array}{c} W^T \\ Z^T \end{array}\right] \\ &= \left[\begin{array}{cc} W & Z \end{array}\right]\left[\begin{array}{cc} W^TW & W^TZ \\ Z^TW & Z^TZ \end{array}\right]^{-1}\left[\begin{array}{c} W^T \\ Z^T \end{array}\right] \\ &\quad - \left[\begin{array}{cc} W & Z \end{array}\right]\left[\begin{array}{cc} W^TW & W^TZ \\ Z^TW & Z^TZ \end{array}\right]^{-1}\left[\begin{array}{c} W^T \\ Z^T \end{array}\right] \\ &= 0\end{aligned}$$

Hence, M_W is orthogonal to $W = X_1 + X_2$ and Z. By analogous reasoning, M_1 is orthogonal to X_1 and Z, and M_2 is orthogonal to X_2 and Z.

Consider the first claim, $\widehat{\beta}_1 = \widehat{\gamma}_1$. This is almost immediately apparent.

$$\widehat{\beta}_1 = \left(X_1^T M_2 X_1\right)^{-1} X_1^T M_2 Y$$

$$\begin{aligned}
\widehat{\gamma}_1 &= \left(W^T M_2 W\right)^{-1} W^T M_2 Y \\
&= \left(\left(X_1 + X_2\right)^T M_2 \left(X_1 + X_2\right)\right)^{-1} \left(X_1 + X_2\right)^T M_2 Y \\
&= \left(X_1^T M_2 X_1 + X_1^T M_2 X_2 + X_2^T M_2 X_1 + X_2^T M_2 X_2\right)^{-1} \\
&\quad \times \left(X_1^T M_2 Y + X_2^T M_2 Y\right)
\end{aligned}$$

Since M_2 annihilates X_2, all terms with $M_2 X_2$ or $X_2^T M_2$ go to zero.

$$\begin{aligned}
\widehat{\gamma}_1 &= \left(X_1^T M_2 X_1 + 0 + 0 + 0\right)^{-1} \left(X_1^T M_2 Y + 0\right) \\
&= \left(X_1^T M_2 X_1\right)^{-1} X_1^T M_2 Y \\
&= \widehat{\beta}_1
\end{aligned}$$

This completes the demonstration of the first claim.

The second claim, $\widehat{\gamma}_2 = \widehat{\beta}_2 - \widehat{\beta}_1$, involves a little more manipulation.

$$\widehat{\gamma}_2 = \left(X_2^T M_W X_2\right)^{-1} X_2^T M_W Y$$

$$\begin{aligned}
\widehat{\beta}_2 - \widehat{\beta}_1 &= \left(X_2^T M_1 X_2\right)^{-1} X_2^T M_1 Y - \left(X_1^T M_2 X_1\right)^{-1} X_1^T M_2 Y \\
&= \left[\left(X_2^T M_1 X_2\right)^{-1} X_2^T M_1 - \left(X_1^T M_2 X_1\right)^{-1} X_1^T M_2\right] Y
\end{aligned}$$

For $\widehat{\gamma}_2 = \widehat{\beta}_2 - \widehat{\beta}_1$, the weights applied to the observed outcomes must be equal

$$\left(X_2^T M_W X_2\right)^{-1} X_2^T M_W = \left(X_2^T M_1 X_2\right)^{-1} X_2^T M_1 - \left(X_1^T M_2 X_1\right)^{-1} X_1^T M_2$$

Post-multiply both sides by X_2. The left hand side becomes

$$\left(X_2^T M_W X_2\right)^{-1} X_2^T M_W X_2 = I$$

and the right hand side is

$$\begin{aligned}
&\left[\left(X_2^T M_1 X_2\right)^{-1} X_2^T M_1 - \left(X_1^T M_2 X_1\right)^{-1} X_1^T M_2\right] X_2 \\
= \;& \left(X_2^T M_1 X_2\right)^{-1} X_2^T M_1 X_2 - \left(X_1^T M_2 X_1\right)^{-1} X_1^T M_2 X_2 \\
= \;& I - 0 = I
\end{aligned}$$

The second term for the right hand side is zero since the projection matrix, M_2, and X_2 are orthogonal. This completes the demonstration of the second claim. Having demonstrated the first two claims, the third claim is satisfied.