

Contents

A	Linear algebra basics	1
A.1	Basic operations	1
A.2	Fundamental theorem of linear algebra	4
A.2.1	Part one	4
A.2.2	Part two	5
A.3	Nature of the solution	7
A.3.1	Exactly-identified	7
A.3.2	Under-identified	8
A.3.3	Over-identified	10
A.4	Matrix decomposition and inverse operations	12
A.4.1	LU factorization	12
A.4.2	Cholesky decomposition	16
A.4.3	Eigenvalues and eigenvectors	17
A.4.4	Singular value decomposition	23
A.4.5	Spectral decomposition	28
A.4.6	quadratic forms, eigenvalues, and positive definiteness	29
A.4.7	similar matrices, Jordan form, and generalized eigen- vectors	29
A.5	Gram-Schmidt construction of an orthogonal matrix	32
A.5.1	QR decomposition	34
A.5.2	Gram-Schmidt QR algorithm	34
A.5.3	Accounting example	35
A.5.4	The Householder QR algorithm	36
A.5.5	Accounting example	36

A.6	Computing eigenvalues	39
A.6.1	Schur's lemma	39
A.6.2	Power algorithm	40
A.6.3	QR algorithm	40
A.6.4	Schur decomposition	42
A.7	Some determinant identities	46
A.7.1	Determinant of a square matrix	46
A.7.2	Identities	47
A.8	Matrix exponentials and logarithms	49
B	Iterated expectations	51
B.1	Decomposition of variance	53
B.2	Jensen's inequality	54
C	Multivariate normal theory	55
C.1	Conditional distribution	57
C.2	Special case of precision	61
C.3	Truncated normal distribution	63
D	Projections and conditional expectations	71
D.1	Gauss-Markov theorem	71
D.2	Generalized least squares (GLS)	74
D.3	Recursive least squares	76
E	Two stage least squares IV (2SLS-IV)	79
E.1	General case	79
E.2	Special case	81
F	Seemingly unrelated regression (SUR)	83
F.1	Classical	84
F.2	Bayesian	84
F.3	Bayesian treatment effect application	85
G	Maximum likelihood estimation of discrete choice models	87
H	Optimization	89
H.1	Linear programming	89
H.1.1	basic solutions or extreme points	90
H.1.2	fundamental theorem of linear programming	91
H.1.3	duality theorems	91
H.1.4	example	92
H.1.5	complementary slackness	92
H.2	Nonlinear programming	93
H.2.1	unconstrained	93
H.2.2	convexity and global minima	93
H.2.3	example	94

H.2.4	constrained — the Lagrangian	94
H.2.5	Karash-Kuhn-Tucker conditions	95
H.2.6	example	96
H.3	Theorem of the separating hyperplane	97
I	Quantum information	99
I.1	Quantum information axioms	99
I.1.1	The superposition axiom	99
I.1.2	The transformation axiom	100
I.1.3	The measurement axiom	100
I.1.4	The combination axiom	101
I.2	Summary of quantum "rules"	103
I.3	Observables and expected payoffs	104
I.4	Density operators and quantum entropy	105
J	Common distributions	109

Appendix D

Projections and conditional expectations

D.1 Gauss-Markov theorem

Consider the data generating process (*DGP*):

$$Y = X\beta + \varepsilon$$

where $\varepsilon \sim (0, \sigma^2 I)$, X is $n \times p$ (with rank p), and $E[X^T \varepsilon] = 0$, or more generally $E[\varepsilon | X] = 0$.

The *Gauss-Markov theorem* states that $b = (X^T X)^{-1} X^T Y$ is the minimum variance estimator of β amongst linear unbiased estimators. Gauss' insight follows from a simple idea. Construct b (or equivalently, the residuals or estimated errors, e) such that the residuals are orthogonal to every column of X (recall the objective is to extract all information in X useful for explaining Y — whatever is left over from Y should be unrelated to X).

$$X^T e = 0$$

where $e = Y - Xb$. Rewriting the orthogonality condition yields

$$X^T (Y - Xb) = 0$$

or the normal equations

$$X^T X b = X^T Y$$

Provided X is full column rank, this yields the usual *OLS* estimator

$$b = (X^T X)^{-1} X^T Y$$

It is straightforward to show that b is unbiased (conditional on the data X).

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

Iterated expectations yields $E[b] = E_X[E[b | X]] = E_X[\beta] = \beta$. Hence, unbiasedness applies unconditionally as well.

$$\begin{aligned} \text{Var}[b | X] &= \text{Var} \left[(X^T X)^{-1} X^T Y | X \right] \\ &= \text{Var} \left[(X^T X)^{-1} X^T (X\beta + \varepsilon) | X \right] \\ &= E \left[\left\{ \beta + (X^T X)^{-1} X^T \varepsilon - \beta \right\} \left\{ (X^T X)^{-1} X^T \varepsilon \right\}^T | X \right] \\ &= (X^T X)^{-1} X^T E[\varepsilon \varepsilon^T] X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} X^T I X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

Now, consider the stochastic regressors case,

$$\text{Var}[b] = \text{Var}_X[E[b | X]] + E_X[\text{Var}[b | X]]$$

The first term is zero since $E[b | X] = \beta$ for all X . Hence,

$$\text{Var}[b] = E_X[\text{Var}[b | X]] = \sigma^2 E \left[(X^T X)^{-1} \right]$$

the unconditional variance of b can only be described in terms of the average behavior of X .

To show that *OLS* yields the minimum variance linear unbiased estimator consider another linear unbiased estimator $b_0 = LY$ (L replaces $(X^T X)^{-1} X^T$). Since $E[LY] = E[LX\beta + L\varepsilon] = \beta$, $LX = I$.

Let $D = L - (X^T X)^{-1} X^T$ so that $DY = b_0 - b$.

$$\begin{aligned} \text{Var}[b_0 | X] &= \sigma^2 \left[D + (X^T X)^{-1} X^T \right] \left[D + (X^T X)^{-1} X^T \right]^T \\ &= \sigma^2 \left(\begin{array}{c} DD^T + (X^T X)^{-1} X^T D^T + DX (X^T X)^{-1} \\ + (X^T X)^{-1} X^T X (X^T X)^{-1} \end{array} \right) \end{aligned}$$

Since

$$LX = I = DX + (X^T X)^{-1} X^T X, DX = 0$$

and

$$\text{Var}[b_0 | X] = \sigma^2 \left(DD^T + (X^T X)^{-1} \right)$$

As DD^T is positive semidefinite, $\text{Var}[b]$ (and $\text{Var}[b | X]$) is at least as small as any other $\text{Var}[b_0]$ ($\text{Var}[b_0 | X]$). Hence, the Gauss-Markov theorem applies to both nonstochastic and stochastic regressors.

Theorem 13 *Rao-Blackwell theorem.* If $\varepsilon \sim N(0, \sigma^2 I)$ for the above *DGP*, b has minimum variance of all unbiased estimators.

Finite sample inferences typically derive from normally distributed errors and t (individual parameters) and F (joint parameters) statistics. Some asymptotic results related to the Rao-Blackwell theorem are as follows. For the Rao-Blackwell *DGP*, *OLS* is consistent and asymptotic normally (*CAN*) distributed. Since *MLE* yields b for the above *DGP* with normally distributed errors, *OLS* is asymptotically efficient amongst all *CAN* estimators. Asymptotic inferences allow relaxation of the error distribution and rely on variations of the laws of large numbers and central limit theorems.

D.2 Generalized least squares (GLS)

For the *DGP*,

$$Y = X\beta + \varepsilon$$

where $\varepsilon \sim N(0, \Sigma)$ and Σ is some general $n \times n$ variance-covariance matrix, then the linear least squares estimator or generalized least squares (*GLS*) estimator is

$$b_{GLS} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$$

with

$$Var [b_{GLS} | X] = (X^T \Sigma^{-1} X)^{-1}$$

If Σ is known, this can be computed by ordinary least squares (*OLS*) following transformation of the variables Y and X via Γ^{-1} where Γ is a triangular matrix such that $\Sigma = \Gamma \Gamma^T$, say via Cholesky decomposition. Then, the transformed *DGP* is

$$\begin{aligned} \Gamma^{-1} Y &= \Gamma^{-1} X \beta + \Gamma^{-1} \varepsilon \\ y &= x \beta + \epsilon \end{aligned}$$

where $y = \Gamma^{-1} Y$, $x = \Gamma^{-1} X$, and $\epsilon = \Gamma^{-1} \varepsilon \sim N(0, I)$. To see where the identity variance matrix comes from, consider

$$\begin{aligned} Var [\Gamma^{-1} \varepsilon] &= \Gamma^{-1} Var [\varepsilon] (\Gamma^{-1})^T \\ &= \Gamma^{-1} \Sigma (\Gamma^{-1})^T \\ &= \Gamma^{-1} \Gamma \Gamma^T (\Gamma^{-1})^T \\ &= \Gamma^{-1} \Gamma \Gamma^T (\Gamma^T)^{-1} \\ &= I \end{aligned}$$

Hence, estimation involves projection of y onto x

$$E[y | x] = xb$$

where

$$\begin{aligned} b &= (x^T x)^{-1} x^T y \\ &= \left(X^T (\Gamma^{-1})^T \Gamma^{-1} X \right)^{-1} X^T (\Gamma^{-1})^T \Gamma^{-1} Y \end{aligned}$$

Since $\Sigma^{-1} = (\Gamma \Gamma^T)^{-1} = (\Gamma^{-1})^T \Gamma^{-1}$, we can rewrite

$$b = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$$

which is the *GLS* estimator for β . Further, $Var [b | X] = E [(b - \beta)(b - \beta)^T]$. Since

$$\begin{aligned}
 b - \beta &= (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y - \beta \\
 &= (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} (X\beta + \varepsilon) - \beta \\
 &= (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} X\beta - \beta \\
 &\quad + (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \varepsilon - \beta \\
 &= \beta + (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \varepsilon - \beta \\
 &= (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \varepsilon
 \end{aligned}$$

$$\begin{aligned}
 Var [b | X] &= E [(b - \beta)(b - \beta)^T] \\
 &= E [(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \varepsilon \varepsilon^T \Sigma^{-1} X (X^T \Sigma^{-1} X)^{-1} | X] \\
 &= (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} E [\varepsilon \varepsilon^T | X] \Sigma^{-1} X (X^T \Sigma^{-1} X)^{-1} \\
 &= (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \Sigma \Sigma^{-1} X (X^T \Sigma^{-1} X)^{-1} \\
 &= (X^T \Sigma^{-1} X)^{-1}
 \end{aligned}$$

as indicated above.

D.3 Recursive least squares

Suppose the analyst expects a series of noisy signals that require filtering to uncover the signals of interest, β , where the *DGP* is

$$Y = X\beta + \varepsilon, \quad \varepsilon | X \sim (0, V)$$

Instead of recalculating everything with each sample, the analyst can employ recursive least squares to achieve the same results. The idea revolves around the design matrix, X_t , Fisher's information matrix, \mathfrak{S}_t , and weights from the variance-covariance matrix, V_t , for the period t draw.

$$\mathfrak{S}_t = \mathfrak{S}_{t-1} + X_t^T V_t^{-1} X_t$$

where the components may be augmented with zeroes so that the matrices conform.

$$\mathfrak{S}_t = \begin{bmatrix} \mathfrak{S}_{t-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & X_t^T V_t^{-1} X_t \end{bmatrix}$$

Let $K_t = \mathfrak{S}_t^{-1} X_t^T V_t^{-1}$ represent the gain, $Y_t - X_t b_{t-1}$ be the innovation, and $b_1 = \mathfrak{S}_1^{-1} X_1^T V_1^{-1} Y_1 = (X_1^T V_1^{-1} X_1)^{-1} X_1^T V_1^{-1} Y_1$ be the first sample least squares estimate.¹ Then, the recursive least squares estimate is

$$b_t = b_{t-1} + K_t (Y_t - X_t b_{t-1})$$

where again b_{t-1} may be augmented with zeroes to conform.

Example 14 (smooth accruals) Suppose the *DGP* for cash flows is

$$\begin{aligned} cf_t &= m_t + e_t \\ m_t &= g m_{t-1} + \varepsilon_t \end{aligned}$$

the variance-covariance matrix, V , is diagonal, and $\nu = \frac{\sigma_\varepsilon}{\sigma_e}$, m_0 and g are known. Then, $accruals_{t-1}$ and cf_t are, collectively, sufficient statistics for the mean of cash flows m_t based on the history of cash flows and $g^{t-1} accruals_t$ is an efficient statistic for m_t

$$\begin{aligned} [\hat{m}_t | cf_1, \dots, cf_t] &= g^{t-1} accruals_t \\ &= \frac{1}{den_t} \left\{ \frac{num_t}{g^2} cf_t + g^{t-1} \nu^2 den_{t-1} accruals_{t-1} \right\} \end{aligned}$$

where $accruals_0 = m_0$, $\begin{bmatrix} den_t \\ num_t \end{bmatrix} = B^t \begin{bmatrix} den_0 \\ num_0 \end{bmatrix} = S \Lambda^t S^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $B = \begin{bmatrix} 1 + \nu^2 & \nu^2 \\ g^2 & g^2 \nu^2 \end{bmatrix}$, Λ is diagonal matrix with the eigenvalues of B , and S is

¹If prior beliefs regarding β are informed then b_0 representing priors is included to construct b_1 in analogous fashion to other samples.

a matrix of the corresponding eigenvectors for B . The variance of accruals is equal to the variance of the estimate of the mean of cash flows multiplied by $g^{2(t-1)}$; the variance of the estimate of the mean of cash flows equals the coefficient on current cash flow multiplied by σ_e^2 , $\text{Var}[\hat{m}_t] = \frac{\text{num}_t}{\text{den}_t g^2} \sigma_e^2$. The

development employs recursive least squares. Let $X_1 = \begin{bmatrix} -\nu \\ 1 \end{bmatrix}$ ($a 2 \times 1$ matrix), $X_2 = \begin{bmatrix} g\nu & -\nu \\ 0 & 1 \end{bmatrix}$ ($a 2 \times 2$ matrix), $X_t = \begin{bmatrix} 0 & \cdots & 0 & g\nu & -\nu \\ 0 & \cdots & 0 & 0 & 1 \end{bmatrix}$ ($a 2 \times t$ matrix with $t-2$ leading columns of zeroes), $Y_1 = \begin{bmatrix} -g\nu m_0 \\ cf_1 \end{bmatrix}$, $Y_2 = \begin{bmatrix} 0 \\ cf_2 \end{bmatrix}$, and $Y_t = \begin{bmatrix} 0 \\ cf_t \end{bmatrix}$. The information matrix for a t -period cash flow history is

$$\begin{aligned} \mathfrak{S}_t &= \mathfrak{S}_{t-1}^a + X_t^T X_t \\ &= \begin{bmatrix} 1 + \nu^2 + g^2 \nu^2 & -g\nu^2 & 0 & \cdots & 0 \\ -g\nu^2 & 1 + \nu^2 + g^2 \nu^2 & -g\nu^2 & \ddots & \vdots \\ 0 & -g\nu^2 & \ddots & -g\nu^2 & 0 \\ \vdots & \ddots & -g\nu^2 & 1 + \nu^2 + g^2 \nu^2 & -g\nu^2 \\ 0 & \cdots & 0 & -g\nu^2 & 1 + \nu^2 \end{bmatrix}, \end{aligned}$$

a symmetric tri-diagonal matrix, where \mathfrak{S}_{t-1}^a is \mathfrak{S}_{t-1} augmented with a row and column of zeroes to conform with \mathfrak{S}_t . For instance, $\mathfrak{S}_1 = [1 + \nu^2]$ and $\mathfrak{S}_1^a = \begin{bmatrix} 1 + \nu^2 & 0 \\ 0 & 0 \end{bmatrix}$. The estimate of the mean of cash flows is derived recursively as

$$\hat{m}_t = \hat{m}_{t-1}^a + K_t (z_t - X_t^a \hat{m}_{t-1}^a)$$

for $t > 1$ where $K_t = \mathfrak{S}_t^{-1} X_t^T$, the gain matrix, and \hat{m}_{t-1}^a is \hat{m}_{t-1} augmented with a zero to conform with \hat{m}_t . The best linear unbiased estimate of the current mean is the last element in the vector \hat{m}_t and its variance is the last row-column element of \mathfrak{S}_t^{-1} multiplied by σ_e^2 .

Example 15 (special case) Suppose $g = \nu = 1$ for the above DGP. Then,

$$\begin{aligned} [\hat{m}_t | cf_1, \dots, cf_t] &= \text{accruals}_t \\ &= \frac{1}{F_{2t+1}} \{F_{2t} cf_t + F_{2t-1} \text{accruals}_{t-1}\} \end{aligned}$$

and variance for the most recent mean estimate (the t th element) is

$$\text{Var}[\hat{m}_t | cf_1, \dots, cf_t]_{tt} = \sigma^2 \frac{F_{2t}}{F_{2t+1}}$$

where $F_t = F_{t-1} + F_{t-2}$, the Fibonacci series.