Contents

Appendix D

Projections and conditional expectations

D.1 Gauss-Markov theorem

Consider the data generating process (*DGP*):

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

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

The *Gauss-Markov theorem* states that $b = \left(X^T X\right)^{-1} X^T Y$ 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

$$X^T \left(Y - Xb\right) = 0$$

or the normal equations

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

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

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

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

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

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

$$\begin{aligned} Var[b \mid X] &= Var\left[\left(X^TX\right)^{-1}X^TY \mid X\right] \\ &= Var\left[\left(X^TX\right)^{-1}X^T(X\beta+\varepsilon) \mid X\right] \\ &= E\left[\left\{\beta+\left(X^TX\right)^{-1}X^T\varepsilon-\beta\right\}\left\{\left(X^TX\right)^{-1}X^T\varepsilon\right\}^T \mid X\right] \\ &= \left(X^TX\right)^{-1}X^TE\left[\varepsilon\varepsilon^T\right]X\left(X^TX\right)^{-1} \\ &= \sigma^2\left(X^TX\right)^{-1}X^TIX\left(X^TX\right)^{-1} \\ &= \sigma^2\left(X^TX\right)^{-1} \end{aligned}$$

Now, consider the stochastic regressors case,

$$Var[b] = Var_X[E[b \mid X]] + E_X[Var[b \mid X]]$$

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

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

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 $\left(X^TX\right)^{-1}X^T$). Since $E[LY] = E[LX\beta + L\varepsilon] = \beta$, $LX = I$.

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

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

Since

$$LX = I = DX + \left(X^TX\right)^{-1}X^TX, DX = 0$$

and

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

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

Theorem 13 *Rao-Blackwell theorem. If* $\varepsilon \sim N\left(0, \sigma^2 I\right)$ *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

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

with

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

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

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

Hence, estimation involves projection of y onto x

$$E\left[y \mid x\right] = xb$$

where

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

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

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

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

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

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

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

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

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, $\Im_t$, and weights from the variance-covariance matrix, V_t, for the period t draw.

$$\Im_t = \Im_{t-1} + X_t^T V_t^{-1} X_t$$

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

$$\Im_t = \begin{bmatrix} \Im_{t-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & X_t^T V_t^{-1} X_t \end{bmatrix}$$

Let $K_t = \Im_t^{-1} X_t^T V_t^{-1}$ represent the gain, $Y_t - X_t b_{t-1}$ be the innovation, and $b_1 = \Im_1^{-1} X_1^T V_1^{-1} Y_1 = \left(X_1^T V_1^{-1} X_1\right)^{-1} X_1^T V_1^{-1} Y_1$ be the first sample least squares estimate.[1] Then, the recursive least squares estimate is

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

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 $\nu = \frac{\sigma_e}{\sigma_\epsilon}$, 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

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

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

[1] If 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, $Var\left[\widehat{m}_t\right] = \frac{num_t}{den_t g^2}\sigma_e^2$. The development employs recursive least squares. Let $X_1 = \begin{bmatrix} -\nu \\ 1 \end{bmatrix}$ (a 2×1 matrix), $X_2 = \begin{bmatrix} g\nu & -\nu \\ 0 & 1 \end{bmatrix}$ (a 2×2 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

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

a symmetric tri-diagonal matrix, where $\Im_{t-1}^a$ is $\Im_{t-1}$ augmented with a row and column of zeroes to conform with $\Im_t$. For instance, $\Im_1 = \left[1+\nu^2\right]$ and $\Im_1^a = \begin{bmatrix} 1+\nu^2 & 0 \\ 0 & 0 \end{bmatrix}$. The estimate of the mean of cash flows is derived recursively as

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

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

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

$$\begin{aligned} \left[\widehat{m}_t | cf_1, ..., cf_t\right] &= accruals_t \\ &= \frac{1}{F_{2t+1}}\left\{F_{2t} cf_t + F_{2t-1} accruals_{t-1}\right\} \end{aligned}$$

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

$$Var\left[\widehat{m}_t | cf_1, ..., cf_t\right]_{tt} = \sigma^2 \frac{F_{2t}}{F_{2t+1}}$$

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