EC327: Financial Econometrics, Spring 2008

Wooldridge, Introductory Econometrics (3rd ed, 2006)

Appendix D:
Summary of matrix algebra

Basic definitions

A **matrix** is a rectangular array of numbers, with m rows and n columns, which are the *row dimension* and *column dimension*, respectively. The matrix **A** will have typical element a_{ij}.

A **vector** is a matrix with one row and/or one column. Thus a *scalar* can be considered a 1×1 matrix. A m-element **row vector** has one row and m columns. A n-element **column vector** has n rows and one column.

A **square matrix** has $m = n$. A **diagonal matrix** is a square matrix with off-diagonal elements equal to 0. It may have m distinct diagonal elements. If those m elements are equal, it is a **scalar matrix**. If they all equal 1, it is an **identity matrix** of order m, customarily written as $\mathbf{I}$ or $\mathbf{I}_m$.

A **symmetric matrix** is a square matrix for which $a_{ij} = a_{ji}\ \forall i, j$. The elements above and below its *main diagonal* are equal. It is often written in *upper triangular* or *lower triangular* form, since there is no need to report more than the main diagonal and sub- (super-) diagonal elements.

A symmetric matrix we often compute in econometrics is the *correlation matrix* of a set of variables. The correlation matrix will have 1s on its main diagonal (since every variable is perfectly correlated with itself) and off-diagonal values between (-1,+1).

Another very important matrix is what Stata calls the VCE, or estimated variance-covariance matrix of the estimated parameters of a regression equation. It is by construction a symmetric matrix with positive elements on the main diagonal (the estimated variances of the parameters b, whose positive square roots are the reported standard errors of those parameters). The off-diagonal elements are the estimated covariances of the estimated parameters, used in computing hypothesis tests or confidence intervals involving more than one parameter. This matrix may be examined after a `regress` command in Stata with the command `estat vce`.

The **transpose** of a matrix reflects the matrix around its main diagonal. We write the transpose of **A** as **A'** (or, less commonly, $\mathbf{A}^T$). If **B = A'**, then $b_{ij} = a_{ji}\ \forall i, j$. If **A** is $m \times n$, **B** is of order $n \times m$. The rows of **A** become the columns of **A**, and *vice versa*.

Several relations involving the transpose of a matrix:
(1) the transpose of a row vector is a column vector;
(2) the transpose of a transpose is the matrix itself;
(3) the transpose of a symmetric matrix is the matrix itself;
(4) The transpose of the sum (difference) is the sum (difference) of the transposes.

A matrix of any row and column dimension with all elements equal to zero is a **zero matrix** or **null matrix**. It plays the role in matrix algebra (or *linear algebra*) that the scalar 0 plays in ordinary algebra, in the sense that adding or subtracting a null matrix has no effect, and multiplying by a null matrix returns a null matrix.

The identity matrix $\mathbf{I}$ plays the role of the number 1 in ordinary algebra: e.g., multiplying by $\mathbf{I}$

has no effect, and we can always insert an $\mathbf{I}$ of appropriate order in a matrix product without changing the product.

Matrix operations

In matrix algebra, algebraic operations are only defined for matrices or vectors of appropriate order. Since vectors are special cases of matrices, we will only speak of matrices. Two matrices **A** and **B** are **equal** if they have the same *order*—the same number of rows and columns—and if $a_{ij} = b_{ij}\ \forall i, j$. Addition or subtraction can be performed if and only if the matrices are of the same order. The sum (difference) $\mathbf{C} = \mathbf{A} \pm \mathbf{B}$ is defined as $c_{ij} = a_{ij} \pm b_{ij}\ \forall i, j$. That is, just as we compare matrices for equality element-by-element, we add (subtract) matrices element-by-element.

Scalar multiplication is defined for any matrix as $\mathbf{C} = \mathrm{k}\ \mathbf{A}$, and involves multiplying each element by that scalar: $c_{ij} = k \times a_{ij}\ \forall i, j$.

Matrix multiplication is defined for matrices $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{n\times q}$ as $\mathbf{C}_{m\times q} = \mathbf{A}\ \mathbf{B}$. This product is defined since the number of *columns* in the first matrix equals the number of *rows* in the second matrix. To define matrix multiplication, we must first define the **dot product** of vectors $\mathbf{u}_{n\times 1}$ and $\mathbf{v}_{n\times 1}$. The product $d = u'v$ is a scalar, defined as $d = \sum_{i=1}^{n} u_i v_i$. This implies that if **u = v**, the scalar d will be the sum of squares of the elements of **u**. We may also compute the **outer product** of these vectors, **u v'**, which for vectors of the same length will create a $n \times n$ matrix.

In matrix multiplication, each element of the result matrix is defined as a dot product. If **C = A B**, c_{ij} is the dot product of the i^{th} row of **A** and the j^{th} column of **B**: $c_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$. These dot products are only defined for vectors of the same length, which gives rise to the constraint on the number of columns of **A**

and the number of rows of **B**. When this constraint is satisfied, the vectors are *conformable* for multiplication.

In ordinary algebra, multiplication is *commutative*: we can write **x y** or **y x** and receive the same result. In matrix algebra, the product of arbitrary matrices will not exist unless they are *conformable*. if one of those products exists, the other will not except in special cases. If **A** and **B** are square matrices of the same order, then both **A B** and **B A** always exist, but they do not yield the same result matrix except under special circumstances. (A natural exception: we can always write **I A** and **A I**, both of which equal **A**. Likewise for the null matrix.)

We can also multiply a matrix by a vector. If we write **C = u A**, we are multiplying the m-element row vector **u** by matrix **A**, which

must have m rows; it may have any number of columns. We are *premultiplying* **A** by **u**. We may also *postmultiply* **A** by vector **v**. If **A** is $m \times n$, **v** must be a n-element column vector.

Some properties of matrix multiplication:
(1) multiplication is *associative* with respect to scalars and matrices, as long as they are conformable:
$k(\mathbf{A} + \mathbf{B}) = k\mathbf{A} + k\mathbf{B}$,
$\mathbf{C}\ (\ \mathbf{A} + \mathbf{B}) = \mathbf{C}\ \mathbf{A} + \mathbf{C}\ \mathbf{B}$,
$(\ \mathbf{A} + \mathbf{B})\ \mathbf{C} = \mathbf{A}\ \mathbf{C}\ + \mathbf{B}\ \mathbf{C}$.
(2) The transpose of a product is the product of the transposes, in reverse order:
$\mathbf{(A\ B)' = B'\ A'}$.
(3) For any matrix **X** the products **X'X**
and **X X'** exist, and are symmetric.

We cannot speak of dividing one matrix by another (unless they are both scalars, and the rules of ordinary algebra apply). Instead, we

define the concept of the **inverse matrix**. A square matrix $\mathbf{A}_{m\times m}$ may possess a unique *inverse*, written $\mathbf{A}^{-1}$, under certain conditions. If it exists, the inverse is defined as that matrix which satisfies $\mathbf{A}\ \mathbf{A}^{-1} = \mathbf{A}^{-1}\ \mathbf{A} = \mathbf{I}_m$, and in that sense the inverse operates like the division operator in normal algebra, where $x \times \frac{1}{x} = 1\ \forall x \neq 0$. A matrix which possesses an inverse is said to be *nonsingular*, or *invertible*, and it has a nonzero *determinant*. A *singular* matrix possesses a zero determinant. We will not go into the computation of inverse matrices or determinants—which is better left to a computer—but we must understand their importance to econometrics and the linear regression model in particular.

Properties of inverses:
(1) The inverse of a product is the product of the inverses, in reverse order: $\mathbf{(A\ B)}^{-1} = \mathbf{B}^{-1}\ \mathbf{A}^{-1}$. The transpose of the inverse is the inverse of the transpose: $(\mathbf{A'})^{-1} = (\mathbf{A}^{-1})'$.

The *trace* of a square matrix, denoted **tr(A)**, is the scalar sum of its diagonal elements. So, for instance, the trace of the identity matrix $\mathbf{I}_n$ is n. The trace of the sum (difference) of matrices is the sum (difference) of the traces, and the trace of a product yielding a square matrix is not dependent on order: that is, if $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{n\times m}$, both the products **AB** and **BA** exist, and have the same trace.

Linear independence and rank

The notion of a matrix possessing an inverse is related to the concept of *linear independence*. A set of n-vectors [$\mathbf{x}_1$, $\mathbf{x}_2$, . . . , $\mathbf{x}_r$] is said to be *linearly independent* if and only if $\alpha' \mathbf{x} = \mathbf{0}$ implies that α is the null vector. If $\alpha' \mathbf{x} = \mathbf{0}$ holds for a set of scalars $\alpha_1, \alpha_2, \ldots, \alpha_r$ that are not all zero, then the set of **x**-vectors are **linearly dependent**. This implies that at least one vector

in this set can be written as a linear combination of the others. In econometric terms, this is the problem of *perfect collinearity*: one or more of the regressors can be expressed as an exact linear combination of the others.

If **A** is a $n \times k$ matrix, the *column rank* of **A** is the maximum number of linearly independent columns of **A**. If rank(**A**) $= k$, **A** has *full column rank*. Row rank is defined similarly. The *rank of a matrix* is the minimum of its row and column ranks. When we consider a data matrix $\mathbf{X}_{n \times k}$, with $n > k$, its rank cannot exceed k. If a square matrix **A** of order k is of full rank (rank=k), then $\mathbf{A}^{-1}$ exists: the matrix is invertible. The rank of a product of matrices cannot exceed either of their ranks, and may be zero.

Quadratic forms and positive definite matrices

If **A** is a $n \times n$ symmetric matrix, then the *quadratic form* associated with **A** is the scalar function

$$Q = x'Ax = \sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{i=1}^{n}\sum_{j>i} a_{ij}x_i x_j$$

where **x** is any n-vector. If $x'Ax > 0$ for all n-vectors **x** except **x=0**, then matrix **A** is said to be *positive definite* (p.d.). If $x'Ax \geq 0$, then **A** is said to be *positive semi-definite* (p.s.d.). A p.d. matrix **A** has all positive diagonal elements and possesses an inverse $\mathbf{A}^{-1}$ which is also p.d. and a positive determinant. For any $\mathbf{X}_{n\times k}$, **X'X** and **XX'** are p.s.d. If $\mathbf{X}_{n\times k}$ with $n > k$ has rank k, then **X'X** is p.d. and non-singular, implying that $\mathbf{X'X}^{-1}$ exists.

This is the relevant concern for regression, where we have a data matrix **X** of n observations on k variables or regressors, with $n > k$. If those regressors are linearly independent, so that **X** is of full rank k, we can invert the matrix **X'X**: a key step in computing linear regression estimates.

Idempotent matrices

If $\mathbf{A}$ is a $n \times n$ symmetric matrix, it is said to be *idempotent* if and only if $\mathbf{A}\ \mathbf{A} = \mathbf{A}$. The identity matrix, playing the role of the number 1 in ordinary algebra, is idempotent. An idempotent matrix has rank equal to its trace, and it is p.s.d.

If we have a data matrix $\mathbf{X}_{n\times k}$ with rank($\mathbf{X}$)=k, then

$$P = X(X'X)^{-1}X' \quad (1)$$
$$M = I_n - X(X'X)^{-1}X' = I_n - P \quad (2)$$

are both symmetric, idempotent matrices. $\mathbf{P}_{n\times n}$ has rank k, while $\mathbf{M}_{n\times n}$ has rank $n-k$, since the trace of $\mathbf{P}$ is that of $\mathbf{I}_k$, and its trace is equal to its rank.

Matrix differentiation

For n-vectors **a** and **x**, define $f(x) = a'x$. Then the derivative of the function with respect to its (vector) argument is

$$\partial f / \partial x = a'$$

a $1 \times n$ vector. For a $n \times n$ symmetric matrix **A** with quadratic form $Q = x'Ax$, the derivative of **Q** with respect to its vector argument is

$$\partial Q / \partial x = 2x'A$$

a $1 \times n$ vector.

Moments and distributions of random vectors

Operations on random variables can be expressed in terms of vectors of random variables. If **y** is a n-element random vector, then its expected value **E[y]** is merely the n-vector of its expected values. This generalizes to a

random matrix. A linear transformation of **y** with non-random matrices yields

$$E[Ay + b] = AE[y] + b$$

If **y** is a n-element random vector, then its *variance-covariance matrix* or VCE is the symmetric matrix

Var(y) =

$$\begin{pmatrix} \sigma_1^2 & & \\ \sigma_{21} & \sigma_2^2 & \\ \vdots & \vdots & \\ \sigma_{n1} & \sigma n2 & \ldots \sigma_n^2 \end{pmatrix}$$

where σ_j^2 is the variance of $\mathbf{y}_j$ and σ_{ij} is the covariance of $\mathbf{y}_i$ and $\mathbf{y}_j$.

Just as we can perform algebra on scalar variances and covariances, we can operate on the VCE of **y**. Some properties:
(1) If **a** is a n-element nonrandom vector, **Var(a'y)**

= [a' Var(y) a] ≥ 0.
(2) If **Var(a'y)** > 0 $\forall a \neq$**0**, **Var(y)** is p.d. and possesses an inverse.
(3) If $\mu =$**E[y]**, $Var(y) = E[(y-\mu)(y-\mu)']$.
(4) If the elements of **y** are uncorrelated, **Var(y)** is a diagonal matrix. This is the assumption of independence of the elements of random vector **y**: for instance, of the error terms of a regression equation.
(5) If in addition **Var($\mathbf{y}_j$) =** σ^2 $\forall j$, then **Var(y)** $= \sigma^2 I_n$. This is the assumption of homoskedasticity of the elements of random vector **y**: for instance, of the error terms of a regression equation.
(6) For nonrandom $\mathbf{A}_{m\times n}$ and $\mathbf{b}_{n\times 1}$, **Var(A y + b) = [A Var(y) A']**.

If **y** is a n-element *multivariate Normal* random vector with mean vector μ and VCE Σ, we write $y \sim N(\mu, \Sigma)$. Properties of the multivariate normal distribution:

(1) If $y \sim N(\mu, \Sigma)$, each element of **y** is Normally distributed.
(2) If $y \sim N(\mu, \Sigma)$, any two elements of **y** are independent if and only if they are uncorrelated ($\sigma_{ij} = 0$).
(3) If $y \sim N(\mu, \Sigma)$ and **A, b** are nonrandom, then **A y + b** $\sim N(A\mu + b, A\Sigma A')$

A χ^2_n random variable is the sum of n squared independent standard Normal variables. If $u \sim N(0, I_n)$, then **u'u** $\sim \chi^2_n$.

A **t**-distributed random variable is the ratio of a standard Normal variable **Z** to a χ^2_n random variable **X**, standardized by its degrees of freedom, where the variables **Z, X** are independent. Let $u \sim N(0, I_n)$, **c** be a nonrandom n-vector and **A** be a nonrandom, $n \times n$ symmetric, idempotent matrix with rank q, with **Ac = 0**. Then the quantity $[c'u/\sqrt{c'c}]/\sqrt{u'Au} \sim t_q$.

A **F**-distributed random variable is the ratio of two independent χ^2 random variables, standardized by their respective degrees of freedom. If $u \sim N(0, I_n)$ and **A, B** are $n \times n$ non-random, symmetric idempotent matrices with rank(**A**) $= k_1$ and rank(**B**) $= k_2$, then $((u'Au)/k_1)/(u'Bu)/k_2) \sim F_{k_2}^{k_1}$.

Appendix E:
The linear regression model in matrix form

OLS estimation

The multiple linear regression model with k regressors can be written as

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + u_t,\ t = 1, 2, \ldots, n$$

where y_t is the dependent variable for observation t and $x_1 \ldots x_k$ are the regressors. β_0 is the intercept term (constant) and $\beta_1 \ldots \beta_k$ are the slope parameters.

We define $\mathbf{x}_t$ as the $1 \times (k+1)$ row vector $(1, x_{t1}, \ldots, x_{tk})$ and the column vector $\beta = (\beta_0, \beta_1, \ldots, \beta_k)'$. Then the model can be written as

$$y_t = \mathbf{x}_t \beta + u_t,\ t = 1, 2, \ldots, n$$

and the entire regression problem as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} \tag{3}$$

where **y** is the $n \times 1$ vector of observations on the dependent variable, $\mathbf{X}_{n\times(k+1)}$ is the matrix of observations on the regressors, including an initial column of 1s, and **u** is the $n \times 1$ vector of unobservable errors.

OLS estimation involves minimizing the sum of squared residuals, where the residuals $e_t = (y_t - x_t\mathbf{b})$ where **b** is the vector of estimated parameters corresponding to the population parameters β. Minimizing the sum of squared e_t is equivalent to minimizing **SSR = e'e** with respect to the $(k+1)$ elements of **b**. A solution to this optimization problem involves a set of $(k+1)$ *first order conditions* (FOC)

$$\partial SSR(\mathbf{b})/\partial \mathbf{b} = 0$$

and setting those conditions equal to zero. The FOCs give rise to a set of $(k+1)$ simultaneous equations in the $(k+1)$ unknowns **b**. In

matrix form, the residuals may be written as $e = y - X\mathbf{b}$, and the FOCs then become

$$\begin{aligned} X'(y - X\mathbf{b}) &= 0 \\ X'y &= (X'X)\mathbf{b} \\ b &= (X'X)^{-1}X'y \end{aligned} \tag{4}$$

if the inverse exists, that is, if and only if **X** is of full (column) rank $(k+1)$. If that condition is satisfied, the *cross-products matrix* **X'X** will be a *positive definite* matrix. For that to be so, it must be the case that the columns of **X** are linearly independent. This rules out the case of *perfect collinearity*, which will arise if one or more of the regressors can be written as a linear combination of the others.

Recall that the first column of **X** is a vector of 1s. This implies that no other column of **X** can take on a constant value, for it would be a multiple of the first column. Likewise, a

situation where the sum of some of the regressors equals a column of ones will violate this assumption. This will occur in the case of the *dummy variable trap* where a complete set of *mutually exclusive and exhaustive* (MEE) indicator variables are included in a regression containing a constant term.

Given a solution for $\mathbf{b}$, the OLS predicted (fitted) values $\widehat{y} = X\mathbf{b}$, and the calculated residuals equal $e = y - \widehat{y}$. From Equation (**??**), the first order condition may be written as $\mathbf{X}'\mathbf{e} = \mathbf{0}$. Since the first column of $\mathbf{X}$ is a vector of 1s, the FOC implies that the residuals sum to zero and have an average value of zero. The remaining FOCs imply that each column of $\mathbf{X}$ (and any linear combinations of those columns) has zero covariance with $\mathbf{e}$. This is algebraically implied by OLS, not an assumption. Since $\widehat{y}$ is such a linear combination, it is also true that $\hat{\mathbf{y}}'\mathbf{e} = \mathbf{0}$: the residuals have zero covariance with the predicted values.

Finite sample properties of OLS

The assumptions underlying the OLS model:

1. *Linear in parameters*: The model can be written as in Equation (**??**), with the observed n-vector **y**, observed $n \times (k+1)$ matrix **X** and n-vector **u** of unobserved errors.

2. *No perfect collinearity*: The matrix **X** has rank $(k+1)$.

3. *Zero conditional mean*: Conditional on the matrix **X**, each element of **u** has zero mean: $E[u_t|\mathbf{X}] = 0 \ \forall t$.

4. *Spherical disturbances*: $Var(u|\mathbf{X}) = \sigma^2 I_n$. This combines the two assumptions of *homoskedasticity*, $Var(u_t|\mathbf{X}) = \sigma^2 \ \forall t$ and independence, $Cov(u_t, u_s|\mathbf{X} = 0 \ \forall t \neq s$. The

first assumption implies that the u_t are *identically* distributed, while the second assumption implies that they are *independently* distributed, or in a time series context, free of serial correlation. Taken together, they allow us to say that u is a *i.i.d.* random variable with a *scalar variance-covariance matrix*.

Given these four assumptions, we may prove several theorems related to the OLS estimator:

1. *Unbiasedness of OLS*: Under assumptions 1, 2 and 3, the OLS estimator b is unbiased with respect to β.

$$\begin{aligned} b &= (X'X)^{-1}X'y \\ &= (X'X)^{-1}X'(X\beta + u) \\ &= (X'X)^{-1}(X'X)\beta + (X'X)^{-1}X'u \end{aligned} \tag{5}$$

Taking the conditional expectation,

$$\begin{aligned} E[b|\mathbf{X}] &= \beta + (X'X)^{-1}X' \; E(u|\mathbf{X}) \quad (6) \\ &= \beta + (X'X)^{-1}X'0 \\ &= \beta \end{aligned}$$

so that b is unbiased.

2. *VCE of the OLS estimator*: Under assumptions 1, 2, 3 and 4,

$$Var(b|\mathbf{X}) = \sigma^2(X'X)^{-1}. \quad (7)$$

From the last line of Equation (**??**), we have

$$\begin{aligned} Var(b|\mathbf{X}) &= Var[(X'X)^{-1}X'u|X] \quad (8) \\ &= (X'X)^{-1}X'[Var(u|X)]X(X'X)^{-1} \end{aligned}$$

Crucially depending on assumption 4, we can then write

$$\begin{aligned} Var(b|\mathbf{X}) &= (X'X)^{-1}X'(\sigma^2 I_n)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}(X'X)(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \quad (9) \end{aligned}$$

This conditional VCE depends on the unknown parameter σ^2, but replacing that with its consistent estimate s^2 it becomes an operational formula.

3. *Gauss–Markov*: Under assumptions 1, 2, 3 and 4, b is the Best Linear Unbiased Estimator of β (b is *BLUE*). Any *linear* estimator of β can be written as

$$\widehat{\beta} = A'y \tag{10}$$

where $\mathbf{A}$ is a $n \times (k+1)$ matrix whose elements are not functions of **y** but may be functions of **X**. Given the model of Equation (**??**), we may write $\widehat{\beta}$ as

$$\widehat{\beta} = A'(X\beta + u) = (A'X)\beta + A'u. \tag{11}$$

We may then write the conditional expectation of $\widehat{\beta}$ as

$$\begin{aligned} E[\widehat{\beta}|X] &= A'X\beta + E[A'u|X] \\ &= A'X\beta + A'E(u|X) \\ &= A'X\beta \end{aligned} \tag{12}$$

The last line following from assumption 3. For $\widehat{\beta}$ to be an unbiased estimator, it must be that $E[\widehat{\beta}|X] = \beta \ \forall \beta$. Because $\mathbf{A'X}$ is a $(k+1) \times (k+1)$ matrix, unbiasedness requires that $\mathbf{A'X = I_{k+1}}$.

From Equation (**??**) we have

$$Var(\widehat{\beta}|X) = A'[Var(u|X)]A = \sigma^2 A'A \quad (13)$$

invoking assumption 4 ($i.i.d.$ disturbances). Therefore

$$\begin{aligned} Var(\widehat{\beta}|X) - & \\ Var(b|X) &= \sigma^2[A'A - (X'X)^{-1}] \\ &= \sigma^2[A'A - A'X(X'X)^{-1}X'A] \\ &= \sigma^2 A'[I_n - X(X'X)^{-1}X']A \\ &= \sigma^2 A'MA \qquad (14) \end{aligned}$$

where $M = [I_n - X(X'X)^{-1}X']$, a symmetric and idempotent matrix which is positive semi-definite (p.s.d.) for any matrix $\mathbf{A}$. But Equation (**??**) represents the difference between the VCE of any arbitrary linear estimator of β and the VCE of the OLS

estimator. That difference will be the null matrix if and only if $A'A = (X'X)^{-1}$: that is, if we choose A to reproduce the OLS estimator. For any other choice of A, the difference will be a positive semi-definite matrix, which implies that at least one of the diagonal elements is larger in the first matrix than in the second. That is, the maximum precision estimator of each element of β can only be achieved by OLS. Any other linear, unbiased estimator of β will have a larger estimated variance for at least one element of β. Thus OLS is *BLUE*: the Best (minimum-variance, or most efficient) Linear Unbiased Estimator of β.

4. *Unbiasedness of* s^2: The unbiased estimator of the error variance σ^2 can be calculated as $s^2 = e'e/(n-k-1)$, with $E[s^2|X] =$

$\sigma^2 \ \forall \ \sigma^2 > 0$.

$$\begin{aligned} e &= y - Xb \\ &= y - X[(X'X)^{-1}X'y] \\ &= My = Mu \end{aligned} \tag{15}$$

where M is defined as above. Since M is symmetric and idempotent,

$$e'e = u'M'Mu = u'Mu \tag{16}$$

which is a scalar, equal to its trace. So

$$\begin{aligned} E[u'Mu|X] &= E[tr(u'Mu)|X] \\ &= E[tr(Muu')|X] \\ &= tr[E(Muu')|X] \\ &= tr[ME[uu'|X]] \\ &= tr(M\sigma^2 I_n) = \sigma^2 tr(M) \\ &= \sigma^2(n-k-1) \end{aligned} \tag{17}$$

because $tr(M) = tr(I_n) - tr[X[(X'X)^{-1}X'] = n - tr(I_{k+1})$. Therefore,

$$E[s^2|X] = E[u'Mu|X)/(n-k-1) = \sigma^2 \tag{18}$$

Statistical inference

With an additional assumption

5. *Normality*: conditional on **X**, the u_t are independently and identically distributed $(i.i.d.)$ as Normal$[0, \sigma^2]$. The vector of errors **u**, conditional on **X**, is distributed multivariate Normal$[0, \sigma^2 I_n]$.

we may present the theorem

Normality of b: Under assumptions 1,2,3,4,5, b conditional on **X** is distributed as multivariate Normal$[\beta, \sigma^2(X'X)^{-1}]$.

with the corollary that under the null hypothesis, t-statistics have a t_{n-k-1} distribution.

Under normality, b is the *minimum variance unbiased estimator* of β, conditional on **X**, in the sense that it reaches the *Cramer-Rao lower*

bound (CRLB) for the VCE of unbiased estimators of β. The CRLB defines the minimum variance possible for any unbiased estimator—linear or nonlinear. Since OLS reaches that bound, it is the most precise unbiased estimator available.