

Multivariate Normal and Chi-squared Distributions*

1 Matrix Algebra Background

1.1 Matrix multiplication

A first useful observation is the following. Let $A_{m \times n} = [a_1 \dots a_j \dots a_n]$ with columns $a_j = (a_{1j}, \dots, a_{mj}) \in \mathbb{R}^m$, $j = 1, \dots, n$. For $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ we have

$$Ax = \begin{pmatrix} a_{11} \dots a_{1n} \\ a_{21} \dots a_{2n} \\ \vdots \\ a_{m1} \dots a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + \dots + a_{1n}x_n \\ a_{21}x_1 + \dots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + \dots + a_{mn}x_n \end{pmatrix} = a_1x_1 + \dots + a_nx_n = \sum_i a_i x_i.$$

In words, Ax is the combination of the columns of A with coefficients given by the elements of x . Observe in particular that for a square $n \times n$ matrix A we have $Ae_j = a_j - e_j$ being the j th unit vector of $\mathbb{R}^n$. More generally:

Proposition 1 (Matrix Products). *Let $A_{m \times n} = [a_1 \dots a_n]$ with columns $a_j \in \mathbb{R}^m$ as before. Then*

$$AA'_{m \times m} = [a_1 \dots a_n] \begin{bmatrix} a'_1 \\ \vdots \\ a'_n \end{bmatrix} = \sum_{i=1}^n a_i a'_i.$$

Proof. Let b_h denote the transpose of the h -th row of A , so that $A' = [b_1 \dots b_m]$; and let b_h have components b_{ih} so that $b_h = (b_{1h} \dots b_{nh})'$. Then

$$\begin{aligned} AA' &= A(b_1 \dots b_m) = (Ab_1 \dots Ab_m) = \left(\sum_i b_{i1}a_i \dots \sum_i b_{im}a_i \right) \\ &= \sum_i (b_{i1}a_i \dots b_{im}a_i) = \sum_i a_i(b_{i1} \dots b_{im}). \end{aligned}$$

*Compiled by Salvatore Modica, August 2026. I consulted www.statlect.com and of course AI.

But $(b_{i1} \dots b_{im})$ is the i -th row of A' , the transpose of the i -th column of A , which is a'_i . $\square$

1.2 Inverse of a square matrix

Given a square $n \times n$ matrix A , its inverse matrix A^{-1} (if it exists) satisfies $A^{-1}A = I$. The definition of determinants is assumed known.

Lemma 1. *A square $n \times n$ matrix A is invertible if and only if its columns are independent.*

Proof. If the columns are independent then the image of the map $x \mapsto Ax$ is all of $\mathbb{R}^n$; and the map is also injective since $Ax = 0 \iff x = 0$. So the map is invertible. If on the other hand the columns are not independent there is a nonzero solution x^* to $Ax = 0$; therefore if the inverse A^{-1} existed we would get $x^* = A^{-1}Ax^* = A^{-1}(0) = 0$, contradiction. $\square$

Remark. The map $x \mapsto Ax$ is easily verified to be linear on $\mathbb{R}^n$.¹ We have tacitly asserted that *the inverse of an invertible linear transformation is linear*, and here is a proof. Invertible means bijective, as usual. Let S be such a map and T its inverse; we have to show additivity and homogeneity. For the latter, for vector u we must show $T(\alpha u) = \alpha T(u)$; but $u = S(x)$ for a unique x so also $\alpha u = S(\alpha x)$; therefore $\alpha T(u) = \alpha x T(\alpha u)$. For additivity take another vector $v = S(y)$; By linearity of S we have $u + v = S(x + y)$, therefore $T(u + v) = x + y = T(u) + T(v)$.

Lemma 2. *The determinant of a square matrix is equal to the product of its eigenvalues.*

Proof. The eigenvalues λ_i are the roots of the polynomial $p(\lambda) = \det(\lambda I - A)$; from $p(\lambda) = (\lambda - \lambda_1) \cdots (\lambda - \lambda_n)$, for $\lambda = 0$ we get $\det(-A) = (-1)^n \prod_i \lambda_i$; but $\det(-A) = (-1)^n \det A$. $\square$

Proposition 2. *A square matrix A is invertible if and only if its determinant is non-zero.*

Proof. Direct from the previous 2 lemmas: A is invertible iff $Ax = 0 \implies x = 0$, that is iff A does not have a null eigenvalue. $\square$

1.3 Symmetric matrices

A square matrix A is *symmetric* if the transpose $A' = A$; it is *orthogonal* if $AA' = I$ (which also implies $A'A = I$). The main result on symmetric matrices is the following. A proof is presented in Appendix 1.

Theorem 1 (Spectral Theorem for Real Symmetric Matrices). *A (real) symmetric matrix A can be decomposed as $S'AS = \Lambda$ with S orthogonal and Λ diagonal, where the elements of S are the eigenvectors of A and the diagonal elements of Λ their corresponding eigenvalues.*

¹In fact $A = [a_1 \dots a_n]$ represents the linear transformation T on $\mathbb{R}^n$ characterized by $Te_j = a_j$ in the canonical basis $\{e_1, \dots, e_n\}$ in the sense that then $Ax = Tx$ for all x .

1.3.1 Symmetric idempotent matrices

A square matrix A is *idempotent* if $A^2 = A$. Also recall that the *rank* of a matrix A (not necessarily square) is the number of its independent columns.

Proposition 3. *A symmetric idempotent matrix $A_{k \times k}$ of rank m can be written as $A = VV'$ with $V_{k \times m} = (v_1 \dots v_m)$, v_i orthonormal eigenvectors corresponding to the nonzero eigenvalues of A .*

Proof. Order the eigenvectors so that in the Spectral Theorem the v_i 's are the first columns of S and correspondingly the first m elements of Λ are 1. Then $SAS' = VV'$. Check this: $S\Lambda = (v_1 \dots v_m \ 0 \dots 0)$ so (using Proposition 1) $SAS' = \sum_{i=1}^m v_i v_i' = VV'$. The v_i are orthonormal because after rearrangement they are the first m columns of S so orthogonality is inherited from that of S . $\square$

1.3.2 Symmetric positive definite matrices

A square matrix A is positive definite if $x'Ax > 0$ for any $x \neq 0$.

Proposition 4. *A symmetric positive definite matrix is invertible.*

Proof. The determinant of a square matrix is equal to the product of its eigenvalues, and none of them can be zero since it is positive definite (for the eigenvector $v \neq 0$ corresponding to a null eigenvalue we would have $v'Av = 0$). $\square$

Theorem 2 (Cholesky Decomposition). *A symmetric positive definite matrix A can be written as $A = LL'$ with L lower triangular with positive diagonal entries.*

A proof is in Appendix 2. Observe that the L in the theorem is invertible since its determinant is the product of its diagonal entries. Another useful fact is the following:

Proposition 5. *If $A_{k \times m}$ has rank m then $A'A$ is positive definite.*

Proof. $x'A'Ax = (Ax)'Ax = \|Ax\|^2 \geq 0$, and equality holds iff $Ax = 0$. But this and full rank of A implies $x = 0$. $\square$

2 Multivariate Normal and Chi-squared Distributions

In what follows X is $k \times 1$ random vector and $\text{Cov}(X) = E[(X - EX)(X - EX)']$.

2.1 Basic facts

A random vector X is *non-degenerate* if $\text{Var}(a'X) > 0$ for all $a \neq 0$. You are supposed to know what independent random variables are.

Proposition 6. *Any covariance matrix $\text{Cov}(X)$ with X non-degenerate is symmetric positive definite. Hence it is invertible (by Proposition 4).*

Proof. Symmetry is obvious. We claim that for $a \in \mathbb{R}^k$ we have $a'\text{Cov}(X)a = \text{Var}(a'X)$. To check this let $V = \text{Cov}(X)$; then $a'Va = \sum_{i,j} a_i a_j V_{ij}$; on the other hand

$$\text{Var}(a'X) \equiv \text{Var}\left(\sum a_i X_i\right) = \sum_{i,j} a_i a_j \text{Cov}(X_i, X_j) = \sum_{i,j} a_i a_j V_{ij}. \quad \square$$

Proposition 7. *If $Y_{p \times 1} = a + BX$ with $a \in \mathbb{R}^p$ and $B_{p \times k}$ then $EY = a + B \cdot EX$ and $\text{Cov}(Y) = B\text{Cov}(X)B'$.*

Proof. Check for variance and take $a = 0$ (otherwise replace with $Y - a$).

$$\begin{aligned} \text{Cov}(Y) &= E[(Y - EY)(Y - EY)'] = E[(BX - B \cdot EX)(BX - B \cdot EX)'] \\ &= BE[(X - EX)(X - EX)']B'. \end{aligned} \quad \square$$

2.2 Multivariate Normal

Definition 1. $Z_{k \times 1}$ is *standard* multivariate normal - written $Z \sim N_k(0, I)$ - if it has density

$$\varphi_k(z) = (2\pi)^{-k/2} \exp(-z'z/2) \quad z \in \mathbb{R}^k$$

Note that $\varphi_k(z) = \varphi(z_1) \cdot \dots \cdot \varphi(z_k)$ with φ standard normal 1-dimensional density, so the components of Z are independent standard normal. Therefore $E(Z) = 0, \text{Cov}(Z) = I$.

Definition 2. Let μ be a $k \times 1$ vector and V a $k \times k$ symmetric positive definite matrix. We say that X is multivariate normal with mean μ and covariance matrix V if it has density

$$\varphi_k(x) = (2\pi)^{-k/2} |\det V|^{-1/2} \exp(-(x - \mu)'V^{-1}(x - \mu)/2) \quad x \in \mathbb{R}^k$$

This is written as $X \sim N_k(\mu, V)$. Note that V is invertible (see Proposition 4) hence its determinant is nonzero (Proposition 2). *Remark:* there are others equivalent definitions, but what matters most is the following:

Proposition 8 (Characterization). *$X \sim N_k(\mu, V)$ iff for some $\Sigma_{k \times k}$ such that $\Sigma\Sigma' = V$ it is $X = \mu + \Sigma Z$ with $Z \sim N_k(0, I)$.*

Proof. Assume first that $X \sim N_k(\mu, V)$. By Proposition 6 V is symmetric positive definite, hence by Theorem 2 there is an invertible Σ such that $V = \Sigma\Sigma'$. Now $V^{-1} = (\Sigma\Sigma')^{-1} = (\Sigma')^{-1}\Sigma^{-1} = (\Sigma^{-1})'\Sigma^{-1}$; and $|\det V|^{-1/2} = |\det \Sigma\Sigma'|^{-1/2} = |\det \Sigma|^{-1}$. Therefore

$$\begin{aligned}\varphi_k(x) &= (2\pi)^{-k/2} |\det V|^{-1/2} \exp\left(-(x - \mu)'V^{-1}(x - \mu)/2\right) \quad x \in \mathbb{R}^k \\ &= (2\pi)^{-k/2} |\det \Sigma|^{-1} \exp\left(-[\Sigma^{-1}(x - \mu)]'\Sigma^{-1}(x - \mu)/2\right)\end{aligned}$$

which by the change of variable formula is exactly the density of $\mu + \Sigma Z$ with $Z \sim N_k(0, I)$. In the other direction, if $X = \mu + \Sigma Z$ with $Z \sim N_k(0, I)$ the above computation (read backwards) yields $X \sim N_k(\mu, V)$ with $V = \Sigma\Sigma'$. $\square$

Remark. This characterization, via Proposition 7, implies that $E(X) = \mu$ and $\text{Cov}(X) = V$, confirming that the parameters in Definition 2 are indeed the mean and covariance matrix of X . By Proposition 6 the restriction to symmetric positive definite V is without loss of generality, since every covariance matrix of a non-degenerate random vector has this form.

Proposition 9 (Linear Transformations). *If $X \sim N_k(\mu, V)$ and $Y = a + BX$ with $a \in \mathbb{R}^p$ and $B_{p \times k}$ then $Y \sim N_p(a + B\mu, BV B')$.*

Proof. All except normality follows from Proposition 7. To check normality: with $X = \mu + \Sigma Z$ as in the theorem we get $Y = a + B\mu + B\Sigma Z$. $\square$

2.3 The univariate case

In the univariate case $k = 1$ the last corollary reduces to: If $X \sim N(\mu, \sigma^2)$ then $Y = aX + b \sim N(a\mu + b, a^2\sigma^2)$ - the subscript omitted. As usual we let $Z \sim N(0, 1)$ denote the standard univariate normal and Φ its cdf. Corollary 9 then implies

Corollary 1. *If $X \sim N(\mu, \sigma^2)$, then its cdf F is given by*

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right),$$

and $\Pr(\mu - \kappa\sigma \leq X \leq \mu + \kappa\sigma) = \Pr(-\kappa < Z < \kappa) = \Phi(\kappa) - \Phi(-\kappa)$ for $\kappa \in \mathbb{R}_+$.

Proof. It suffices to observe that by Proposition 9 $(X - \mu)/\sigma \sim N(0, 1)$. $\square$

2.4 Chi-squared distributions

The Chi-squared distribution with k degrees of freedom χ_k^2 is defined as the distribution of $Z'Z$ with $Z \sim N_k(0, I)$ - that is the sum of the squares of k independent standard normals.

Proposition 10. *If $X \sim N_k(\mu, V)$ then $(X - \mu)'V^{-1}(X - \mu) \sim \chi_k^2$.*

Proof. Take $\mu = 0$ without loss of generality. By Theorem 2 (Cholesky) we may write $V = \Sigma \Sigma'$. Then $X'(\Sigma \Sigma')^{-1}X = X'(\Sigma^{-1})'\Sigma^{-1}X = Z'Z$ with $Z = \Sigma^{-1}X$; and $\text{Cov}(Z) = \Sigma^{-1}V(\Sigma^{-1})' = I$ (Proposition 9) so $Z \sim N_k(0, I)$ whence $Z'Z \sim \chi_k^2$. $\square$

The Chi-squared distributions arise from the standard normal via quadratic transformations of idempotent matrices.

Proposition 11. *If $X \sim N_k(0, I)$ and $A_{k \times k}$ is symmetric idempotent with rank m then $X'AX \sim \chi_m^2$.*

Proof. Let $Y = X'AX$. Write $A = VV'$ as in Proposition 3, with $V_{k \times m}$. Then $Y = X'VV'X = (V'X)'V'X = Z'Z$ with $Z = V'X$. Since $V'V = I_m$ we get $Z \sim N_m(0, I)$ from Proposition 8, so $Z'Z \sim \chi_m^2$. $\square$

3 Applications

Normal random variable are ubiquitous in statistic, probability, finance... We end these notes with a few typical applications to elementary Statistics.

3.1 Sample mean and variance of normal data

The setting here is that we draw a sample of n realization of a variable of interest from a population, assuming that in the population the variable has a normal distribution. Formally we are given a set of independent (scalar) variables $X_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, n$; $\bar{X} = \sum_i X_i/n$ is the sample mean.

Proposition 12. *With $X_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, n$ independent, let*

$$W = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad V = \frac{\sum_i (X_i - \bar{X})^2}{\sigma^2}$$

denote respectively standardized sample mean and variance. Then $W \sim N(0, 1)$ and $V \sim \chi_{n-1}^2$. W and V are independent.

Proof. It suffices to assume $\mu = 0, \sigma^2 = 1$ - otherwise just take $\tilde{X}_i = (X_i - \mu)/\sigma$. Independence of the X_i 's implies that $X \equiv (X_1, \dots, X_n) \sim N_n(0, I_n)$. Use Gram-Schmidt² to construct an orthonormal matrix $B_{n \times n}$ whose first row is $(1/\sqrt{n})_{1 \times n}$ (all elements equal to $1/\sqrt{n}$) and let $Y = BX$. Then $Y \sim N(0, BB') = N(0, I_n)$ using Proposition 9 and the orthogonality of B , so that the Y_i 's are independent standard normals. Observe that $Y_1 = (1/\sqrt{n})_{1 \times n} X = \sqrt{n} \bar{X} = W$, establishing the first claim. Next, we have $Y'Y = X'X$

²The procedure is spelled out in the remark at the end of the "Projections" section below.

(since B is orthogonal) that is $\sum_i Y_i^2 = \sum_i X_i^2$; but $Y_1^2 = n\bar{X}^2$, therefore $\sum_{i=2}^n Y_i^2 = \sum_i X_i^2 - n\bar{X}^2 = \sum_i (X_i - \bar{X})^2 = V$; this proves the second claim. Independence follows from the fact that $W = Y_1$ and V depends on $(Y_2, \dots, Y_n)$. $\square$

The last two topics are Maximum Likelihood and OLS estimation. In both cases we first need to know how to compute derivatives of certain simple functions.

3.2 Gradients of Linear and Quadratic forms

Let $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$, with f thought of as an $m \times 1$ column vector. Recall that the gradient $\nabla f(x)$ of f is the $m \times n$ matrix of the partial derivatives, with elements $\nabla f(x)_{ij} = \partial f_i(x)/\partial x_j \equiv f_{ij}$. Its j th column is $\nabla f(x)_j = \partial f(x)/\partial x_j$.³

3.2.1 Gradients of linear forms

Let $A_{m \times n} = (a_1 \dots a_n)$ with columns $a_j \in \mathbb{R}^m$. To compute the gradient of the map $x \in \mathbb{R}^n \mapsto Ax \in \mathbb{R}^m$ observe that $Ax = \sum_i a_i x_i$ implies $\partial Ax/\partial x_j = a_j$, so - as direct generalization of the scalar case - we have

$$\nabla(Ax) = A.$$

3.2.2 Gradients of quadratic forms

With squared $A_{m \times m} = (a_1 \dots a_m)$ and again $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$, we want the gradient of $f' Af: \mathbb{R}^n \rightarrow \mathbb{R}$. This gradient $\nabla(f' Af)$ is $1 \times n$.

Proposition 13. $\nabla(f' Af) = f'(A' + A)\nabla f$. It reduces to $2f' A \nabla f$ if A is symmetric.

Proof. We have $f' Af = \sum_{j=1}^m (\sum_{h=1}^m f_h a_{hj}) f_j = \sum_{h,j} a_{hj} f_h f_j$, whence

$$\begin{aligned} \frac{\partial f' Af}{\partial x_i} &= \sum_{h,j} a_{hj} (f_{hi} f_j + f_h f_{ji}) = \sum_j f_j \sum_h a_{hj} f_{hi} + \sum_h f_h \sum_j a_{hj} f_{ji} \\ &= \sum_j f_j (A' \nabla f)_{ji} + \sum_h f_h (A \nabla f)_{hi} = (f' A' \nabla f)_i + (f' A \nabla f)_i \\ &= (f'(A + A') \nabla f)_i. \end{aligned}$$

$\square$

3.3 Maximum Likelihood Estimation of Normal Mean

We observe a sample of n data which we assume independently drawn from a Normal distribution, $X_1, X_2, \dots, X_n \sim N_k(\mu, V)$ where V is given. We want to estimate the mean vector

³You may have known $\nabla f(x)$ as the Jacobian of f .

μ by maximizing its log-likelihood. It turns out that isolating the terms containing μ yields the objective function:

$$\ln L(\mu) = -\frac{1}{2} \sum_{i=1}^n (X_i - \mu)' V^{-1} (X_i - \mu).$$

Expanding the single term gives

$$(X_i - \mu)' V^{-1} (X_i - \mu) = X_i' V^{-1} X_i - 2\mu' V^{-1} X_i + \mu' V^{-1} \mu,$$

and since this is a concave function to find the maximum we have to solve the appropriate First Order Conditions. To this end we must compute the gradients of the linear form $-2\mu' V^{-1} X_i$ and of the quadratic form $\mu' V^{-1} \mu$. Since V^{-1} is symmetric we have

$$\nabla_{\mu}(-2\mu' V^{-1} X_i) = -2V^{-1} X_i \quad \text{and} \quad \nabla_{\mu}(\mu' V^{-1} \mu) = 2V^{-1} \mu.$$

The maximum of $\ln L$ is then found by setting $\nabla_{\mu} \ln L(\mu) = 0$:

$$0 = \nabla_{\mu} \ln L(\mu) = -\frac{1}{2} \sum_{i=1}^n (-2V^{-1} X_i + 2V^{-1} \mu)$$

that is $V^{-1} \sum_{i=1}^n X_i = nV^{-1} \mu$. Since V (being a covariance matrix) is invertible so is V^{-1} , therefore premultiplying by V we arrive at the maximum likelihood estimator

$$\hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Thus $\hat{\mu}_{MLE}$ is simply the sample mean vector.

3.4 Geometry of OLS and Distribution of Residuals

The simplest regression model is specified as a linear relation $y = X\beta + u$, where $y \in \mathbb{R}^n$ is a $n \times 1$ vector of observations on the dependent variable, $X_{n \times m}$ is a matrix of observations on m explanatory variables (“regressors”), $\beta_{m \times 1}$ is a vector of coefficients to be estimated, and $u \sim N_n(0, \sigma^2 I)$ is a vector of independent normal random disturbances. It is assumed that $n \geq m$ and that X has rank m .

The Ordinary Least Squares criterion seeks the parameter vector $\hat{\beta}$ that makes $X\beta$ closest to y ; that is the β which minimizes

$$\|y - X\beta\|^2 = (y - X\beta)'(y - X\beta) = y'y - 2\beta'X'y + \beta'X'X\beta$$

(observe that $y'X\beta = \beta'X'y$ is a scalar). This is a convex function, to minimize which we

must again set the gradient to zero. Now

$$\nabla_{\beta}(-2\beta'X'y) = -2X'y \quad \text{and} \quad \nabla_{\beta}(\beta'X'X\beta) = 2X'X\beta$$

therefore the FOC is $-2X'y + 2X'X\beta = 0$, which yields the so-called “normal equations” $X'X\beta = X'y$. Since X is of full rank, by Proposition 5 $X'X$ is positive definite; it is also symmetric, hence invertible (Proposition 4). Therefore the OLS estimate is

$$\hat{\beta} = (X'X)^{-1}X'y.$$

Proposition 14. $\hat{\beta} \sim N_m(\beta, \sigma^2(X'X)^{-1})$.

Proof. $u \sim N_n(0, \sigma^2 I)$, therefore $\hat{\beta}$ is a linear transformation of a normal variable:

$$\hat{\beta} = (X'X)^{-1}X'(X\beta + u) = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u = \beta + (X'X)^{-1}X'u$$

hence normal by Proposition 9. Proposition 7 implies directly that $E(\hat{\beta}) = \beta$, and that

$$\begin{aligned} \text{Cov}(\hat{\beta}) &= \text{Cov}((X'X)^{-1}X'u) = [(X'X)^{-1}X'] \text{Cov}(u) [(X'X)^{-1}X']' \\ &= [(X'X)^{-1}X'] (\sigma^2 I_n) [(X'X)^{-1}X']' = \sigma^2 (X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1}, \end{aligned}$$

where we used the easily verified fact that $[(X'X)^{-1}]' = (X'X)^{-1}$. □

The estimated y (the variable of interest) is

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y \equiv Hy.$$

We next want to look closer at the residuals

$$e_{n \times 1} \equiv y - \hat{y},$$

to do which we must understand what the (quite special) $H_{n \times n}$ matrix above represents.

Projection Matrices

Let $V_{n \times m} = (v_1, \dots, v_m)$ with $v_i \in \mathbb{R}^k$ independent (so $n \geq m$). Let $\mathcal{C}(V) = \{Vb \in \mathbb{R}^n : b \in \mathbb{R}^m\}$ be the column span of V (an m -dimensional subspace of $\mathbb{R}^n$). For $x \in \mathbb{R}^n$ let us find the $y \in \mathcal{C}(V)$ at minimum distance from x . Intuitively this is the orthogonal projection of x onto

$\mathcal{C}(V)$, and we shall confirm it. The problem is

$$\min_{y \in \mathcal{C}(V)} \|x - y\|^2.$$

Incidentally, this has a solution because the distance function is continuous and since $0 \in \mathcal{C}(V)$ the inf problem can be reduced to the compact set $K = \{y \in \mathcal{C}(V) : \|x - y\|^2 \leq \|x\|^2\}$. Since $y \in \mathcal{C}(V)$ iff $y = Vb$ for some $b \in \mathbb{R}^m$ the problem is $\min_{b \in \mathbb{R}^m} \|x - Vb\|^2 = \min_{b \in \mathbb{R}^m} (x - Vb)'(x - Vb)$. The distance is also convex; the FOC for this is $0 = \nabla_b(x - Vb)'(x - Vb) = -2(x - Vb)'V$; after transposing this is $V'x = V'Vb$. Since V is full rank then $V'V$ (an $m \times m$ matrix) is invertible⁴ so $b = (V'V)^{-1}V'x$; hence the solution $Vb \in \mathcal{C}(V)$ is given by

$$V(V'V)^{-1}V'x \equiv Hx, \quad H = V(V'V)^{-1}V'.$$

The $n \times n$ matrix $H_{n \times n}$ is symmetric and idempotent, and it has rank m . It is easy to check the first two properties; to derive its rank observe that since V and $(V'V)^{-1}$ are full rank we have $Hx = 0 \iff V'x = 0$, so that H has the same null space as V' ; this has dimension $n - m$, so the rank of H is $n - (n - m) = m$.⁵ Since $(Hx)'(x - Hx) = x'H'x - x'H'Hx = x'Hx - x'Hx = 0$ we see that Hx is indeed the orthogonal projection of x onto $\mathcal{C}(V)$.

The converse is also true: any symmetric idempotent $H_{n \times n}$ of rank m is the projection matrix onto an m -dimensional subspace. Indeed, from Corollary 3 we can write $H = VV'$ with $V_{n \times m} = (v_1 \dots v_m)$ and v_i 's orthonormal. Since $V'V = I_m$ we have $H = V(V'V)^{-1}V'$, the projection matrix onto the column span of V .

Remark (The Gram-Schmidt Procedure). This is a way to construct an orthonormal basis of $\mathbb{R}^n$. Pick any vector v_1 ; then pick a second vector $\tilde{v}_2$ independent of it, project it onto the 1-dimensional space spanned by v_1 , let $\text{Proj}(\tilde{v}_2)$ be this projection and let $v_2 = \tilde{v}_2 - \text{Proj}(\tilde{v}_2)$; then take $\tilde{v}_3$ independent of the first two, project it onto their 2-dimensional span and let v_3 be the difference between $\tilde{v}_3$ and its projection. And so on until the last, n -th independent vector. Then normalize each vector obtained by dividing by its norm.

Back to OLS

Since $\hat{y} = Hy$, we have learnt that the estimate $\hat{y}$ is the orthogonal projection of the observed y onto the column space of X - if you visualize it is a pretty unsurprising result. Recall that $e_{n \times 1} \equiv y - \hat{y}$ denote the residuals. We know that orthogonality means e is orthogonal to the

⁴For this it suffices to show that $V'Vx = 0 \implies x = 0$ for $x \in \mathbb{R}^m$. If $V'Vx = 0$ then

$$0 = x'V'Vx = (Vx)'(Vx) = \|Vx\|^2$$

that is $Vx = 0$. Since V has independent columns this implies $x = 0$.

⁵You are supposed to know the rank-nullity theorem: for $A_{n \times m}$ we have $\text{Rank}(A) + \text{Nullity}(A) = m$.

column span of X . And indeed

$$X'e = X'(y - Hy) = X'(I_n - H)y = X'y - X'Hy = 0.$$

Our last step will be to derive the distribution of the residual variance.

Proposition 15. *Let $H = X(X'X)^{-1}X'$ be the projection matrix onto $\mathcal{C}(X)$. Then the rank of $I_n - H$ is $n - m$.*

Proof. Observe that $x \in \text{Null}(I_n - H) \iff (I_n - H)x = 0 \iff Hx = x$, which occurs iff $x \in \mathcal{C}(X)$. Therefore $\text{Null}(I_n - H) = \mathcal{C}(X)$, which implies $\text{Dim}(\text{Null}(I_n - H)) = m$. Conclusion follows from the rank-nullity relation $\text{Rank}(I_n - H) + \text{Dim}(\text{Null}(I_n - H)) = n$. $\square$

From this one easily obtains the distribution of the (scaled) sum of the squares of the residuals - the residual variance of the estimate.

Proposition 16. $\frac{e'e}{\sigma^2} \sim \chi_{n-m}^2$.

Proof. We know that $I_n - H$ is symmetric idempotent (immediate to check) of rank $n - m$. Next observe that $HX = X$; therefore

$$e = y - \hat{y} = (I - H)y = (I - H)(X\beta + u) = (I - H)X\beta + (I - H)u = (I - H)u$$

whence $e'e = u'(I_n - H)'(I_n - H)u = u'(I_n - H)u$. Since $u/\sigma \sim N_n(0, I_n)$ the assertion follows from Proposition 11. $\square$

On this result depends the computation of the crucial goodness-of-fit metric $\bar{R}^2$, the “adjusted R^2 ”. The unadjusted R^2 is defined as

$$R^2 = 1 - \frac{e'e}{\sum_i (y_i - \bar{y})^2}.$$

At one extreme $R^2 = 1$ denotes a perfect fit: $\hat{y}_i = y_i$ all i ; at the other extreme $R^2 = 0$ means the estimate $\hat{y}_i$ is as good as $\bar{y}$ - that is the regression exercise has zero value.

The unadjusted R^2 fails to account for the fact that $e'e/\sigma^2 \sim \chi_{n-m}^2$. This fact makes $e'e/(n - m)$ the natural unbiased estimate of σ^2 . Letting $\hat{\sigma}^2 = e'e/(n - m)$ and inserting the unbiased estimate of the total variance $s_y^2 = (n - 1)^{-1} \sum_i (y_i - \bar{y})^2$ we arrive at the *adjusted* R^2 (which is the value the statistical software packages report), defined as

$$\bar{R}^2 = 1 - \frac{\hat{\sigma}^2}{s_y^2}.$$

Appendix 1. Proof of Theorem 1 (the Spectral Theorem).

Recall that for a complex matrix $M = A + iB$ the conjugate transpose is defined as $M^* = A' - iB'$. The theorem in the text follows from the following more general result:

Theorem (The Schur Decomposition Theorem). *Let A be an $n \times n$ square matrix with complex entries. There exists a unitary matrix U ($U^*U = I$) and an upper-triangular matrix T such that $A = UTU^*$, and the diagonal entries of T are precisely the eigenvalues of A .*

Proof. The proof proceeds by induction on the dimension of the vector space. Base Case: $n = 1$. Any 1×1 matrix $A = [a]$ is already upper-triangular. We can choose $U = [1]$, which is unitary, and $T = [a]$. The claim holds trivially. For the inductive step assume now that the Schur Decomposition holds true for all $(n - 1) \times (n - 1)$ matrices. We must show that it then holds for any $n \times n$ matrix A .

First find an eigenvector. By the Fundamental Theorem of Algebra the characteristic polynomial of A has at least one complex root. Therefore, A has at least one eigenvalue λ_1 and a corresponding non-zero eigenvector v_1 .

Next normalize v_1 so that $\|v_1\| = 1$, and by the Gram-Schmidt process extend $\{v_1\}$ to an orthonormal basis for $\mathbb{C}^n$, say $\{v_1, v_2, \dots, v_n\}$. Let V_1 be the $n \times n$ unitary matrix whose columns are these basis vectors: $V_1 = [v_1 \dots v_n]$. Now examine the matrix product $V_1^*AV_1$. Because v_1 is an eigenvector of A (meaning $Av_1 = \lambda_1 v_1$), the first column of the resulting matrix has the form

$$V_1^*Av_1 = V_1^*(\lambda_1 v_1) = \lambda_1(V_1^*v_1) = \begin{bmatrix} \lambda_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Therefore the full block matrix structure of $V_1^*AV_1$ is

$$V_1^*AV_1 = \begin{bmatrix} \lambda_1 & w^* \\ 0 & A_1 \end{bmatrix}$$

where w^* is a $1 \times (n - 1)$ row vector, 0 is an $(n - 1) \times 1$ zero column vector, and A_1 is an $(n - 1) \times (n - 1)$ matrix.

By our induction hypothesis there exists an $(n - 1) \times (n - 1)$ unitary matrix U_1 and an upper-triangular matrix T_1 such that $A_1 = U_1 T_1 U_1^*$, which implies $U_1^* A_1 U_1 = T_1$. Define a new $n \times n$ matrix V_2 as a block diagonal matrix:

$$V_2 = \begin{bmatrix} 1 & 0 \\ 0 & U_1 \end{bmatrix}$$

Because U_1 is unitary, it is straightforward to verify that also V_2 is unitary ($V_2^*V_2 = I$). Let $T = V_2^*(V_1^*AV_1)V_2$ and compute this using block multiplication:

$$T = \begin{bmatrix} 1 & 0 \\ 0 & U_1^* \end{bmatrix} \begin{bmatrix} \lambda_1 & w^* \\ 0 & A_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & U_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & U_1^* \end{bmatrix} \begin{bmatrix} \lambda_1 & w^*U_1 \\ 0 & A_1U_1 \end{bmatrix} = \begin{bmatrix} \lambda_1 & w^*U_1 \\ 0 & U_1^*A_1U_1 \end{bmatrix} = \begin{bmatrix} \lambda_1 & w^*U_1 \\ 0 & T_1 \end{bmatrix}.$$

Since T_1 is upper-triangular the entire matrix T is upper-triangular.

To conclude let $U = V_1V_2$. Because the product of two unitary matrices is unitary, U is unitary.⁶ Rearranging our definition of T we obtain $T = (V_1V_2)^*A(V_1V_2) = U^*AU$, that is $A = UTU^*$, as was to be shown. $\square$

Corollary (Spectral Theorem for Real Symmetric Matrices). *A (real) symmetric matrix A can be decomposed as $S'AS = \Lambda$ with S orthogonal and Λ diagonal, where the elements of S are the eigenvectors of A and the diagonal elements of Λ their corresponding eigenvalues.*

Proof. The above Theorem states yields the factorization

$$A = STS^* \quad \text{or equivalently} \quad S^*AS = T$$

for any complex matrix A , where S is a unitary matrix ($S^*S = I$) and T is an upper-triangular matrix whose diagonal elements are the eigenvalues of A . Now consider real symmetric A .

First we claim that T must be diagonal. If A is real and symmetric, it is also a complex Hermitian matrix ($A^* = A$). Therefore

$$T^* = (S^*AS)^* = S^*A^*S = S^*AS = T.$$

Because $T^* = T$, the upper-triangular matrix T is equal to its own conjugate transpose. The only way an upper-triangular matrix can match its conjugate transpose is if all its off-diagonal entries are zero. Therefore T must be a diagonal matrix, which we call Λ .

Next claim: A 's eigenvalues are real. Because $T = \Lambda$ is equal to its own conjugate transpose ($\Lambda^* = \Lambda$), every entry on the diagonal must equal its own complex conjugate ($\lambda_j = \overline{\lambda_j}$). This proves that all eigenvalues of a real symmetric matrix are real.

Since A and its eigenvalues Λ are completely real, the linear systems used to calculate the eigenvectors contain only real coefficients. Hence we can choose a completely real set of eigenvectors. We can normalize and orthogonalize them using the Gram-Schmidt process. This allows us to replace the complex unitary matrix S with a purely real orthogonal matrix, switching the notation from S^* to the standard real transpose S' . Putting it all together, $S^*AS = T$ simplifies to $S'AS = \Lambda$. $\square$

⁶This is because for complex matrices M, N we have $(MN)^* = \overline{(MN)}' = \overline{N'M'} = \overline{N'}' \overline{M'} = N^*M^*$.

Appendix 2. Proof of Theorem 2 (Cholesky Decomposition).

Theorem (Cholesky Decomposition). *A symmetric positive definite $n \times n$ matrix A can be written as $A = LL'$ with L lower triangular with positive diagonal entries.*

Proof. Again by induction. The case $n = 1$ is easy: A must be a positive number a and we take L to be the number $\sqrt{a}$. Now assume the result true for dimension $n - 1$; we must show that it then holds for n . A symmetric positive definite $n \times n$ matrix A is of the form

$$A = \begin{bmatrix} \alpha & v' \\ v & B \end{bmatrix}$$

where α is a positive scalar ($\alpha = e_1' A e_1 > 0$ since A is positive definite), v is $(n - 1) \times 1$ and v' is there by symmetry, and B is an $(n - 1) \times (n - 1)$ symmetric matrix. We look for an $n \times n$ lower triangular matrix L with positive diagonal entries such that $LL' = A$. Such a matrix must have the form

$$L = \begin{bmatrix} \beta & 0_{1 \times (n-1)} \\ w & M \end{bmatrix}$$

where $\beta > 0$, w is $(n - 1) \times 1$ and M is a $(n - 1) \times (n - 1)$ lower triangular matrix with positive entries. Compute LL' :

$$LL' = \begin{bmatrix} \beta & 0 \\ w & M \end{bmatrix} \begin{bmatrix} \beta & w' \\ 0 & M' \end{bmatrix} = \begin{bmatrix} \beta^2 & \beta w' \\ \beta w & ww' + MM' \end{bmatrix}.$$

Imposing $LL' = A$ implies $\beta^2 = \alpha$ hence $\beta = \sqrt{\alpha}$; $\beta w = v$ hence $w = v/\sqrt{\alpha}$; and lastly, that it must be $ww' + MM' = B$, or $MM' = B - vv'/\alpha$: can this be met with M lower triangular with positive entries? since $B - vv'/\alpha$ is $(n - 1) \times (n - 1)$ the answer is yes - by the induction hypothesis - provided that $B - vv'/\alpha$ is positive definite (symmetry is clear). To verify this let $c \in \mathbb{R}^{n-1}$ arbitrary non-zero and define $x \in \mathbb{R}^n$ as

$$x = \begin{bmatrix} -\frac{1}{\alpha}v'c \\ c \end{bmatrix}.$$

Note that $c \neq 0 \Rightarrow x \neq 0$, and since A is positive definite $x'Ax > 0$. But

$$\begin{aligned} x'Ax &= \begin{bmatrix} -\frac{1}{\alpha}c'v & c' \end{bmatrix} \begin{bmatrix} \alpha & v' \\ v & B \end{bmatrix} \begin{bmatrix} -\frac{1}{\alpha}v'c \\ c \end{bmatrix} = \begin{bmatrix} -\frac{1}{\alpha}c'v & c' \end{bmatrix} \begin{bmatrix} -v'c + v'c \\ -\frac{1}{\alpha}vv'c + Bc \end{bmatrix} \\ &= \begin{bmatrix} -\frac{1}{\alpha}c'v & c' \end{bmatrix} \begin{bmatrix} 0 \\ (B - \frac{1}{\alpha}vv')c \end{bmatrix} = c' \left(B - \frac{1}{\alpha}vv' \right) c. \end{aligned}$$

The result now follows. □