REGRESSION-BASIC PREREQUISITE KNOWLEDGE

Statistics and Matrix Algebra Required for Regression

Reference: N. R. Draper and H. Smith, Applied Regression Analysis, 3rd Edition, Chapter 0

@ S. S. Roy | Date: 23 July 2026

1. Introduction

Regression analysis requires a fixed set of statistical and matrix tools. Fitting a line is only the first step; every subsequent question — whether a slope is significantly different from zero, how wide its confidence interval is, whether the model as a whole explains anything — is answered using the three distributions, the estimation–testing framework, and the matrix algebra summarised in this tutorial. Each concept is defined below together with its application in regression, before being treated in detail in its own section.

Concept	Definition	Application in regression
Degrees of freedom (df, ν)	The number of independent pieces of information remaining after parameters have been estimated from the data: df = n − (number of estimated parameters).	Determines which t-curve is used for tests and intervals; F requires two df values.
Normal distribution $N(\mu, \sigma^2)$	A symmetric bell-shaped distribution fully determined by its mean μ and standard deviation σ.	The random errors ε in the regression model are assumed normal; this assumption underlies all t and F procedures.
Gamma function Γ(q)	A generalisation of the factorial to non-integer arguments, defined by an integral and satisfying Γ(q) = (q−1)Γ(q−1).	Appears as the normalising constant inside the t and F density formulas; it is evaluated, never integrated, in practice.
t-distribution t(ν)	A family of bell-shaped distributions, one for each df value ν, with heavier tails than the normal; t(∞) = N(0,1).	Confidence intervals and significance tests for individual regression coefficients when σ is estimated by s.
F-distribution F(m, n)	The distribution of a ratio of two independent variance estimates, indexed by two df values; non-negative and right-skewed.	Testing equality of two variances; the overall significance test of a regression model (upper-tailed).
Confidence interval / t-test	Interval: estimate ± t × standard error. Test: t = (estimate − test value) / standard error.	Applied to the slope β_1, the intercept β_o, and predicted responses.
Matrix algebra	Rules for rectangular arrays: transpose, multiplication, inverse, determinant.	Multiple regression is written and solved entirely in matrix form, $Y = X\beta + \varepsilon$, $b = (X'X)^{-1}X'Y$.

2. Degrees of Freedom

Definition. Degrees of freedom (df, denoted ν) is the number of values in a calculation that are free to vary after certain quantities have been estimated from the same data. Each parameter estimated from the data removes one degree of freedom:

df = n − (number of parameters estimated from the data)

Example. Consider 5 numbers whose mean is known to be 10, so that their total must equal 50. Any 4 of the numbers can be chosen freely, but the 5th is then forced to whatever value makes the total 50. Five numbers with one estimated quantity (the mean) therefore carry 5 − 1 = 4 degrees of freedom.

Example in regression. A straight-line fit $Y = b_0 + b_1X$ estimates two parameters from the data. With n = 5 observations, the residuals carry n − 2 = 3 degrees of freedom, and the estimate s of σ is said to be based on 3 df.

- Interpretation: df measures the amount of independent information available for estimating the error variance σ^2. Larger df produces a more reliable s and a t-curve closer to the normal.
- The t-distribution carries one df value (ν). The F-distribution carries two — one for the variance estimate in the numerator, one for the denominator — because each estimate has its own df.

3. The Normal Distribution

Definition. The normal distribution is a symmetric, bell-shaped probability distribution determined completely by two quantities: the mean μ (centre) and the standard deviation σ (spread). Practically the entire distribution (99.73%) lies inside the range $\mu - 3\sigma \le x \le \mu + 3\sigma$. The frequency function (0.1.1) is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \qquad -\infty \le x \le \infty$$

Application. The normal distribution occurs frequently in the natural world, either for data as they come or for transformed data; the heights of a large randomly selected group of people are approximately normal. In regression, the random errors ε are assumed to be normally distributed — this assumption is what justifies the t and F procedures of later sections.

Notation. $x \sim N(\mu, \sigma^2)$ is read "x is normally distributed with mean μ and variance σ^2." Most manipulations are carried out with the standard (unit) normal N(0, 1), for which μ = 0 and σ = 1. A general normal variable x is converted to a standard normal variable z by (0.1.2):

$$z = \frac{x - \mu}{\sigma}$$

z expresses distance from the mean in units of standard deviations. The total area under every normal curve equals 1.

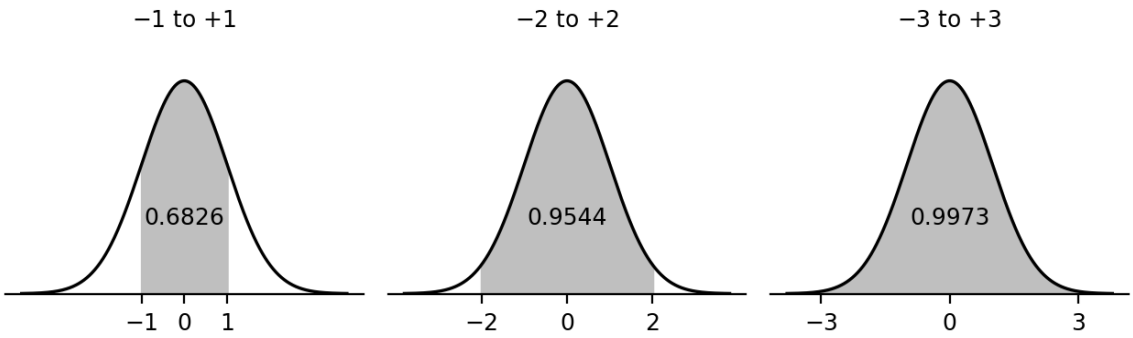

Figure 1. The standard normal N(0,1): area captured within ±1, ±2 and ±3 standard deviations.

Standard areas of the N(0, 1) distribution:

Range of z	Area inside	Area in each tail	Interpretation
−1 to +1	0.6826	0.1587	≈ 68% of the distribution
−2 to +2	0.9544	0.0228	≈ 95%
−3 to +3	0.9973	0.00135 *	“practically all”
−1.645 to +1.645	0.90	0.05	90% limits
−1.96 to +1.96	0.95	0.025	95% limits
−2.57 to +2.57	0.99	0.005	99% limits

** 0.00135 = (1 − 0.9973)/2. All values in this table are obtainable from a standard normal table.*

4. The Gamma Function

Definition. The gamma function Γ(q) is a generalisation of the factorial to non-integer arguments. Formally it is defined by the integral:

$$\Gamma(q) = \int_0^\infty e^{-x}\, x^{q-1}\, dx$$

Application. Γ(q) occurs inside the density formulas of the t-distribution (0.1.3) and the F-distribution (0.1.4), where it acts as the normalising constant that makes the total area under each curve equal to 1. In those applications the arguments are integers or half-integers, so the integral is never evaluated directly — the recursion below reduces every case to a simple product.

Working rule (generalised factorial property):

$$\Gamma(q) = (q-1)\,\Gamma(q-1) = (q-1)(q-2)\,\Gamma(q-2) = \cdots$$

Two starting values close the recursion:

$$\Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi} \qquad \text{and} \qquad \Gamma(1) = 1$$

Consequently, integer arguments give plain factorials — Γ(n) = (n − 1)! — and half-integer arguments give products ending in √π.

Example 1.

$$\Gamma(5) = 4 \times \Gamma(4) = 4 \times 3 \times \Gamma(3) = 4 \times 3 \times 2 \times 1 \times \Gamma(1) = 24$$

Example 2.

$$\Gamma\left(\tfrac{5}{2}\right) = \tfrac{3}{2} \times \Gamma\left(\tfrac{3}{2}\right) = \tfrac{3}{2} \times \tfrac{1}{2} \times \Gamma\left(\tfrac{1}{2}\right) = \frac{3\sqrt{\pi}}{4}$$

5. The t-Distribution

Definition. The t-distribution is a family of symmetric, bell-shaped distributions, one member for each value of ν, the degrees of freedom. Its density (0.1.3) is:

$$f_\nu(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}, \qquad -\infty \le t \le \infty$$

Application. The t-distribution is used whenever the standard deviation σ is unknown and is estimated by s from the data — the standard situation in regression. Confidence intervals and significance tests for individual regression coefficients (Sections 7–9) are all based on t(ν), where ν is the df of the estimate s.

Shape. A t(ν) curve resembles the standard normal but is heavier in the tails and correspondingly lower in the middle, since the total area must remain 1. The heavier tails reflect the additional uncertainty introduced by estimating σ rather than knowing it.

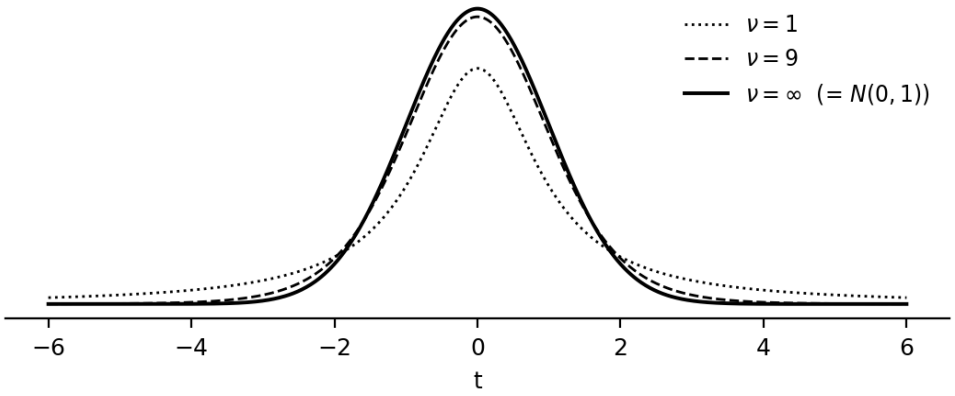

Figure 2. t-distributions for ν = 1, 9 and ∞. t(∞) is exactly N(0, 1).

As ν increases, the distribution becomes more normal; t(∞) is precisely the N(0, 1) distribution. Once ν exceeds about 30, the difference between t(ν) and N(0, 1) is so small that it is conventional — though not mandatory — to use N(0, 1) instead. Regression on small data sets typically has few degrees of freedom, so the t-distribution remains the working tool.

6. The F-Distribution

Definition. The F-distribution is the distribution of a ratio of two statistically independent variance estimates. It depends on two separate degrees of freedom, m (numerator) and n (denominator). Its density (0.1.4) is:

$$f_{m,n}(F) = \frac{\Gamma\left(\frac{m+n}{2}\right)\left(\frac{m}{n}\right)^{m/2}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{F^{m/2-1}}{(1 + mF/n)^{(m+n)/2}}, \qquad F \ge 0$$

Application. The F-distribution is used to compare two variances, and later to test the overall significance of a regression model. In the standard introduction, the null hypothesis of equal variances is tested against a two-sided alternative:

$$H_0 : \ \sigma_1^2/\sigma_2^2 = 1 \qquad \text{versus} \qquad H_1 : \ \sigma_1^2/\sigma_2^2 \neq 1$$

The test statistic is $F = s_1^2/s_2^2$, where s_1^2 and s_2^2 are statistically independent estimates of σ_1^2 and σ_2^2 with ν_1 and ν_2 degrees of freedom respectively. If the two samples are independent and normal, $(s_1^2/s_2^2)/(\sigma_1^2/\sigma_2^2)$ follows the $F(\nu_1, \nu_2)$ distribution; hence when $\sigma_1^2 = \sigma_2^2$ the plain ratio s_1^2/s_2^2 itself follows $F(\nu_1, \nu_2)$.

Shape. The distribution is non-negative (it is a ratio of squared quantities), rises from zero — sometimes quite steeply — to a peak, and falls away strongly skewed to the right.

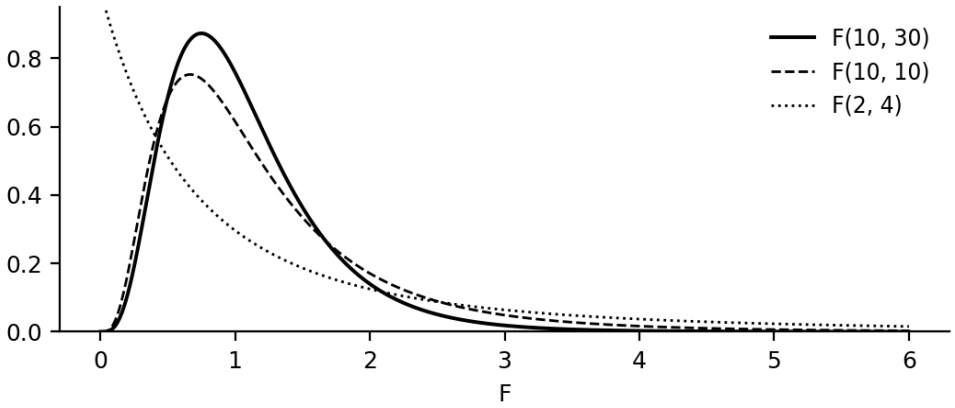

Figure 3. Selected F(m, n) distributions, showing the strong right skew.

One-tailed use in regression: in basic statistics the equal-variance test is two-tailed. In regression applications it is typically a one-tailed, upper-tailed test, because the variance estimate that could be too large — but cannot be too small — is placed in the numerator, and the estimate regarded as a reliable measure of the true σ^2 is placed in the denominator. The hypotheses are then H_0: $\sigma_1^2 = \sigma_2^2$ against H_1: $\sigma_1^2 > \sigma_2^2$, and F-tables list upper-tail percentage points at 10%, 5% and 1%.

The three distributions compared

Distribution	Parameters	Shape	Application in regression
Normal $N(\mu, \sigma^2)$	μ and σ	symmetric bell	distributional assumption on the errors ε
$t(\nu)$	one df, ν	bell, heavier tails	intervals and tests for a single coefficient
$F(m, n)$	two dfs, m and n	non-negative, right-skewed	variance comparison; overall model test

7. Confidence Intervals and t-Tests

Four symbols underlie the entire estimation–testing framework:

Symbol	Name	Meaning
θ	parameter (“thing”)	Any quantity to be estimated — a slope, an intercept, a mean response
$\hat{\theta}$	estimate (“estimate of thing”)	The estimate of θ computed from the data
$\sigma_{\hat{\theta}}$	standard deviation of $\hat{\theta}$	The true spread of the estimator; usually unknown, because it contains σ
$\mathrm{se}(\hat{\theta})$	standard error	The estimated standard deviation of $\hat{\theta}$, based on ν degrees of freedom

$\hat{\theta}$ typically follows a normal distribution — either exactly, when the underlying observations are themselves normal, or approximately, by the effect of the Central Limit Theorem. The standard error $\mathrm{se}(\hat{\theta})$ is obtained by substituting an estimate of the unknown standard deviation (based on ν degrees of freedom) into the formula for $\sigma_{\hat{\theta}}$; this is the reason s replaces σ throughout regression formulas, and the reason a df value accompanies every standard error.

(a) Confidence interval

A 100(1 − α)% confidence interval for θ is given by (0.2.1):

$$\hat{\theta} \pm t(\nu, 1-\alpha/2)\ \mathrm{se}(\hat{\theta})$$

where t(ν, 1 − α/2) is the percentage point of a t-variable with ν degrees of freedom leaving probability α/2 in the upper tail. Equation (0.2.1) in words (0.2.2):

$$\left\{\begin{matrix}\text{Estimate}\\ \text{of thing}\end{matrix}\right\} \pm \left\{\begin{matrix}t \text{ point leaving } \alpha/2 \text{ in the}\\ \text{upper tail, based on } \nu \text{ df}\end{matrix}\right\} \times \left\{\begin{matrix}\text{Standard error}\\ \text{of estimate}\end{matrix}\right\}$$

(b) t-Test

To test $\theta = \theta_o$, where θ_o is a specified value presumed valid ($\theta_o = 0$ in most tests of regression coefficients), the statistic (0.2.3) is evaluated:

$$t = \frac{\hat{\theta} - \theta_0}{\mathrm{se}(\hat{\theta})}$$

or, in words (0.2.4):

$$t = \frac{\{\text{Estimate of thing}\} - \{\text{Test value of thing}\}}{\{\text{Standard error of estimate of thing}\}}$$

The observed t is placed on the t(ν) curve; the tail probability beyond it, δ, is evaluated and doubled for a two-tailed test. The doubled area 2δ is the p-value: the observed value could just as well have come out negative, and the mirror-image point supplies the second δ.

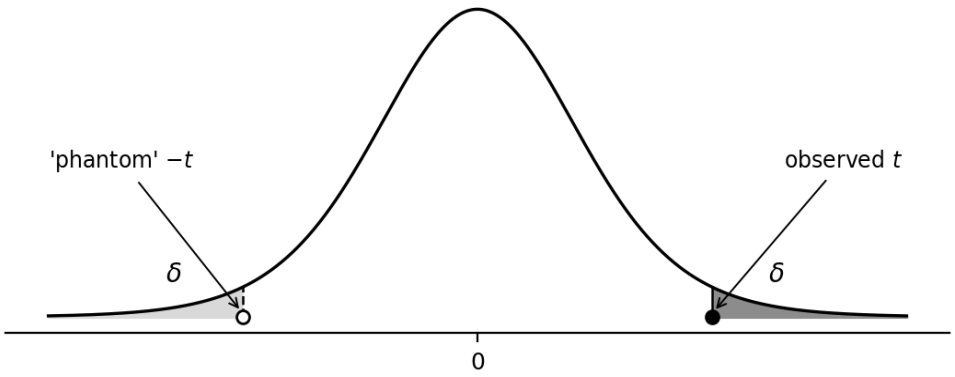

Figure 4. A two-tailed t-test: the observed t (solid dot) has tail area δ; the mirror-image "phantom" point supplies the second δ, giving a two-tailed probability of 2δ.

Decision convention: if $2\delta < 0.05$, t is termed significant and the hypothesis $\theta = \theta_o$ is rejected; if $2\delta > 0.05$, t is non-significant and the hypothesis is not rejected. The alternative hypothesis is $\theta \neq \theta_o$, a two-sided alternative.

On the significance level α: the value 0.05 is a convention, not an absolute standard. Using $\alpha = 0.05$ means accepting a 1-in-20 risk of a wrong decision; $\alpha = 0.10$ (1 in 10) or $\alpha = 0.01$ (1 in 100) are equally legitimate choices, provided the chosen level is applied consistently throughout the testing. A result at $2\delta = 0.049$ versus one at 0.051 differs only by an arbitrary boundary; promising experimental leads deserve follow-up even when the arbitrary standard has not quite been attained. The α value is a guidepost, not a boundary.

(c) Applications of formulas (0.2.1)–(0.2.4) — Table 0.1

Every application of the four formulas requires identifying θ, $\hat{\theta}$, θ_o, $se(\hat{\theta})$ and the t percentage point; the formulas themselves never change. The cases arising in the straight-line model $Y = \beta_o + \beta_1 X + \varepsilon$ are:

Situation	θ	$\hat{\theta}$	$se(\hat{\theta})$
Slope of the line	β_1	b_1	$s / S\_XX^{\wedge}½$
Intercept of the line	β_o	b_o	$s \{ \Sigma X_i^2 / (n \cdot S\_XX) \}^{\wedge}½$
Mean response at $X = X_o$	$E(Y)$ at X_o	$\hat{Y}_o = b_o + b_1 X_o$	$s \{ 1/n + (X_o - \bar{X})^2 / S\_XX \}^{\wedge}½$

where $S\_XX = \Sigma(X_i - \bar{X})^2$ and s is the estimate of σ obtained from the fit; the symbol s replaces σ of the corresponding standard-deviation formulas. All three standard errors are s multiplied by a quantity built from the X-values.

8. Elements of Matrix Algebra

Definition. Matrix algebra is the arithmetic of rectangular arrays of numbers. **Application.** Multiple regression with several predictors is formulated and solved entirely in matrix form; the transpose, product, inverse and determinant defined below are precisely the operations used in the estimation formula $b = (X'X)^{-1}X'Y$ and in the construction of confidence regions.

(a) Vocabulary

Term	Meaning	Example
Matrix	A rectangular array with p rows and q columns (a p × q matrix)	A = [[4, 1, 3, 7], [−1, 0, 2, 2], [6, 5, −2, 1]] is 3 × 4
Row vector	A matrix with only one row	a′ = [1, 6, 3, 2, 1], length five
Column vector	A matrix with only one column	b = (−1, 0, 1)′, length three
Scalar	A 1 × 1 "vector" — an ordinary number	7

Convention (not universal): capital letters denote matrices, lower-case letters denote vectors, often in boldface; brackets may be square-ended or curved. The plural of matrix is matrices.

(b) Basic operations

Operation	Rule	Condition
Equality	Identical entry in every position	Dimensions must be identical
Sum / Difference	Element-by-element addition or subtraction	Dimensions must be identical, otherwise undefined
Transpose (M′)	Rows of M become the columns of M′, in the same order	Always defined
Symmetry	M is symmetric if M′ = M	M must be square
Multiplication (AB)	$c_{ij} = \Sigma\, a_{il}\ b_{l\ j}$ — the inner product of row i of A with column j of B	A is p × q, B must be q × s (conformable); result is p × s

Multiplication example — a 2 × 3 matrix times a 3 × 3 matrix gives a 2 × 3 result:

$$\begin{bmatrix} 1 & 2 & 1 \\ -1 & 3 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 0 & -1 \\ -2 & 1 & 3 \end{bmatrix} = \begin{bmatrix} 7 & 3 & 4 \\ 11 & -2 & -6 \end{bmatrix}$$

Verification of the top-left entry: 1(1) + 2(4) + 1(−2) = 7. Every other entry is the corresponding row-by-column inner product.

Two properties of matrix multiplication: (1) AB and BA, even when both are conformable, do not in general give the same result — the order of matrices is crucial, unlike the order of numbers in a scalar product. In AB, B is premultiplied by A, equivalently A is postmultiplied by B. (2) When several matrices and vectors are multiplied together, the bracketing that leads to the least work should be chosen: for W(p×p) Z′(p×n) y(n×1), computing W(Z′y) requires far fewer cross-products than (WZ′)y.

(c) Special matrices and vectors

Symbol	Name	Content	Role

Symbol	Name	Content	Role
I_n	Identity (unit) matrix	n × n, 1s on the diagonal, 0s elsewhere	The matrix counterpart of the number 1; the subscript n is omitted when the size is clear
0	Zero vector or matrix	Every entry zero	The counterpart of the number 0; size clear from context
1	Vector of ones	(1, 1, …, 1)′	1′1 equals the squared length of 1; 11′ is a square matrix of all 1s

Orthogonality. A vector a = (a_1, …, a_n)′ is orthogonal to a vector b = (b_1, …, b_n)′ if the sum of the products of their elements is zero:

$$\sum_{i=1}^{n} a_i b_i = \mathbf{a}'\mathbf{b} = \mathbf{b}'\mathbf{a} = 0$$

(d) Inverse matrix

Definition. The inverse M^{-1} of a square matrix M is the unique matrix satisfying:

$$\mathbf{M}^{-1}\mathbf{M} = \mathbf{I} = \mathbf{M}\mathbf{M}^{-1}$$

Existence depends on **linear dependence**: the columns m_1, m_2, …, m_n of an n × n matrix are linearly dependent if there exist constants c_1, c_2, …, c_n, not all zero, such that

$$c_1\mathbf{m}_1 + c_2\mathbf{m}_2 + \cdots + c_n\mathbf{m}_n = \mathbf{0}$$

and similarly for rows. Then:

- A square matrix with linearly dependent rows or columns is **singular** and possesses no inverse. A square matrix that is not singular is **nonsingular** and can be inverted.
- If M is symmetric, so is M^{-1}. (In regression, the matrix to be inverted, X′X, is always symmetric.)

Obtaining an inverse

The inversion procedure is defined through an example. To invert M = [3 4 5 ; 1 2 6 ; 7 1 9], M^{-1} is written as a matrix of nine unknowns (a, b, c, …, h, k) subject to $M^{-1}M = I$:

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & k \end{bmatrix} \begin{bmatrix} 3 & 4 & 5 \\ 1 & 2 & 6 \\ 7 & 1 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Multiplying out and matching every entry against I produces three sets of three linear simultaneous equations — one set per column of unknowns:

$$\begin{array}{lll} 3a + b + 7c = 1, & 3d + e + 7f = 0, & 3g + h + 7k = 0, \\ 4a + 2b + c = 0, & 4d + 2e + f = 1, & 4g + 2h + k = 0, \\ 5a + 6b + 9c = 0, & 5d + 6e + 9f = 0, & 5g + 6h + 9k = 1. \end{array}$$

Solving the nine equations yields the inverse (with the common factor 1/103 removed, as described in Section (f)):

$$\mathbf{M} = \begin{bmatrix} 3 & 4 & 5 \\ 1 & 2 & 6 \\ 7 & 1 & 9 \end{bmatrix}, \qquad \mathbf{M}^{-1} = \frac{1}{103} \begin{bmatrix} 12 & -31 & 14 \\ 33 & -8 & -13 \\ -13 & 25 & 2 \end{bmatrix}$$

For an n × n matrix there are, in general, n sets of n simultaneous linear equations. Accelerated methods adapted for computers obtain inverses with great speed even for large matrices; hand inversion is obsolete except in simple cases such as the 2 × 2 formula:

$$\mathbf{M}^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \begin{bmatrix} d/D & -b/D \\ -c/D & a/D \end{bmatrix}, \qquad D = ad - bc$$

Determinant route to the inverse (verifiable against the 2 × 2 formula above): replace each element m_{ij} of M by (1) the determinant of the submatrix obtained by crossing out the row and column containing m_{ij}, (2) with the sign attached from the + − + − count, (3) divided by det M. When every element has been replaced, **the resulting matrix is transposed**; that transpose is M^{-1}.

(e) Determinants

Definition. The determinant is a single number associated with a square matrix. Determinants occur naturally in the solution of linear simultaneous equations and in the inversion of matrices. For a 2 × 2 matrix:

$$\det \mathbf{M} = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

For a 3 × 3 matrix [a b c ; d e f ; g h k], expanding by the first row — each first-row element multiplied by the determinant of the submatrix left after crossing out its row and column, with alternating signs + − + — gives:

$$a \begin{vmatrix} e & f \\ h & k \end{vmatrix} - b \begin{vmatrix} d & f \\ g & k \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix} = aek - afh - bdk + bfg + cdh - ceg$$

The determinant can be written as an expansion of any row or column by the same technique. The signs attached are counted + − + − from the top left-hand corner element, alternating along a row or column (not diagonally):

$$\begin{bmatrix} + & - & + \\ - & + & - \\ + & - & + \end{bmatrix}$$

Geometrically, the determinant measures the volume of the parallelepiped defined by the vectors in the rows (or columns) of the matrix; this is the reason det(X′X) later determines the size of a joint confidence region for regression parameters.

(f) Common factors

If every element of a matrix has a common factor, it can be taken outside the matrix; conversely, multiplying a matrix by a constant c multiplies every element by c:

$$\begin{bmatrix} 4 & 6 & -2 \\ 8 & 6 & 2 \end{bmatrix} = 2 \begin{bmatrix} 2 & 3 & -1 \\ 4 & 3 & 1 \end{bmatrix}$$

Determinant of a scaled matrix: for a square p × p matrix with common factor c, the determinant carries the factor $\mathbf{c^p}$, not c:

$$\begin{vmatrix} 4 & 6 \\ 8 & 6 \end{vmatrix} = 2^2 \begin{vmatrix} 2 & 3 \\ 4 & 3 \end{vmatrix} = 4\,(6 - 12) = -24$$

9. Solved Example 1 — Confidence Interval and t-Test on a Fitted Line

Problem. Five observations of a response Y are taken at five settings of a predictor X. A straight line $Y = b_0 + b_1X$ is fitted. Required: (a) S_{XX}, (b) the fitted line, (c) the estimate s, (d) a 95% confidence interval for the slope β_1, (e) a test of H_0: $\beta_1 = 0$.

X	0	1	2	3	4
Y	3	6	8	11	12

Solution

Step 1 — means. $\bar{X} = (0+1+2+3+4)/5 = 2$, $\bar{Y} = (3+6+8+11+12)/5 = 40/5 = 8$.

Step 2 — sums of squares and products (deviations from the means):

$X_i - \bar{X}$	−2	−1	0	1	2	Sum
$Y_i - \bar{Y}$	−5	−2	0	3	4	—
$(X_i - \bar{X})^2$	4	1	0	1	4	$S_{XX} = 10$
$(X_i - \bar{X})(Y_i - \bar{Y})$	10	2	0	3	8	$S_{XY} = 23$

Step 3 — fitted line. $b_1 = S_{XY} / S_{XX} = 23/10 = 2.3$, and $b_0 = \bar{Y} - b_1\bar{X} = 8 - 2.3(2) = 3.4$:

$$\hat{Y} = 3.4 + 2.3\,X$$

Step 4 — estimate s of σ. Fitted values and residuals:

X	0	1	2	3	4
$\hat{Y} = 3.4 + 2.3X$	3.4	5.7	8.0	10.3	12.6
Residual $(Y - \hat{Y})$	−0.4	0.3	0.0	0.7	−0.6
Residual²	0.16	0.09	0.00	0.49	0.36

The sum of squared residuals is 1.10 on $n - 2 = 3$ degrees of freedom (two df are removed by the estimation of b_0 and b_1; compare Section 2). Therefore:

$$s^2 = \frac{1.10}{3} = 0.3667, \qquad s = 0.6055 \qquad (\nu = 3 \text{ df})$$

Step 5 — standard error of the slope. From Table 0.1:

$$\operatorname{se}(b_1) = \frac{s}{\sqrt{S_{XX}}} = \frac{0.6055}{\sqrt{10}} = 0.1915$$

Step 6 — 95% confidence interval, formula (0.2.1). With α = 0.05 and ν = 3, the t-table gives t(3, 0.975) = 3.182:

$$2.3 \;\pm\; 3.182 \times 0.1915 \;=\; 2.3 \;\pm\; 0.609 \;=\; (1.69,\; 2.91)$$

Interpretation: with 95% confidence, the true slope lies between 1.69 and 2.91. The interval is wide because at only 3 degrees of freedom the t-multiplier 3.182 is far larger than the large-sample value 1.96.

Step 7 — t-test of H_o: $\beta_1 = 0$, formula (0.2.3). Here $\hat{\theta} = 2.3$ and $\theta_o = 0$:

$$t = \frac{2.3 - 0}{0.1915} = 12.01 \qquad \text{on 3 df}$$

Since |12.01| exceeds the critical value 3.182, the result is significant at the 5% level (exact two-tailed probability 2δ ≈ 0.0012). H_o is rejected: the slope is non-zero, and X contributes genuinely to the explanation of Y.

Relationship between interval and test: the confidence interval (1.69, 2.91) excludes 0, and the t-test rejects $\beta_1 = 0$. The two procedures always agree in this way — a two-sided t-test at level α rejects $\theta = \theta_o$ exactly when the 100(1−α)% confidence interval excludes θ_o.

10. Solved Example 2 — The Same Fit in Matrix Form

Problem. For the three points X = 1, 2, 3 with Y = 2, 4, 5: set up the matrices X and Y; compute X′X, its determinant and inverse; obtain the coefficient vector $b = (X'X)^{-1}X'Y$; and verify the result against the ordinary least-squares formulas.

Solution

Step 1 — matrices. The column of 1s carries the intercept:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \;(3 \times 2), \qquad \mathbf{Y} = \begin{bmatrix} 2 \\ 4 \\ 5 \end{bmatrix} \;(3 \times 1)$$

Step 2 — transpose and product. X′ is 2 × 3, so X′X is conformable and yields a 2 × 2 result:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}$$

Entry by entry: top-left = 1+1+1 = 3 (which is n); top-right = 1+2+3 = 6 (ΣX); bottom-right = 1+4+9 = 14 (ΣX^2). X′X is symmetric, in agreement with Section 8(d).

Step 3 — determinant. det(X′X) = (3)(14) − (6)(6) = 42 − 36 = 6. The determinant is non-zero, so the matrix is nonsingular and invertible.

Step 4 — inverse, by the 2 × 2 formula. Diagonal entries swapped, off-diagonal entries negated, all divided by the determinant:

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{6} \begin{bmatrix} 14 & -6 \\ -6 & 3 \end{bmatrix} = \begin{bmatrix} 7/3 & -1 \\ -1 & 1/2 \end{bmatrix}$$

Verification — the product with the original matrix must return I:

$$\frac{1}{6}\begin{bmatrix}14 & -6\\ -6 & 3\end{bmatrix}\begin{bmatrix}3 & 6\\ 6 & 14\end{bmatrix} = \frac{1}{6}\begin{bmatrix}6 & 0\\ 0 & 6\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \checkmark$$

Step 5 — X′Y.

$$\mathbf{X'Y} = \begin{bmatrix}1 & 1 & 1\\ 1 & 2 & 3\end{bmatrix}\begin{bmatrix}2\\ 4\\ 5\end{bmatrix} = \begin{bmatrix}11\\ 25\end{bmatrix}$$

Step 6 — coefficients.

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y} = \begin{bmatrix}7/3 & -1\\ -1 & 1/2\end{bmatrix}\begin{bmatrix}11\\ 25\end{bmatrix} = \begin{bmatrix}77/3 - 25\\ -11 + 12.5\end{bmatrix} = \begin{bmatrix}0.667\\ 1.5\end{bmatrix}$$

Hence $b_0 = 0.667$ and $b_1 = 1.5$, giving $\hat{Y} = 0.667 + 1.5X$.

Step 7 — cross-check with the ordinary formulas. $\bar{X} = 2$, $\bar{Y} = 11/3$. $S_{XX} = 1 + 0 + 1 = 2$ and $S_{XY} = (-1)(-5/3) + 0 + (1)(4/3) = 3$. Then $b_1 = 3/2 = 1.5$ and $b_0 = 11/3 - 1.5(2) = 0.667$. The matrix route and the scalar route agree exactly.

Purpose of the matrix formulation: the formula $b = (X'X)^{-1}X'Y$ belongs to a later chapter; it is included here to show where the transpose, conformability, symmetry, determinant and inverse of Section 8 are applied. With one predictor the matrix route duplicates the scalar formulas; with many predictors it is the only practical method, and every step remains identical — only the matrix dimensions grow.

11. One-Page Recap

Idea	Summary
Degrees of freedom	df = n − (parameters estimated); measures independent information; t carries one df, F carries two
Normal	Symmetric; z = (x − μ)/σ; ±1σ → 68%, ±2σ → 95%, ±3σ → 99.7%; ±1.96 → exactly 95%; assumption on the errors ε
Gamma	Γ(q) = (q−1)Γ(q−1), with Γ(1) = 1 and Γ(½) = √π; the normalising constant inside t and F densities
t-distribution	One df ν; heavier-tailed than the normal; t(∞) = N(0,1); used whenever σ is estimated by s
F-distribution	Two dfs; non-negative, right-skewed; ratio of two variance estimates; upper-tailed in regression
Confidence interval	estimate ± t(ν, 1−α/2) × standard error
t-test	t = (estimate − test value) / standard error; doubled tail area 2δ is the p-value; compare with α
Significance level	α = 0.05 is a convention — a guidepost, not a boundary
Matrix multiplication	A(p×q) B(q×s) defined only when inner dimensions match; result p×s; AB ≠ BA in general

Idea	Summary
Transpose / symmetry	M′ interchanges rows and columns; M symmetric if M′ = M; X′X is always symmetric
Inverse	$M^{-1}M = I$; exists only for square nonsingular matrices (det ≠ 0); found from n sets of n simultaneous equations, or by the determinant route
Determinant	2×2: ad − bc; expansion along any row or column with + − + signs; common factor c emerges as c^p

Next step: with these tools in place, the straight-line model $Y = \beta_o + \beta_1 X + \varepsilon$ can be developed — least-squares fitting, t-tests on the coefficients, an F-test on the whole model, and the generalisation to many predictors in the matrix form $Y = X\beta + \varepsilon$.

Exercise Problem.
The breaking strength of steel rods is normally distributed. A sample of **n = 8** rods gives a sample mean $\bar{X}$ = 52 units and a sample standard deviation s = 4 units. For the process as a whole, the true values are known historically to be $\mu = 50$ and $\sigma = 4$.
(a) State the degrees of freedom associated with s.
(b) Using the historical values, find the z-score of a rod measuring 58 units, and the proportion of rods measuring between 42 and 58 units.
(c) Evaluate Γ(4) and Γ(7/2).
(d) Using the sample results, construct a 95% confidence interval for the true mean μ.
Solution
(a) Degrees of freedom. One parameter (the mean) is estimated from the data, so one degree of freedom is lost:
df = n − 1 = 8 − 1 = **7**
(b) Normal distribution. With $\mu = 50$ and $\sigma = 4$, apply $z = (x - \mu)/\sigma$:
z at x = 58 → (58 − 50)/4 = **+2**
z at x = 42 → (42 − 50)/4 = **−2**
The range 42 to 58 is exactly ±2σ about the mean. From the standard normal table, the area between z = −2 and z = +2 is **0.9544**, so approximately **95.44%** of rods lie between 42 and 58 units.
(c) Gamma function. Apply $\Gamma(q) = (q - 1)\Gamma(q - 1)$ with $\Gamma(1) = 1$ and $\Gamma(\tfrac{1}{2}) = \sqrt{\pi}$.
$\Gamma(4) = 3 \times \Gamma(3) = 3 \times 2 \times \Gamma(2) = 3 \times 2 \times 1 \times \Gamma(1) =$ **6** (= 3!)
$\Gamma(7/2) = (5/2) \times (3/2) \times (1/2) \times \Gamma(1/2) =$ **$15\sqrt{\pi}/8$** ≈ 3.3234
(d) t-distribution — 95% confidence interval. Since σ is estimated by s, the t-distribution applies with ν = 7 df. The standard error is:
$se(\bar{X}) = s/\sqrt{n} = 4/\sqrt{8} = 1.4142$
From the t-table, t(7, 0.975) = 2.365. Applying the interval formula:
$52 \pm 2.365 \times 1.4142 = 52 \pm 3.34 =$ **(48.66, 55.34)**
Note the effect of small df: the multiplier 2.365 is appreciably larger than the large-sample normal value of 1.96, which widens the interval.

Practice Exercises (keys only)
Exercise 1. A straight line $Y = \beta_o + \beta_1 X$ is fitted to n = 12 observations. (a) State the df of the estimate s. (b) Evaluate Γ(6). (c) State the t multiplier for a 95% confidence interval on β_1. (d) Compare it with the corresponding normal value.

Key: (a) 12 − 2 = **10 df** (two parameters estimated). (b) $\Gamma(6) = 5! = \mathbf{120}$. (c) $t(10, 0.975) = \mathbf{2.228}$. (d) Larger than 1.96; the gap narrows as df increases.

Exercise 2. Examination marks follow $N(60, 8^2)$. (a) Find the z-score of a student scoring 76. (b) Find the percentage of students scoring between 44 and 76. (c) Find the mark exceeded by only 2.5% of students. (d) Evaluate $\Gamma(9/2)$.

Key: (a) $z = (76 - 60)/8 = \mathbf{+2}$. (b) 44 gives $z = -2$, so the range is $\pm 2\sigma \rightarrow \mathbf{95.44\%}$. (c) Upper 2.5% corresponds to $z = 1.96$, so mark = 60 + 1.96(8) = **75.68**. (d) $\Gamma(9/2) = (7/2)(5/2)(3/2)(1/2)\sqrt{\pi} = \mathbf{105\sqrt{\pi}/16} \approx 11.6317$.

Exercise 3. A sample of n = 25 observations gives $\bar{X} = 100$ and $s = 10$. (a) State the df. (b) Compute $se(\bar{X})$. (c) Construct a 95% confidence interval for μ, given $t(24, 0.975) = 2.064$. (d) State which distribution and which multiplier would apply if $\sigma = 10$ were known instead of estimated.

Key: (a) 25 − 1 = **24 df**. (b) $se(\bar{X}) = 10/\sqrt{25} = \mathbf{2}$. (c) $100 \pm 2.064(2) = 100 \pm 4.13 = \mathbf{(95.87, 104.13)}$. (d) The standard normal N(0, 1), multiplier **1.96** — giving the narrower interval (96.08, 103.92).