Mathematical Foundations and Detailed Derivations for the Analysis Pipeline

Automated Exposition

December 4, 2025

Abstract

This document integrates and exhaustively explains every mathematical and analytical component used in the provided analysis code. The exposition covers: the Pearson correlation and permutation test; ordinary least squares and residual sums of squares; the Bayesian Information Criterion (BIC) and its Laplace approximation to Bayes factors; the Savage–Dickey density ratio and practical estimation of posterior density at a point using Gaussian kernel density estimation (KDE); Bayesian linear regression modeling (centered and non-centered parameterizations); Student–t likelihood for robust regression; Markov chain Monte Carlo (MCMC) diagnostics including divergences, Gelman–Rubin $\hat{R}$, and effective sample size; posterior predictive checks (PPC); leave-one-out cross-validation (LOO) and Pareto-smoothed importance sampling (PSIS); and the conservative decision heuristics used to combine evidence. Each formula is derived or motivated and practical numerical considerations are discussed.

Contents

1 Notation and basic setup **4**

2 Pearson correlation: definition and algebraic properties **4**
- 2.1 Definition . . . 4
- 2.2 Relation to linear regression . . . 4
- 2.3 Properties . . . 5

3 Permutation test for Pearson correlation **5**
- 3.1 Null hypothesis and permutation principle . . . 5
- 3.2 Test statistic . . . 5
- 3.3 Permutation distribution . . . 5
- 3.4 Two-sided p-value with small-sample correction . . . 5
- 3.5 Remarks on interpretation . . . 5

4 Ordinary least squares, residual sum of squares, and R^2 **6**
- 4.1 OLS estimators . . . 6
- 4.2 Residual sum of squares (RSS) . . . 6
- 4.3 Coefficient of determination R^2 . . . 6

5 Bayesian Information Criterion (BIC) and Laplace approximation to Bayes factors **6**
5.1 BIC definition . . . 6
5.2 Laplace approximation and Bayes factor . . . 7
5.3 Interpretation of BF values . . . 7

6 Savage–Dickey density ratio: exact identity and practical estimation **8**
6.1 Savage–Dickey identity . . . 8
6.2 Prior density at zero for a normal prior . . . 8
6.3 Posterior density at zero: estimation from samples . . . 8
6.4 Bandwidth selection and practical issues . . . 8
6.5 Interpretation and limitations . . . 9

7 Bayesian linear regression: model specification and parameterizations **9**
7.1 Standard (centered) parameterization . . . 9
7.2 Non-centered parameterization . . . 9
7.3 Why non-centered helps . . . 9

8 Student–t likelihood for robust regression **9**
8.1 Student–t density . . . 9
8.2 Robustness properties . . . 10
8.3 Priors for ν . . . 10

9 MCMC sampling diagnostics **10**
9.1 Divergences in Hamiltonian Monte Carlo (NUTS) . . . 10
9.2 Gelman–Rubin $\hat{R}$ statistic . . . 10
9.3 Effective sample size (ESS) . . . 11

10 Posterior predictive checks (PPC) **11**
10.1 Posterior predictive distribution . . . 11
10.2 Generating replicated datasets . . . 11
10.3 Summary statistics and discrepancy measures . . . 12

11 Leave-one-out cross-validation (LOO) and PSIS **12**
11.1 LOO predictive density . . . 12
11.2 Importance sampling identity . . . 12
11.3 Pareto-smoothed importance sampling (PSIS) . . . 12
11.4 LOO model comparison . . . 12

12 Highest density interval (HDI) and credible intervals **13**
12.1 Definition . . . 13
12.2 HDI exclusion of zero . . . 13

13 Combining evidence: heuristics used in the pipeline **13**

14 Numerical and practical considerations **14**
14.1 Finite precision and floors . . . 14
14.2 Monte Carlo variability . . . 14
14.3 Standardization of predictors . . . 14

15 Worked symbolic derivations and small proofs **14**
15.1 Derivation: relation between r and OLS slope 14
15.2 Derivation: BIC for Gaussian linear model 14
15.3 Derivation: Savage–Dickey identity (sketch) 15

16 Summary of recommended practical workflow (algorithmic steps) **15**

17 Appendix: useful formulae and constants **16**

1 Notation and basic setup

Let $\{(x_i, y_i)\}_{i=1}^n$ denote paired observations. Vectors $x = (x_1, \ldots, x_n)^\top$ and $y = (y_1, \ldots, y_n)^\top$ are column vectors. We use the following standard sample statistics:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i.$$

Define centered variables

$$\tilde{x}_i = x_i - \bar{x}, \qquad \tilde{y}_i = y_i - \bar{y}.$$

Define sums of squares and cross-products

$$S_{xx} = \sum_{i=1}^n \tilde{x}_i^2, \quad S_{yy} = \sum_{i=1}^n \tilde{y}_i^2, \quad S_{xy} = \sum_{i=1}^n \tilde{x}_i \tilde{y}_i.$$

We denote by $\|\cdot\|$ the Euclidean norm and by $\mathbf{1}\{\cdot\}$ the indicator function.

2 Pearson correlation: definition and algebraic properties

2.1 Definition

The sample Pearson correlation coefficient r between x and y is

$$\boxed{r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.}$$

This is the sample estimate of the population Pearson correlation ρ.

2.2 Relation to linear regression

Consider the simple linear regression model with intercept:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

The ordinary least squares (OLS) slope estimator (when regressing y on x with intercept) is

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}.$$

The sample correlation relates to $\hat{\beta}$ via

$$\hat{\beta} = r \cdot \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}}.$$

Equivalently,

$$r = \hat{\beta} \cdot \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}}.$$

2.3 Properties

- $r \in [-1, 1]$.
- $r = 1$ or -1 indicates perfect linear relationship.
- r is invariant to separate location shifts of x and y and to separate positive scaling.

3 Permutation test for Pearson correlation

3.1 Null hypothesis and permutation principle

The permutation test assesses the null hypothesis H_0: "no association between x and y" (exchangeability of y relative to x). Under H_0, the joint distribution of (x, y) factorizes so that the labels of y can be permuted without changing the distribution of the data.

3.2 Test statistic

Use the observed Pearson correlation r_{obs} as the test statistic:

$$r_{\text{obs}} = r(x, y).$$

3.3 Permutation distribution

Generate N_{perm} random permutations π_j of indices $\{1, \ldots, n\}$. For each permutation compute

$$r^{(j)} = r\big(x,\, y_{\pi_j}\big).$$

This yields an empirical null distribution of the statistic under H_0.

3.4 Two-sided p-value with small-sample correction

Count exceedances:

$$C = \sum_{j=1}^{N_{\text{perm}}} \mathbf{1}\big\{|r^{(j)}| \geq |r_{\text{obs}}|\big\}.$$

The permutation p-value with the standard Monte Carlo correction is

$$\boxed{p = \frac{C + 1}{N_{\text{perm}} + 1}.}$$

This correction ensures that p is never zero and yields an unbiased estimator of the permutation p-value in the Monte Carlo sense.

3.5 Remarks on interpretation

- The permutation test is nonparametric and exact (conditional on the observed x) if all permutations are enumerated; with Monte Carlo sampling it is approximate.
- The test assesses association but not causation.

- The test is sensitive to monotone linear association as measured by Pearson r; it is not optimal for nonlinear associations.

4 Ordinary least squares, residual sum of squares, and R^2

4.1 OLS estimators

In matrix form, for a model with intercept and single predictor x, define design matrix

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \in \mathbb{R}^{n \times 2}.$$

The OLS estimator is

$$\hat{\beta}_{\text{vec}} = (X^\top X)^{-1} X^\top y = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}.$$

4.2 Residual sum of squares (RSS)

Predicted values $\hat{y} = X\hat{\beta}_{\text{vec}}$. Residuals $e = y - \hat{y}$. Residual sum of squares:

$$\text{RSS} = \sum_{i=1}^{n} e_i^2 = \|y - X\hat{\beta}_{\text{vec}}\|^2.$$

4.3 Coefficient of determination R^2

Total sum of squares:

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = S_{yy}.$$

Explained sum of squares $\text{ESS} = \text{TSS} - \text{RSS}$. Then

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

For simple linear regression, $R^2 = r^2$.

5 Bayesian Information Criterion (BIC) and Laplace approximation to Bayes factors

5.1 BIC definition

For a model M with k free parameters (counting intercepts and variance parameters as appropriate) and sample size n, the Bayesian Information Criterion (Schwarz, 1978) is

$$\boxed{\text{BIC} = -2 \log \hat{L} + k \log n}$$

where $\hat{L}$ is the maximized likelihood. For Gaussian linear models with residual sum of squares RSS, the log-likelihood (up to additive constants not depending on parameters) is

$$\log L \propto -\frac{n}{2} \log\left(\frac{\mathrm{RSS}}{n}\right).$$

Thus a commonly used expression for BIC in the Gaussian regression context is

$$\boxed{\mathrm{BIC} = n \log\left(\frac{\mathrm{RSS}}{n}\right) + k \log n.}$$

This is the form used in the code.

5.2 Laplace approximation and Bayes factor

The marginal likelihood (evidence) for model M is

$$p(y \mid M) = \int p(y \mid \theta, M) p(\theta \mid M)\, d\theta.$$

Laplace's method approximates this integral by expanding the log posterior around its mode and approximating by a Gaussian integral. The BIC arises as an asymptotic approximation to $-2 \log p(y \mid M)$ up to constants. For two models M_0 (null) and M_1 (alternative), the Bayes factor in favor of M_0 over M_1 is

$$\mathrm{BF}_{01} = \frac{p(y \mid M_0)}{p(y \mid M_1)}.$$

Using the BIC approximation,

$$-2 \log p(y \mid M) \approx \mathrm{BIC} + \text{constant},$$

so

$$\log \mathrm{BF}_{01} \approx -\frac{1}{2}\mathrm{BIC}_0 + \frac{1}{2}\mathrm{BIC}_1,$$

hence

$$\boxed{\mathrm{BF}_{01} \approx \exp\left(\frac{\mathrm{BIC}_0 - \mathrm{BIC}_1}{2}\right).}$$

This is a large-sample approximation; it depends on the choice of parameterization and is less reliable for small n or weak priors.

5.3 Interpretation of BF values

Common interpretive thresholds (Jeffreys, Kass & Raftery) are:

- $\mathrm{BF}_{01} > 10$: strong evidence for M_0.
- $\mathrm{BF}_{01} \in (3, 10]$: moderate evidence for M_0.
- $\mathrm{BF}_{01} \in (1/3, 3)$: inconclusive.
- $\mathrm{BF}_{01} < 1/10$: strong evidence for M_1.

These are heuristic and context dependent.

6 Savage–Dickey density ratio: exact identity and practical estimation

6.1 Savage–Dickey identity

Consider nested models where M_0 is obtained from M_1 by fixing a parameter β to a point value (commonly 0). Suppose the prior under M_1 factorizes as $p(\beta, \phi) = p(\beta)p(\phi)$ and the prior under M_0 is $p(\phi)$ (i.e., the prior for nuisance parameters ϕ is the same). Then the Bayes factor comparing M_0 to M_1 can be written as the ratio of prior to posterior marginal densities of β at the point:

$$\boxed{\mathrm{BF}_{01} = \frac{p_{\mathrm{prior}}(\beta = 0)}{p_{\mathrm{post}}(\beta = 0)}.}$$

This is the Savage–Dickey density ratio. It is exact under the stated conditions.

6.2 Prior density at zero for a normal prior

If the prior for β is $\beta \sim \mathcal{N}(0, \sigma^2_{\mathrm{prior}})$, then the prior density at zero is

$$\boxed{p_{\mathrm{prior}}(0) = \frac{1}{\sqrt{2\pi}\,\sigma_{\mathrm{prior}}}.}$$

6.3 Posterior density at zero: estimation from samples

The posterior density $p_{\mathrm{post}}(\beta)$ is typically not available in closed form when using MCMC. Given posterior samples $\{\beta^{(s)}\}_{s=1}^{S}$, estimate $p_{\mathrm{post}}(0)$ using a kernel density estimator (KDE). For a Gaussian kernel $K(u) = (2\pi)^{-1/2}\exp(-u^2/2)$ and bandwidth $h > 0$,

$$\boxed{\hat{p}_{\mathrm{post}}(0) = \frac{1}{Sh}\sum_{s=1}^{S} K\left(\frac{0 - \beta^{(s)}}{h}\right) = \frac{1}{Sh}\sum_{s=1}^{S} \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{(\beta^{(s)})^2}{2h^2}\Big).}$$

Then the Savage–Dickey estimate is

$$\widehat{\mathrm{BF}}_{01} = \frac{p_{\mathrm{prior}}(0)}{\hat{p}_{\mathrm{post}}(0)}.$$

6.4 Bandwidth selection and practical issues

- Bandwidth h controls bias–variance tradeoff. Common rules include Silverman's rule of thumb:

$$h_{\mathrm{Silverman}} = 0.9\,\min\{\hat{\sigma},\ \mathrm{IQR}/1.34\}\,S^{-1/5},$$

  where $\hat{\sigma}$ is the sample standard deviation of $\{\beta^{(s)}\}$ and IQR is the interquartile range.

- If posterior samples are few or concentrated, KDE can be unstable; a fallback is to approximate the posterior by a normal distribution with mean $\hat{\mu}$ and standard deviation $\hat{s}$, giving

$$\hat{p}_{\mathrm{post}}(0) \approx \frac{1}{\sqrt{2\pi}\,\hat{s}}\exp\Big(-\frac{\hat{\mu}^2}{2\hat{s}^2}\Big).$$

- Numerical floors (e.g., 10^{-300}) are applied to avoid division by zero or overflow when densities are extremely small.

6.5 Interpretation and limitations

- Savage–Dickey requires that the prior under the alternative factorizes and that the null is a point restriction of the alternative.
- The BF depends strongly on the prior scale σ_{prior}; different reasonable priors can yield different BFs.
- KDE-based estimates inherit Monte Carlo variability from the posterior samples.

7 Bayesian linear regression: model specification and parameterizations

7.1 Standard (centered) parameterization

Model:

$$y_i \sim p(y_i \mid \alpha, \beta, \sigma) = \mathcal{N}(\alpha + \beta x_i,\ \sigma^2).$$

Priors (example choices used in the code):

$$\alpha \sim \mathcal{N}(0, 5^2), \qquad \beta \sim \mathcal{N}(0, \sigma_{\text{prior}}^2), \qquad \sigma \sim \text{HalfNormal}(2.5).$$

This is the centered parameterization: β is directly sampled.

7.2 Non-centered parameterization

Non-centered parameterization is often used to improve sampling geometry when priors are weakly informative or hierarchical structures exist. Introduce $\beta_{\text{raw}} \sim \mathcal{N}(0, 1)$ and set

$$\beta = \beta_{\text{raw}} \cdot \sigma_{\text{prior}}.$$

This reparameterization decouples scale from standard normal draws and can reduce funnel-shaped posterior geometries that cause sampler difficulties.

7.3 Why non-centered helps

When the posterior is concentrated in a narrow region relative to the prior scale, the centered parameterization can produce strong posterior correlations between β and σ (or other parameters), leading to slow mixing and divergences. The non-centered parameterization often yields more isotropic posterior geometry for the sampler.

8 Student–t likelihood for robust regression

8.1 Student–t density

The Student–t distribution with location μ, scale σ, and degrees of freedom $\nu > 0$ has density

$$p(y \mid \mu, \sigma, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}\,\sigma}\left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}.$$

For regression, set $\mu_i = \alpha + \beta x_i$.

8.2 Robustness properties

- For small ν (e.g., $\nu \in (1, 10)$), the Student–t has heavier tails than the Gaussian, reducing the influence of outliers on parameter estimates.
- As $\nu \to \infty$, the Student–t converges to the Gaussian with variance σ^2.
- The scale parameter σ in the Student–t is not the same as the Gaussian variance; the variance exists only for $\nu > 2$ and equals $\sigma^2 \cdot \nu/(\nu - 2)$.

8.3 Priors for ν

A common weakly informative prior for ν is an exponential or gamma distribution shifted to favor moderate tail heaviness (e.g., $\nu \sim \text{Exponential}(\lambda)$ or $\nu \sim \text{Exponential}(1/30)$ as in the code). This allows the data to inform tail behavior.

9 MCMC sampling diagnostics

Reliable inference from MCMC requires diagnostics. We discuss divergences, Gelman–Rubin $\hat{R}$, and effective sample size.

9.1 Divergences in Hamiltonian Monte Carlo (NUTS)

- Hamiltonian Monte Carlo (HMC) simulates Hamiltonian dynamics to propose new states. Numerical integrators (e.g., leapfrog) approximate continuous trajectories.
- A *divergence* occurs when the numerical integrator fails to accurately follow the true Hamiltonian trajectory, often due to regions of high curvature in the posterior (e.g., funnels).
- Divergences indicate that the sampler may not be exploring the posterior correctly; zero divergences is the desired outcome.
- Remedies include reparameterization (non-centered), increasing target_accept (reducing step size), using more tuning, or changing priors/likelihood (e.g., Student–t).

9.2 Gelman–Rubin $\hat{R}$ statistic

Setup. Run m parallel chains, each of length n after warmup. Let θ_{ij} denote the j-th draw from chain i. Define per-chain means $\bar{\theta}_{i\cdot}$ and overall mean $\bar{\theta}_{\cdot\cdot}$.

$$\bar{\theta}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n}\theta_{ij}, \qquad \bar{\theta}_{\cdot\cdot} = \frac{1}{m}\sum_{i=1}^{m}\bar{\theta}_{i\cdot}.$$

Between-chain variance B and within-chain variance W.

$$B = \frac{n}{m-1}\sum_{i=1}^{m}(\bar{\theta}_{i\cdot} - \bar{\theta}_{\cdot\cdot})^2, \qquad W = \frac{1}{m}\sum_{i=1}^{m} s_i^2,$$

where $s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(\theta_{ij} - \bar{\theta}_{i\cdot})^2$.

Estimate of marginal posterior variance

$$\widehat{\mathrm{Var}}^{+} = \frac{n-1}{n}W + \frac{1}{n}B.$$

Potential scale reduction factor

$$\boxed{\hat{R} = \sqrt{\frac{\widehat{\mathrm{Var}}^{+}}{W}}.}$$

Interpretation: $\hat{R} \to 1$ as chains converge. Practical thresholds: $\hat{R} \leq 1.01$ is excellent; $\hat{R} \leq 1.05$ is often considered acceptable in applied work; values > 1.1 indicate problems.

9.3 Effective sample size (ESS)

MCMC draws are autocorrelated; ESS estimates the number of independent samples equivalent to the correlated chain. For a single chain, ESS can be estimated using the autocorrelation function ρ_k:

$$\mathrm{ESS} \approx \frac{n}{1 + 2\sum_{k=1}^{\infty}\rho_k}.$$

Arviz and other packages compute multichain ESS using spectral estimators. Larger ESS implies more precise Monte Carlo estimates.

10 Posterior predictive checks (PPC)

10.1 Posterior predictive distribution

Given posterior draws $\{\theta^{(s)}\}_{s=1}^{S}$, the posterior predictive distribution for a new or replicated observation y^{rep} is

$$p(y^{\mathrm{rep}} \mid y) = \int p(y^{\mathrm{rep}} \mid \theta)\, p(\theta \mid y)\, d\theta \approx \frac{1}{S}\sum_{s=1}^{S} p(y^{\mathrm{rep}} \mid \theta^{(s)}).$$

10.2 Generating replicated datasets

For each posterior draw $\theta^{(s)}$, simulate $y^{\mathrm{rep},(s)}$ from the likelihood $p(y \mid \theta^{(s)})$. This yields replicated datasets $\{y^{\mathrm{rep},(s)}\}$.

10.3 Summary statistics and discrepancy measures

Compute summary statistics $T(y)$ (e.g., mean, standard deviation) for observed data and for each replicate $T(y^{\text{rep},(s)})$. The PPC p-value (one-sided) is estimated as

$$\Pr\left(T(y^{\text{rep}}) \le T(y) \mid y\right) \approx \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}\{T(y^{\text{rep},(s)}) \le T(y)\}.$$

Interpretation: extreme values (close to 0 or 1) indicate systematic misfit for that statistic.

11 Leave-one-out cross-validation (LOO) and PSIS

11.1 LOO predictive density

The leave-one-out predictive density for observation i is

$$p(y_i \mid y_{-i}) = \int p(y_i \mid \theta)\, p(\theta \mid y_{-i})\, d\theta.$$

Exact LOO requires refitting the model n times; approximate methods reuse full-data posterior draws.

11.2 Importance sampling identity

Using full-data posterior draws $\{\theta^{(s)}\}$, we can write

$$p(y_i \mid y_{-i}) = \int p(y_i \mid \theta)\frac{p(\theta \mid y_{-i})}{p(\theta \mid y)} p(\theta \mid y)\, d\theta = \int p(y_i \mid \theta)\frac{1}{p(y_i \mid \theta)}\frac{p(\theta)p(y_{-i} \mid \theta)}{p(\theta)p(y \mid \theta)} p(\theta \mid y)\, d\theta,$$

which simplifies to an importance sampling estimator with weights proportional to $1/p(y_i \mid \theta^{(s)})$.

11.3 Pareto-smoothed importance sampling (PSIS)

Raw importance weights can be unstable. PSIS fits a generalized Pareto distribution to the upper tail of the importance weights and smooths them, producing stabilized weights. The Pareto shape parameter k diagnoses reliability:

- $k < 0.5$: reliable.
- $0.5 \le k < 0.7$: usable with caution.
- $k \ge 0.7$: unreliable; LOO estimates may be biased.

Arviz and other libraries compute PSIS-LOO and report Pareto k values.

11.4 LOO model comparison

Compute expected log predictive density (ELPD) or LOO information criterion for each model; prefer the model with higher ELPD (or lower LOOIC). Differences can be assessed with standard errors.

12 Highest density interval (HDI) and credible intervals

12.1 Definition

A $100(1-\alpha)\%$ highest density interval (HDI) for a posterior distribution is the narrowest interval $[a, b]$ such that

$$\int_a^b p(\beta \mid y)\, d\beta = 1 - \alpha$$

and for all $x \in [a, b]$ and $x' \notin [a, b]$, $p(x \mid y) \geq p(x' \mid y)$. For unimodal posteriors this is the shortest credible interval.

12.2 HDI exclusion of zero

If the 95% HDI does not contain zero (i.e., either $a > 0$ or $b < 0$), this is often interpreted as evidence that β is credibly nonzero at the 95% level. The code counts how many prior settings yield HDIs excluding zero.

13 Combining evidence: heuristics used in the pipeline

The code aggregates multiple evidence sources:

- **Permutation p-value** for Pearson correlation (frequentist).
- **Savage–Dickey BF** estimates across a grid of prior scales; geometric mean of BF values is used to summarize across priors:

$$\mathrm{BF}_{\mathrm{geo}} = \exp\Big(\frac{1}{G}\sum_{g=1}^{G} \log \mathrm{BF}_g\Big),$$

  where G is the number of prior settings.
- **BIC approximation BF** as a large-sample check.
- **HDI exclusion counts** across priors.
- **PPC summaries** (proportion of replicates with mean or sd less than or equal to observed).
- **LOO preference for robust model** counts.

A conservative decision rule in the code:

- If $\mathrm{BF}_{\mathrm{geo}} \geq 10$: conclude "no relation" (evidence for null).
- If $\mathrm{BF}_{\mathrm{geo}} \leq 0.1$: conclude "relation" (evidence for alternative).
- If HDI exclusion proportion ≥ 0.5 and permutation $p < 0.05$: conclude "relation".
- Otherwise: "inconclusive".

These thresholds are heuristic and chosen for conservatism.

14 Numerical and practical considerations

14.1 Finite precision and floors

Densities and likelihoods can be extremely small; numerical floors (e.g., 10^{-300}) are used to avoid division by zero and floating point underflow.

14.2 Monte Carlo variability

All sample-based estimates (KDE, posterior summaries, PPC proportions, PSIS-LOO) have Monte Carlo error. Increasing draws and chains improves precision.

14.3 Standardization of predictors

Standardizing x (centering and scaling) improves numerical stability and interpretability of priors on β. If x is standardized to mean 0 and sd 1, a prior $\beta \sim \mathcal{N}(0, \sigma^2_{\text{prior}})$ has a direct interpretation in terms of expected change in y per standard deviation change in x.

15 Worked symbolic derivations and small proofs

15.1 Derivation: relation between r and OLS slope

Starting from OLS slope:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}.$$

Divide numerator and denominator by $\sqrt{S_{xx}S_{yy}}$:

$$\hat{\beta} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \cdot \frac{\sqrt{S_{xx}S_{yy}}}{S_{xx}} = r \cdot \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}}.$$

Thus $r = \hat{\beta} \cdot \sqrt{S_{xx}/S_{yy}}$.

15.2 Derivation: BIC for Gaussian linear model

For Gaussian errors with variance σ^2, the log-likelihood is

$$\log L(\theta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\text{RSS}.$$

Maximizing over σ^2 yields $\hat{\sigma}^2 = \text{RSS}/n$. Plugging in,

$$\log L(\hat{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{\text{RSS}}{n}\right) - \frac{n}{2}.$$

Up to constants independent of model, $-2\log L(\hat{\theta})$ is proportional to $n\log(\text{RSS}/n)$. Adding the penalty $k\log n$ yields the BIC formula used earlier.

15.3 Derivation: Savage–Dickey identity (sketch)

Let M_1 be the full model with parameter β and nuisance ϕ. The marginal posterior of β under M_1 is

$$p(\beta \mid y, M_1) = \frac{p(\beta) \int p(y \mid \beta, \phi) p(\phi)\, d\phi}{p(y \mid M_1)}.$$

Evaluating at $\beta = 0$,

$$p(0 \mid y, M_1) = \frac{p(0) \int p(y \mid 0, \phi) p(\phi)\, d\phi}{p(y \mid M_1)}.$$

But the numerator integral is exactly $p(y \mid M_0)$ (the marginal likelihood under the null). Rearranging gives

$$\frac{p(y \mid M_0)}{p(y \mid M_1)} = \frac{p(0)}{p(0 \mid y, M_1)}.$$

This is the Savage–Dickey identity.

16 Summary of recommended practical workflow (algorithmic steps)

1. **Data cleaning and alignment.** Remove non-finite values and align paired observations.
2. **Frequentist baseline.** Fit OLS, compute slope, R^2, Pearson r, and permutation p-value with N_{perm} permutations.
3. **Bayesian grid.** For a grid of prior scales $\{\sigma_{\text{prior},g}\}$, fit Bayesian linear models using non-centered parameterization and optionally Student–t likelihoods if sampling issues arise.
4. **Diagnostics.** For each fit, check divergences (should be 0), $\hat{R}$ (should be near 1), and ESS.
5. **Posterior summaries.** Extract posterior samples of β, compute means, sds, and 95% HDIs.
6. **Savage–Dickey BF.** Estimate posterior density at 0 via KDE (or normal approximation fallback) and compute BF for each prior scale.
7. **Aggregate BF.** Compute geometric mean of BF values across priors to summarize sensitivity to prior scale.
8. **PPC and LOO.** Compute posterior predictive summaries and PSIS-LOO comparisons between Gaussian and robust (Student–t) fits.
9. **Decision heuristic.** Combine BF, HDI counts, permutation p-value, PPC, and LOO preferences to form a conservative verdict.

17 Appendix: useful formulae and constants

Normal density: $\mathcal{N}(x \mid \mu, \sigma^2) = \dfrac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \dfrac{(x-\mu)^2}{2\sigma^2} \Big)$.

Student–t density: $t(x \mid \mu, \sigma, \nu) = \dfrac{\Gamma\big(\frac{\nu+1}{2}\big)}{\Gamma\big(\frac{\nu}{2}\big)\sqrt{\nu\pi}\,\sigma} \Big(1 + \dfrac{(x-\mu)^2}{\nu\sigma^2}\Big)^{-(\nu+1)/2}$.

Silverman bandwidth (rule of thumb): $h = 0.9 \min\{\hat{\sigma},\ \mathrm{IQR}/1.34\}\, S^{-1/5}$.

Concluding remarks

This document has integrated and expanded every mathematical element present in the analysis pipeline, providing definitions, derivations, practical estimation formulas, and numerical considerations. The pipeline intentionally combines frequentist permutation inference, asymptotic BIC approximations, and sample-based Bayesian evidence (Savage–Dickey) together with robust modeling and predictive checks to produce a conservative, multi-angle assessment of evidence for or against a linear relationship between two series.