TRAINING SCHOOL

PART B: Analytical Methodologies for Hematological Disorder Diagnostics and Prognostics

3rd - 4th June 2026 Virtual Meeting
ORGANIZER: Working Group 03 Dr Helen Latsoudis (Leader)
Dr Giacomo Cavalca (Co-Leader)

A. Core Statistical Concepts for a Proper Study Design of a Rare Hematological Disease (Prof. Esin AVCI)

Outline

- Small Sample Sizes (The "Low N" Problem)
- Sparse Data (The "Empty Space" Problem)
- Handling Small Sample Sizes and Sparse Data
- Descriptive Statistics Table For Sparse Data
- Inferential Statistics Table for Sparse Data
- Probability Types in Rare Hematology Disease

Small Sample Sizes (The "Low N" Problem)

There is no universal number that defines "small," but in regression analysis, a dataset with fewer than 30 to 50 observations is generally considered small.

Why it causes problems:

- High Sampling Variance
- Overfitting
- Low Statistical Power

Sparse Data (The "Empty Space" Problem)

It means our data matrix is mostly filled with zeroes, missing values, or uninformative blanks

This can happen in two distinct ways:

A. High-Dimensional Sparsity (Too Many Columns, p≫N)

Example: In DNA analysis, you might have blood samples from **100 patients (Rows)**, but you are tracking **20,000 genes (Columns)**. Most of those genes have zero correlation with the disease, creating a massive, sparse matrix.

B. Categorical / Event Sparsity (Rare Events)

Example: A pharmaceutical company runs a large-scale clinical trial with **10,000 leukemia patients (N = 10,000)** to test a new chemotherapy drug. Out of the 10,000 patients, **only 3 patients** experience this event. The rest of the 9,997 rows in that column are 0

Handling Small Sample Sizes and Sparse Data

1. Switch to Bayesian Statistics: Frequentist statistics (p-values) often fail due to small n. Bayesian approaches incorporate prior knowledge to provide a more meaningful analysis of therapeutic impact.

2. Maximize "Borrowing" and Data Pooling: When data is sparse, we cannot rely solely on the patients sitting in our local clinic. We have to "borrow" data from other sources.

3. Choose Highly Efficient Endpoints: If the study measures vague or rare outcomes (e.g., "overall mortality over 10 years"), the data will be incredibly sparse. We must choose endpoints that yield rich data from every single patient.

4. Use Exact Statistical Tests: Standard statistical tests (like the Chi-square test) rely on large-sample approximations. When our data cells contain zeros or very low numbers, these approximations break down completely.

5. Reduce Variable Dimensionality: When data is sparse, we cannot throw 20 different variables into a model. We use Propensity Score Adjustment or Principal Component Analysis (PCA).

6. Choosing the Right Data Transformation: If we are passing sparse data into statistical models or visualizations, apply transformations to stabilize the variance (log Transformation (log(x + 1))

7. The Rule of Three: This is a mathematical shortcut for sparse binary data. If zero events occur in n subjects, the 95% upper confidence bound for the rate is approximately 3/n.

Descriptive Statistics Table For Sparse Data

Metric Type	Traditional Metric (What to Avoid)	Recommended Sparse Metric (What to Use)	Why it works & What it tells you
Center / Location	Arithmetic Mean	Conditional Mean (Non-zero Mean)	The standard mean will pull toward zero, making the values seem tiny. The conditional mean calculates the average only when an event actually occurs.
		Median / Mode	Highly robust. In very sparse data, the mode will instantly tell you the most common baseline (usually 0).
		Geometric Mean (For Highly Skewed Biomarkers)	It effectively log-transforms the data natively, squishing extreme values together so that sparse, highly skewed laboratory data can be summarized meaningfully.
Distribution Shape	Standard Skewness & Kurtosis	Gini Coefficient	Standard skewness formulas break down with massive zero inflation. The Gini Coefficient (0 to 1) effectively measures the inequality or "concentration" of the rare values.
Visual Summary	Standard Histogram / Boxplot	Zero-Hurdle Histogram	A standard boxplot will compress into a flat line at zero. A zero-hurdle chart separates the "zero vs. non-zero" count into a bar, then plots a detailed histogram only for the remaining positive values.

Descriptive Statistics Table For Sparse Data Cont.

Metric Type	Traditional Metric (What to Avoid)	Recommended Sparse Metric (What to Use)	Why it works & What it tells you
Spread / Dispersion	Standard Deviation (SD) / Variance	Sparsity Ratio (Density)	Instead of measuring distance from a skewed mean, calculate: Sparsity = Count of Zeros/Total Observations
		Median Absolute Deviation (MAD)	Measures the median of the absolute deviations from the data's median. It completely ignores extreme, rare spikes that bloat standard deviation.
		Interquartile Range (IQR)	The IQR measures the spread of the middle 50% of your data (from the 25th percentile to the 75th percentile), completely cutting out the extreme low and high outliers. It is paired with the median to show variation reliably.
		Minimum and Maximum (Full Range)	Providing the exact Min and Max values alongside the Median/IQR allows readers to see the exact footprint of your data, highlighting just how varied the rare disease expression was in your tiny cohort.

Example of Sparse Data

The Clinical Data

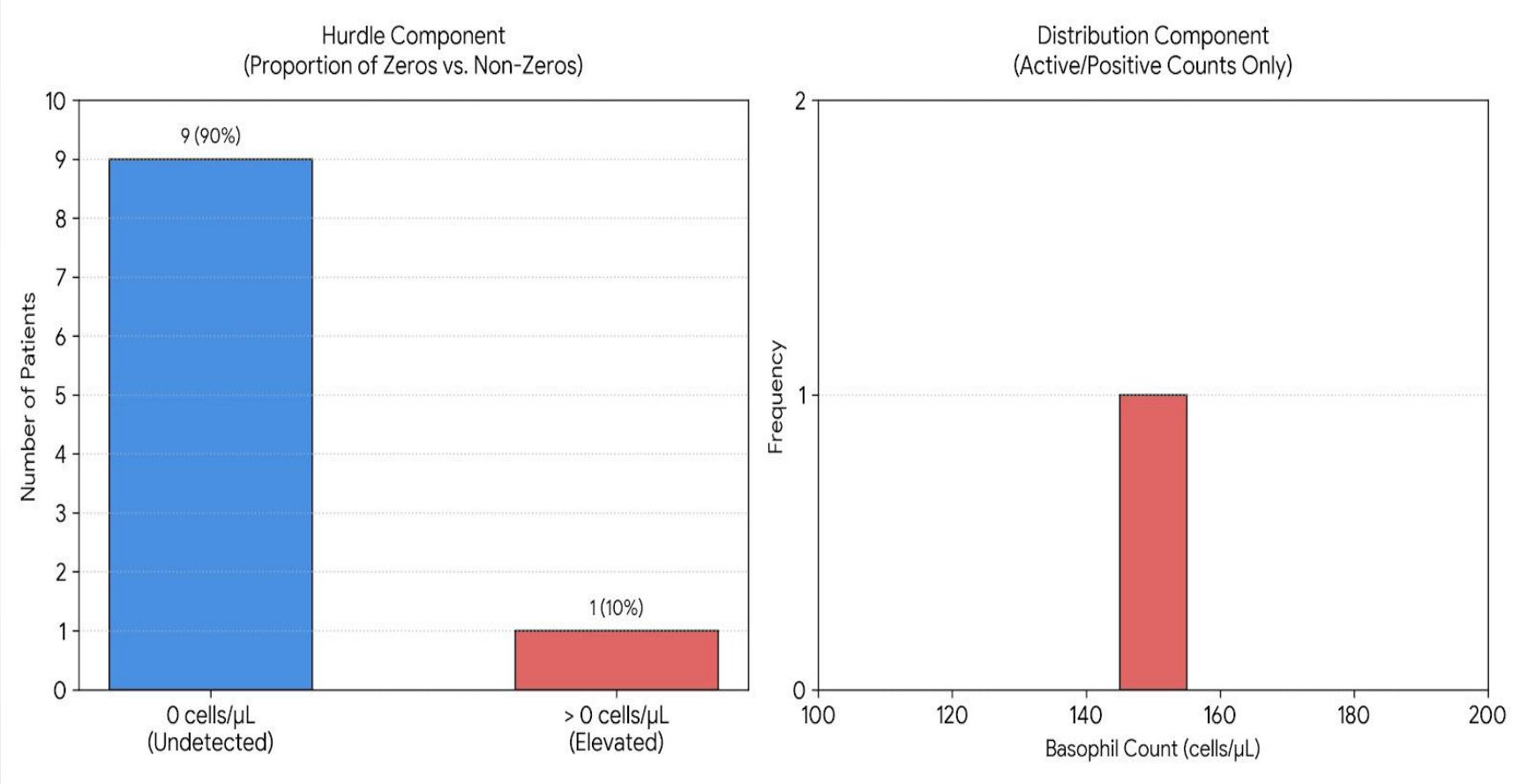
Assume we measure the absolute Basophil count (cells/μL) for 10 patients on a specialized ward:

Data = [0, 0, 0, 0, 0, 0, 0, 0, 0, 150]

Descriptive statistics for this sparse data:

- ✓ Total Sample Size (n): 10
- ✓ Conditional Mean (Active Cases): 150 cells/μL (n=1)
- ✓ Sample Median: 0 cells/μL
- ✓ Geometric mean (G): 0 cells/μL
- ✓ Gini Coefficient: 0.9
- ✓ Sparsity / Zero-inflation: 90% undetected (n=9)
- ✓ Median Absolute Deviation (MAD): 0 cells/μL
- ✓ Interquartile Range (IQR): [0, 0]
- ✓ [Min, Max]: [0, 150]

Zero-Hurdle Histogram



Inferential Statistics Table for Sparse Data

What are you trying to infer?	Standard Statistic (Fails)	Sparse Data Savior (Succeeds)	How it Works / Best Use Case
If a treatment group out-performed a control group (counts)	Chi-Square Test	Fisher's Exact Test	How it works: Calculates the exact hypergeometric probability of the observed table configuration rather than approximating it. Best Use Case: Small clinical trials or rare event comparisons (e.g., assessing treatment side effects when only 2 or 3 patients experience them).
If a biomarker significantly changed (continuous values)	Paired t-test	Wilcoxon Signed-Rank Test (or Permutation)	How it works: Discards the raw values and instead ranks the absolute differences between pairs, effectively stripping away the skewing power of zeros and massive outliers. Best Use Case: Before-and-after laboratory measurements for highly targeted therapies where the baseline count for most patients is zero.
If multiple independent patient cohorts differ in a sparse biomarker (3+ groups)	One-Way ANOVA	Kruskal-Wallis Test (with Post-hoc Dunn's)	How it works: Converts all continuous values into a single global ranking pool across all groups. It evaluates whether the median ranks significantly differ without assuming a normal distribution. Best Use Case: Comparing a rare cell or cytokine count across 3 or more distinct disease variants (e.g., Type 1 vs. Type 2 vs. Type 3

Inferential Statistics Table for Sparse Data cont.

What are you trying to infer?	Standard Statistic (Fails)	Sparse Data Savior (Succeeds)	How it Works / Best Use Case
The true risk ratio of a rare genetic mutation	Standard Logistic Regression	Firth's Penalized Regression	How it works: Introduces a penalty term based on the Fisher information matrix into the likelihood function to prevent coefficients from blowing up or failing to converge. Best Use Case: Epidemiology and genomics when you have "sparse cells" or "complete separation" (e.g., when a rare mutation appears only in the disease group and never in the control group).
Predictors of rare event frequencies & count intensities	Standard Poisson Regression	Zero-Inflated Poisson (ZIP) or Negative Binomial	How it works: Standard Poisson assumes the Mean = Variance. Sparse counts usually suffer from overdispersion. ZIP models the zeros separately from the actual continuous counts using a two-part joint probability system. Best Use Case: Modeling the number of rare clinical events (e.g., seizure frequency per year, or hospital readmission counts for an orphan disease).
The true real-world efficacy probability of an orphan drug	Frequentist p-value	Bayesian Posterior Estimation	How it works: Combines a mathematical "prior" distribution (built from historical data, animal models, or expert consensus) with the sparse current trial data to compute an updated probability curve. Best Use Case: Ultra-rare disease

Probability Types in Rare Hematology Disease

1. **The Probability of Diagnosis: Bayes' Theorem**

Using $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ to calculate the **Positive Predictive Value (PPV)**.

The Example: A test for a rare blood cancer is 99% accurate, but the disease only affects 1 in 10,000 people.

The Reality: If a patient tests positive, the probability they actually have the disease is only about **1%**.

2. Transition Probabilities (Markov Models)

Estimating the probability of moving from "Stable→"Crisis" →"Remission."

By calculating the **transition probability** at each month, researchers can predict the "Event-Free Survival" for a cohort.

Example: Modeling the life cycle of **Myelodysplastic Syndrome (MDS)**.

The States: State A (Stable), State B (Transfusion Dependent), State C (Transformation to Leukemia).

The Probability: We calculate that a patient has a **15% monthly probability** of staying in State A and a **2% probability** of jumping to State C.

The Use: This helps insurance companies and hospitals predict how many beds or units of blood they will need over the next 5 years for a rare cohort.

3. Power and Sample Size (The Frequentist Trap)

In a common disease, we want a power (1-β) of 80% to detect a treatment effect. In rare diseases, the probability of achieving this with a small group is nearly zero unless the effect size is massive. we calculate the exact probability of the observed data occurring by chance.

Example: A trial for a new drug for **Pure Red Cell Aplasia**.

The Problem: You only have **8 patients** in the world willing to join the study.

The Frequentist Failure: A standard t-test requires 30+ people to be valid.

The Solution (Fisher's Exact Test): You look at the 8 patients. 7 improved on the drug, 1 did not.

The Math: You calculate the *exact* probability that 7/8 people would get better by pure "luck." If that probability is less than 5%, the drug is considered a success despite the small group.

4. Poisson Distribution

When looking at complications like the probability of a rare thrombotic event (blood clot) in a PNH patient we use the Poisson distribution.

$$P(k\ event\ in\ interval) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Example: Monitoring **Intracranial Hemorrhage (Brain Bleeds)** in patients with **Rare Severe Hemophilia**.

The Scenario: In a group of 50 patients, you expect **λ= 0.5** bleeds per year (1 bleed every 2 years).

The Event: Suddenly, 3 bleeds happen in a single month.

The Math: The Poisson formula shows there is only a **0.01% probability** of this happening by chance.

The Result: You immediately stop the study to check if the new medication is actually *causing* the bleeds, rather than assuming it's just "bad luck."

5. Bayesian Posterior Probabilities

This is the "gold standard" for rare disease study design. It allows us to "update" our belief as data trickles in.

Example:

Prior: Based on 5 previous patients, we think the drug has a 40% chance of working.

New Data: We treat 2 more patients, and they both improve.

Posterior: Our "belief" in the drug's success is updated to 65% based on the combined probability.

Why it works: It prevents a study from "failing" just because it didn't meet a rigid frequentist sample size.

TRAINING SCHOOL

PART B: Analytical Methodologies for Hematological Disorder Diagnostics and Prognostics

3rd - 4th June 2026 Virtual Meeting
ORGANIZER: Working Group 03 Dr Helen Latsoudis (Leader)
Dr Giacomo Cavalca (Co-Leader)

B. Bayesian Methods (Prof. Esin AVCI)

Outline

- Core principles of Bayesian inference (prior, likelihood, posterior)
- Bayesian linear regression (linear, logistic)
- Bayesian logistic regression
- Bayesian survival analyses (time to event models)
- Simulations and interpretation of graphical outputs (posterior plots, probability curves)

Bayesian Inference

- Frequentist stats treat every trial like it's the first day of the World ignoring all previous lab work, animal studies, or registries.

- Bayesian inference is a framework for learning from data by **updating beliefs** in a mathematically coherent way.

Updated belief (Posterior) = Previous knowledge (prior) + Data Evidence (Likelihood)

- It is especially valuable in medicine and particularly in **rare hematological diseases** because:

✓ Small sample sizes
✓ Heterogeneous patient populations
✓ Limited clinical trials
✓ Strong reliance on expert knowledge

Prior (Previous knowledge)

The **prior distribution** represents what is known (or believed) about a parameter *before* observing current data.

Can come from:

- ✓ Previous studies
- ✓ Expert opinion
- ✓ Registry data
- ✓ Meta-analysis

Example: Suppose we are studying a rare leukemia subtype (e.g., Acute Promyelocytic Leukemia). Previous studies suggest that the probability of remission with a certain therapy is around **70%**.

Types of Priors

Priors exist on a spectrum based on how much influence they have on the final result:

- **Non-informative (Flat) Priors:** These suggest that all values are equally likely (e.g., a Normal distribution with a massive variance).

  Effect: The model relies almost 100% on the data. The results will be very close to standard frequentist (p-value) results.

- **Weakly Informative Priors:** These provide a broad range of plausible values but "rule out" the impossible (e.g., a person's age cannot be 500).

  Effect: They act as a stabilizer, preventing the model from producing extreme or erratic estimates, especially in small samples.

- **Informative Priors:** These are based on previous clinical trials, expert consensus, or physiological facts.

  Effect: They "nudge" the model. If our dataset is small (like our 12 events), the informative prior provides the extra "weight" needed to reach a stable conclusion.

Likelihood (Data Evidence)

Likelihood: What Does the Data Say?

- The **likelihood** represents the probability of observing the data given a parameter value.
- It comes directly from your **current study data**
- It reflects how compatible different parameter values are with observed outcomes

Posterior: Updated Belief

- The posterior distribution combines prior knowledge and observed data

- It answers: “What is the most plausible treatment effect now?”

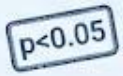
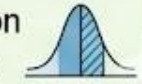
The Posterior usually calculated via software using Markov Chain Monte Carlo or MCMC sampling) merges the prior and the data.

Credible Interval

In Bayesian statistics, a **credible interval** is a range of values within which an unobserved parameter (like a population mean or a success rate) falls with a certain probability, given the observed data.

Example: If you calculate a 95% credible interval for a drug's effectiveness to be between 70% and 85%, it means **there is a 95% probability that the true effectiveness lies between 70% and 85%.**

Frequentist vs. Bayesian Approaches in Rare Disease Clinical Trials

Attribute	Frequentist Approach (Traditional)	Bayesian Approach (Modern Rare Disease)
Fundamental Question	How likely is the data, given a true null hypothesis (p-value)?	How likely is the hypothesis, given the new and prior data?
Handling Prior Knowledge	Prior data is ignored; analysis relies solely on the current trial. 	Explicitly incorporates prior beliefs or data from registries, previous trials, or biology. 
Primary Goal	Achieve statistical significance ($p < 0.05$). 	Update probability of treatment effect or success. 
Sample Size ('n') Requirements	Often requires large sample sizes to find significant effects in rare diseases. 	Can adaptively borrow information from smaller sample sizes effectively. 
Interim Monitoring	Strict rules; stopping early requires large penalties (e.g., 'O'Brien-Fleming'). 	Seamless; can assess predictive probability of success or futility continuously. 
Result Interpretation	Provides point estimates with Confidence Intervals (e.g., 95% CI). 	Provides a Posterior Distribution (e.g., 95% Credible Intervals). 

Example: Case: Predicting Treatment Response in Aplastic Anemia

Prior: Let's say prior belief: "Treatment success ≈ 30%"

- ✓ **Weak prior (little confidence)**
  Equivalent to "10 prior patients"
- ✓ **Informative prior (high confidence)**
  Equivalent to "100 prior patients"

Likelihood: You treat 8 patients, and 5 respond the remission. The data suggest a **62% response rate**

Posterior: Posterior estimate ≈ **40–50% Weak prior**
Posterior estimate ≈ **30–40% Informative prior**

- ✓ More conservative
- ✓ Clinically safer
- ✓ Less misleading

Frequentist Conclusion: "Response = 62%" (unstable due to small sample)

Bayesian Linear Regression

Why Use It?

- ✓ **Handling Uncertainty:** Instead of just saying "the weight is 5.2," you get a distribution. This tells you how confident the model is.
- ✓ **Small Datasets:** If you have very small data, standard models tend to overfit. Bayesian models use the "prior" to stay grounded and prevent wild predictions.
- ✓ **Incorporating Expert Knowledge:** If you already know from previous studies that a certain variable should have a positive effect, you can bake that into the **Prior**.
- ✓ **Automatic Regularization:** Certain priors (like a Gaussian prior centered at zero) act exactly like Ridge or Lasso regression, helping to simplify the model.

The Core Philosophy

In a standard linear regression, we assume the relationship is:

$$y = X\beta + \epsilon$$

We usually look for a single value of β that minimizes the error. In Bayesian regression, we treat β as a **random variable**. We start with a guess, see the data, and update that guess.

How It Works (The Math)

We use **Bayes' Theorem** to find the posterior distribution of our parameters:

$$P(\beta|y,X) = \frac{P(y|X,\beta)P(\beta)}{P(y|X)}$$

- **The Prior** $P(\beta)$**:** What we believe about the coefficients before seeing any data (e.g., "I think the slope is around 0").
- **The Likelihood** $P(y|X,\beta)$**:** How well the data is explained by specific parameters.
- **The Posterior** $P(\beta|y,X)$**:** Our updated belief about the parameters after combining the prior and the data.

Bayesian Logistic Regression

The Core Philosophy

Bayesian Logistic Regression is used when your outcome is categorical (e.g., "Ill" vs. "Healthy")

While standard logistic regression gives you a single probability, the Bayesian version gives you a **distribution of probabilities**, allowing you to quantify how sure you are about a patient's diagnosis.

The Mathematical Model

The model follows the same logistic (logit) form as standard regression, but the coefficients (β) are treated as random variables:

$$P(y_i = 1|X_i, \beta) = logit^{-1}(X_i\beta) = \frac{1}{1 + e^{-X_i\beta}}$$

- **Prior** $P(\beta)$**:** Usually a Normal distribution. You can set a "Strong Prior" if you have established medical thresholds for anemia.
- **Likelihood:** A Bernoulli distribution (since the outcome is 0 or 1).
- **Posterior:** Calculated via MCMC sampling because the logistic function doesn't have a simple "closed-form" solution.

Bayesian Survival Regression

The Core Philosophy

Survival analysis is used when we’re interested in the time until an event occurs (e.g., equipment failure, medical recovery, or customer churn). Adding a Bayesian framework to this allows us to incorporate prior knowledge and handle uncertainty more intuitively than traditional frequentist methods.

To understand the Bayesian approach, we first look at the fundamental "survival" building blocks:

- **Survival Function *S(t)*:** The probability that the event has not yet happened by time *t*.
- **Hazard Function *h(t)*:** The instantaneous rate at which the event occurs, given that the subject has survived up to time *t*.
- **Censoring:** A unique feature of survival data where we know an event *hasn't* happened yet (e.g., a study ends before a patient recovers), but we don't know exactly when it will.

The Mathematical Model

In mathematical framework of **Bayesian survival model**, we treat the regression coefficients (β) and any baseline hazard parameters as random variables rather than fixed constants.

Instead of maximizing a likelihood function to get a single point estimate, we combine the survival likelihood with **prior distributions** to solve for the **posterior distribution** of our parameters using Bayes' Theorem:

$$p(\beta, \theta | Data) \propto L(Data | \beta, \theta) \times p(\beta) \times p(\theta)$$

Where β represents the covariate coefficients, θ represents the parameters defining the baseline hazard, and L(Data | β,θ) is the survival likelihood that incorporates censoring.

Example: Let's say we want to predict **Hemoglobin (Hb)** levels based on **Packed Cell Volume (PCV)**

$$\mathbf{Hb} = \beta_0 + \beta_1 \times \mathbf{PCV}$$

b0 = c(0, 0.33) (The Prior Mean)

B0 = c(0.01, 100) (The Prior Precision)

Output

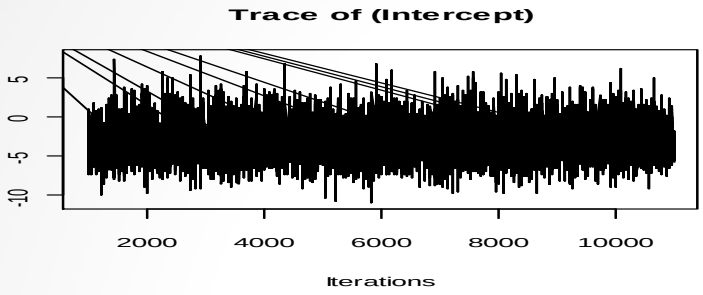
Iterations = 1001:11000
Sample size per chain = 10000
Empirical mean and standard deviation for each variable,
plus standard error of the mean:

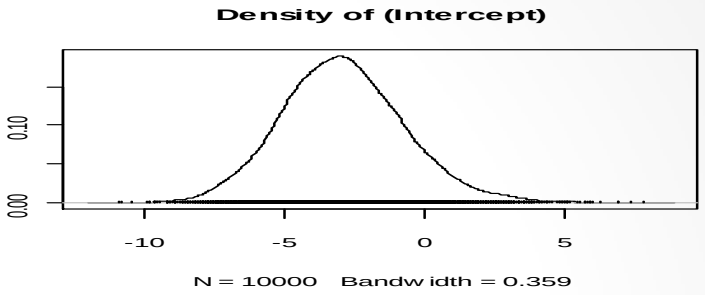
	Mean	SD	Naive SE	Time-series SE
(Intercept)	-2.9386	2.21148	0.0221148	0.0235811
PCV	0.4111	0.05745	0.0005745	0.0006112
sigma2	0.6901	0.45400	0.0045400	0.0055437

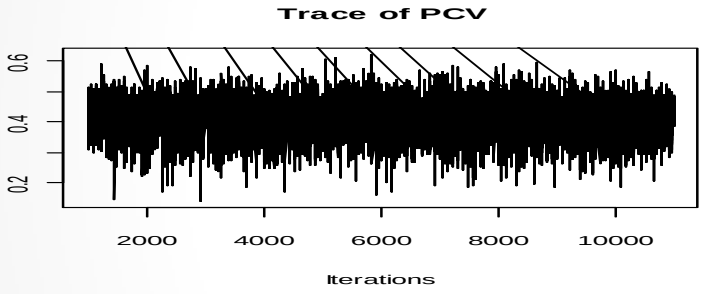
2. Quantiles for each variable:

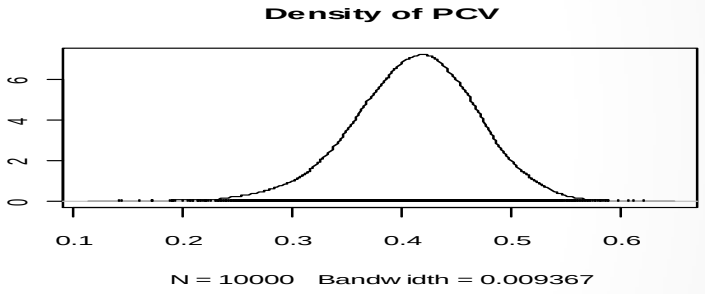
	2.5%	25%	50%	75%	97.5%
(Intercept)	-7.0454	-4.4313	-3.0304	-1.5677	1.7309
PCV	0.2882	0.3753	0.4137	0.4500	0.5179
sigma2	0.2420	0.4077	0.5676	0.8221	1.8703

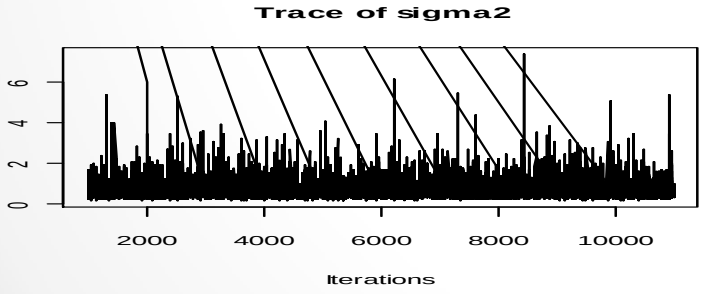
Diagnostic plots

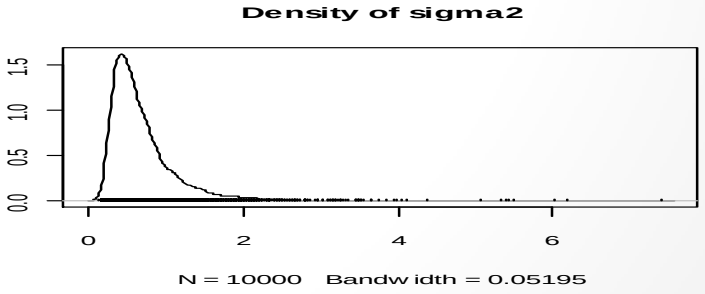












Checklist table for evaluating Bayesian regression output

Category	Output / Metric	Checklist Item	Evaluation Criteria (What to check)
1. Metadata & Tuning	Warmup / Burn-in	Verify Adaption Phase	Ensure the burn-in period was long enough for the sampler to find the high-probability region. Warmup samples must be discarded and never used for final inference.
2. Convergence Diagnostics	The Trace Plots	Verify Chain Mixing	Independent sampling chains must overlap and travel together horizontally, creating a flat, tightly bound "fuzzy caterpillar" appearance.
	The Density Plots	Verify Posterior Shape	Distribution curves should be smooth, continuous, and stable. Jagged spikes, gaps, or multiple isolated peaks (multimodality) suggest the chains struggled to explore the parameter space.
3. Parameter Estimates	Mean / Median	Evaluate Point Estimates	Look at the sign (+/-) and magnitude of the coefficient. Does the direction of the effect make practical and theoretical sense?
	SD (Standard Deviation)	Assess Posterior Uncertainty	This represents the overall uncertainty of the parameter distribution. A massive SD relative to the mean indicates a highly volatile effect.
	95% Credible Interval	Determine Effect Certainty	Check if the interval contains 0. If 0 is entirely outside the interval boundaries, there is a >95% probability that the predictor has a real, directional effect.
4. Performance & Uncertainty Accuracy	Naive SE	Measure Pure Simulation Error	Calculated as SD / $\sqrt{N}$(where N is total iterations). It assumes all MCMC samples are completely independent, which is rarely true due to autocorrelation.
	Time-series SE (MCSE)	Adjust for Autocorrelation	Must be small (typically < 5% of the parameter's standard deviation). This metric adjusts the standard error to account for the fact that sequential MCMC samples are correlated.

Prior Precision

Scenario	Prior Precision ($\tau=1/\sigma 2$)	The Core Challenge / Behavior
Non-Informative / Flat	$\tau < 0.0001$	**The "No Anchor" Danger:** The prior is so flat that the mathematical weight is practically zero. The posterior is entirely dependent on the data, meaning small sample sizes will result in massive, unstable confidence ranges.
Weakly Informative	$0.01 \leq \tau \leq 1.0$	Soft Regularization: This acts as a protective boundary. It rules out physically or mathematically impossible values but remains weak enough to let the data dictate the final exact number.
Highly Informative	$\tau > 1.0$	The Weight of History: The prior precision dominates the equation. If your cutoff pushes precision above 1.0, you are explicitly telling the model: "I am highly confident the true coefficient lives in a tiny window."


Thank You
for your attention