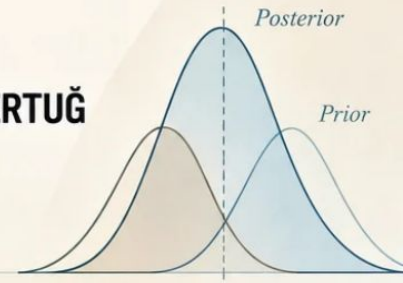
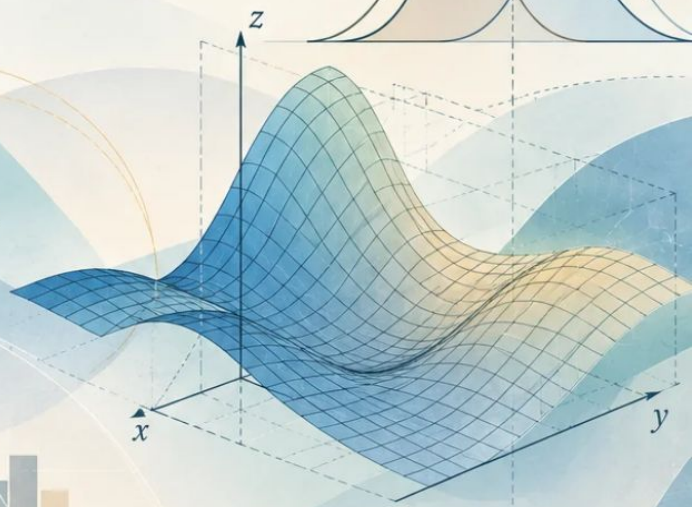
ADVANCES IN STATISTICAL SCIENCE:

REGRESSION ANALYSIS, BAYESIAN METHODS, AND APPLIED RESEARCH

EDİTÖR
PROF. DR. ZELİHA KAYGISIZ ERTUĞ







$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



$$p(\theta|D) = \frac{p(D|\theta)\,p(\theta)}{p(D)}$$

BİDGE yayınları

BİDGE Yayınları

Advances in Statistical Science: Regression Analysis, Bayesian Methods, and Applied Research

Editor: Prof. DR. ZELİHA KAYGISIZ ERTUĞ

ISBN: 978-625-5057-90-7

1st Edition
Page Layout By: Gözde YÜCEL
Publication Date: 2026-10-25
BİDGE Yayınları

All rights reserved. No part of this work may be reproduced in any form or by any means, except for brief quotations for promotional purposes with proper source attribution, without the written permission of the publisher and the editor.

Certificate No: 71374
All rights reserved © BİDGE Yayınları
www.bidgeyayinlari.com.tr - bidgeyayinlari@gmail.com
Krc Bilişim Ticaret ve Organizasyon Ltd. Şti.
Güzeltepe Mahallesi Abidin Daver Sokak Sefer Apartmanı No: 7/9 Çankaya / Ankara

CONTENTS

EFFECTS OF MULTICOLLINEARITY IN THE LINEAR REGRESSION MODEL: A CONTROLLED MONTE CARLO SIMULATION AND AN APPLIED ANALYSIS OF THE LONGLEY DATA 1

BERRİN ERKAN

BAYESIAN HIERARCHICAL REGRESSION FOR MULTIGROUP DATA: SHARED AND GROUP-SPECIFIC EFFECTS THROUGH PARTIAL POOLING 26

FATİH SAĞLAM, BURÇİN ŞEYDA ÇORBA

BAYESIAN ROBUST REGRESSION FOR CENSORED SURVIVAL DATA: A COMPARISON OF NORMAL, STUDENT-T, AND CAUCHY ACCELERATED FAILURE TIME MODELS USING PBC DATA 64

BURÇİN ŞEYDA ÇORBA, FATİH SAĞLAM

DEĞİŞEN AİLE, YAVAŞLAYAN NÜFUS: TÜRKİYE'NİN 16 YILLIK DEMEOGRAFİK EĞİLİMLERİ 80

CENGİZ GAZELOĞLU

CHAPTER 1

EFFECTS OF MULTICOLLINEARITY IN THE LINEAR REGRESSION MODEL: A CONTROLLED MONTE CARLO SIMULATION AND AN APPLIED ANALYSIS OF THE LONGLEY DATA

BERRİN ERKAN[1]

Introduction

Linear regression is widely employed to examine the conditional association between an outcome and one or more explanatory variables. However, variables measuring different aspects of the same phenomenon often move together. Strong correlations between the price level and nominal income, years of education and work experience, or various financial indicators are not surprising. Dependencies among explanatory variables can complicate the interpretation of what each regression coefficient separately measures. This problem is generally known as *multicollinearity*.

Multicollinearity is often misunderstood as a violation that necessarily invalidates regression analysis. In fact, the classical linear model does not require explanatory variables to be

[1] Assist. Prof. Dr., Çanakkale Onsekiz Mart University, Department of Statistics, Orcid: 0000-0001-5889-7461

uncorrelated with one another. Under full column rank, exogeneity, and the standard assumptions about errors, strong but imperfect collinearity does not automatically make OLS biased. The primary difficulty is that there may be insufficient independent variation in the data to estimate individual coefficients precisely. Standard errors may therefore increase, coefficient signs may become sensitive to the sample, and the statistical power of individual hypothesis tests may decline (Montgomery, 2021) (Belsley, 1980).

At the same time, a model can have a high in-sample R^2 even when several individual coefficients are not statistically significant. The model can use predictors jointly to explain variation in Y while being unable to assign that shared explanatory information reliably to individual predictors. Prediction for new observations does not invariably deteriorate under severe collinearity either: performance may be preserved if test observations follow the same covariance structure, whereas it can deteriorate sharply if the predictors' pattern of joint variation changes (Dormann, 2013).

The purpose of this investigation is not merely to state these distinctions in theory, but to demonstrate them with a deliberately controlled simulation and to complement the results with a diagnostic example using observed data. In this way, expected theoretical effects can be considered separately from diagnostic symptoms appearing in a real dataset.

CONCEPTUAL AND THEORETICAL FRAMEWORK

Let y denote the response vector, X the explanatory-variable matrix, and ε the error vector. The classical model is

$$y = \beta_0 \mathbf{1}_n + X\beta + \varepsilon, \qquad E(\varepsilon \mid X) = 0, \qquad \mathrm{Var}(\varepsilon \mid X) = \sigma^2 I_n.$$

The model contains p slope coefficients. When the columns of the full design matrix, including the intercept, are linearly independent, the OLS solution is unique. Under exact multicollinearity, one explanatory variable is an exact linear combination of others, so a unique classical OLS coefficient vector cannot be identified. Under strong but imperfect collinearity, a solution exists, yet may be numerically and statistically sensitive. Dependence among explanatory variables and correlation between regressors and the error term are different problems: the latter may violate exogeneity and bias OLS, while the former alone does not imply such bias.

For centered data, the OLS estimator and its conditional covariance matrix are

$$\hat{\beta} = (X_c^T X_c)^{-1} X_c^T y_c, \qquad \mathrm{Var}\big(\hat{\beta} \mid X\big) = \sigma^2 (X_c^T X_c)^{-1}.$$

This equation shows why small eigenvalues of $X_c^T X_c$ provide little information about particular linear combinations of coefficients. Both inflated standard errors and unfavorable condition indices arise from this underlying algebra. The Gauss–Markov property of OLS as the best linear unbiased estimator continues to hold in the presence of strong collinearity provided the design matrix retains full rank and the assumptions are satisfied. However, “best” applies within the class of linear unbiased estimators; it does not establish universal superiority to biased estimators such as ridge.

Variance inflation and the VIF

Let R_j^2 be the coefficient of determination from regressing the jth explanatory variable on the remaining predictors. The variance inflation factor is

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}, \qquad \mathrm{Var}\left(\hat{\beta}_j \mid X\right) = \frac{\sigma^2}{\sum_i \left(x_{ij} - \bar{x}_j\right)^2} \mathrm{VIF}_j.$$

Holding the marginal spread of the relevant predictor and the error variance fixed, the variance of its slope estimate increases as that predictor becomes more explainable by the other variables. A VIF of one indicates no linear variance inflation from the other predictors, while higher values indicate greater inflation. Thresholds such as 5 or 10 are not by themselves adequate grounds for deleting variables (O'Brien, 2007). The inferential implications of a given VIF also depend on sample size, noise level, and effect magnitude.

For the correlation structure $\Sigma_\rho = (1 - \rho) I_p + \rho \mathbf{1}\mathbf{1}^T$ used here, the population VIF at $p = 6$ is

$$\mathrm{VIF}_{\mathrm{pop}} = \frac{1 + (p - 2)\rho}{(1 - \rho)\{1 + (p - 1)\rho\}}.$$

For example, when $\rho = 0.99$, the population VIF is approximately 83.36. The mean of the *maximum* VIF in realized samples may be higher: the theoretical population VIF and the largest VIF in a finite sample are different statistics.

The condition index and its spectral interpretation

Let Z be the explanatory-variable matrix after centering and scaling by sample standard deviations. The condition index is defined here as $\kappa(Z) = s_{\max}(Z) / s_{\min}(Z)$. As this ratio increases, even small perturbations in the data can translate into large changes in the estimated coefficients. (Belsley, 1980) systematically

discussed the joint use of condition indices and variance-decomposition proportions.

For the equicorrelation matrix with $p = 6$, the eigenvalues are $1 + 5\rho$ and $1 - \rho$, the latter repeated five times. Thus, a population-scale spectral condition ratio is approximately $\sqrt{(1 + 5\rho)/(1 - \rho)}$. At $\rho = 0.99$ this is about 24.39, although the realized ratio between the largest and smallest singular values in a sample may differ. Pairwise correlations alone can also fail to detect a linear dependence involving three or more predictors jointly.

Individual tests and confidence intervals

For $H_0: \beta_1 = 0$, the classical t-statistic under appropriate model assumptions is $t = \hat{\beta}_1 / \widehat{SE}(\hat{\beta}_1)$. When the true β_1 differs from zero, a larger standard error tends to reduce the absolute t-statistic, potentially lowering test power. A genuinely influential variable may appear statistically insignificant because the available data do not distinguish its effect with sufficient precision; nonsignificance is not evidence that the effect is absent.

The two-sided 95% confidence interval is $\hat{\beta}_1 \pm t_{0.975,n-p-1} \widehat{SE}(\hat{\beta}_1)$. If the model assumptions hold and sampling is random, multicollinearity need not impair the coverage probability of this conventional interval, though it may increase its expected length. This distinction shows circumstances in which multicollinearity makes inference less precise rather than inherently incorrect.

Why can the overall F-test and R^2 behave differently?

The overall F-test evaluates the joint null hypothesis $H_0: \beta_1 = \cdots = \beta_p = 0$, whereas each individual t-test assesses whether a specific coefficient is zero conditional on the other predictors. The targets of these tests are different. Multiple predictors

can jointly convey substantial information even when it is difficult to assign that information to individual variables. A strong joint F-test and weak individual t-tests can therefore occur simultaneously (Montgomery, 2021).

The in-sample goodness-of-fit measure $R^2 = 1 - SSE/SST$ cannot mechanically decrease when predictors are added to an OLS regression with an intercept; a high R^2 is therefore not evidence of stable individual coefficients. Adjusted R^2 applies a simple penalty for the number of variables but does not automatically make collinear coefficients identifiable. Furthermore, if population signal variance changes as correlation varies in an experiment, a difference in R^2 cannot be attributed solely to collinearity. In the present study, signal strength is deliberately held constant.

Sources of collinearity and possible responses

Collinearity can arise from common time trends, similar measurements, subcomponents of the same quantity, compositional variables summing to a constant, polynomial terms, a narrow sampling range, or a large number of predictors relative to observations. Depending on the context, suitable responses may include collecting additional observations, expanding the range of predictor values, combining variables based on subject-matter knowledge, or reporting a well-identified contrast. Ridge and other regularization methods are options when prediction is the priority. In studies requiring causal or structural interpretation, however, removing a theoretically essential variable solely to achieve a smaller VIF may create a new model-specification error.

RESEARCH QUESTIONS AND DESIGN

The research examines four questions: (1) how the standard error and sign stability of an OLS coefficient change as multicollinearity rises; (2) how much the t-test power declines for a truly nonzero coefficient; (3) how 95% confidence interval coverage and the power of the overall F-test behave under the same conditions; and (4) whether in-sample R^2 responds systematically to increasing collinearity when population signal strength is held constant. An additional applied question concerns what VIF, the condition index, and a leave-one-out perturbation experiment reveal in real data.

Monte Carlo data-generating process

For each observation, $X_i \sim N_6(0, \Sigma_\rho)$, $\varepsilon_i \sim N(0, \sigma^2)$, and $y_i = X_i^T \beta + \varepsilon_i$. Predictors and errors are independent. The factorial design consists of 24 cells with $n \in \{50,100,200\}$, $\rho \in \{0,0.50,0.90,0.99\}$, and $\sigma \in \{1,2\}$. Each cell is replicated 1,000 times. The true coefficients are $\beta = (1,1,-0.5,0,0,0)^T$ and the true intercept is zero. Thus, the alternative is known to satisfy $\beta_1 = 1$ for the first coefficient; the reported rejection proportion for its t-test is its empirical *statistical power*.

The variance of the explainable signal is unchanged as correlation among the predictors increases:

$$\mathrm{Var}(X_i^T \beta) = \beta^T \Sigma_\rho \beta = (1-\rho) \parallel \beta \parallel^2 + \rho(\mathbf{1}^T \beta)^2 = 2.25.$$

Consequently, the population $R^2 = 2.25/(2.25 + \sigma^2)$ is 0.6923 for $\sigma = 1$ and 0.3600 for $\sigma = 2$. This design eliminates the risk of conflating the effect of multicollinearity with the changing R^2 observed in earlier pilot analyses. In-sample R^2 can differ from the population R^2 because of sampling uncertainty and the number of predictors included.

Computed Statistics

In every replication, six slopes and an intercept are estimated by OLS/SVD. The calculations include the standard error of β_1, an indicator of rejection in the individual two-sided t-test, an indicator of whether the 95% confidence interval contains the true coefficient, the overall F-test, R^2, adjusted R^2, maximum VIF, and the condition index. Cell-level means or proportions are then reported. The residual degrees of freedom for OLS standard errors are $n - p - 1$. "Mean maximum VIF" is computed by first taking the maximum of the six VIFs in each replication and then averaging these maxima across replications.

The analysis uses 95% Wilson intervals for proportions and Monte Carlo standard errors for means of continuous quantities. Confidence interval coverage near 95% does not mean that a particular realized interval contains the fixed true coefficient with a posterior probability of 95%; it refers to the long-run coverage frequency over repeated samples. Because the true parameters are known in the simulation, this frequency can be measured directly. All calculations use the fixed root seed 20261018, and the generators for separate factorial cells are separated through seed sequences. The covariance matrix is sampled using Cholesky decomposition, and SVD-based solutions are preferred for coefficient estimation and standard-error calculations. Direct inversion of $X^T X$ is avoided because it can introduce unnecessary rounding error under severe collinearity. The design does not include exact collinearity ($\rho = 1$) or high-dimensional cases with $p \geq n$. Findings are conditional on Gaussian predictors and independent homoskedastic Gaussian errors.

MONTE CARLO RESULTS

Increasing VIF and Condition Indices

Table 1 shows how the diagnostic measures change as the strength of collinearity increases when $n = 50$ and $\sigma = 1$. At $\rho = 0$, the population VIF is 1.00 and the mean sample maximum VIF is 1.21. At $\rho = 0.99$, these quantities rise to 83.36 and 126.03, respectively. The mean sample condition index also increases from 1.64 to 32.91. Thus, the implications of dependence among the predictors for regression variance and for numerical conditioning change dramatically.

Table 1. Multicollinearity diagnostics and inference for β_1 ($n = 50$, $\sigma = 1$; $B = 1{,}000$).

ρ	Population VIF	Mean max. VIF	Mean condition index	Mean SE(β_1)	t-test power
0.00	1.00	1.21	1.64	0.153	100.0%
0.50	1.71	2.38	3.50	0.200	99.4%
0.90	8.36	12.22	9.83	0.446	59.5%
0.99	83.36	126.03	32.91	1.405	10.8%

In an equicorrelation structure, every pairwise correlation is equal, so pairwise correlations also reveal strong dependence. In real applications, however, multivariate dependencies may be present even when all pairwise correlations are relatively low. VIF and singular values should therefore be interpreted together with knowledge of what the explanatory variables represent.

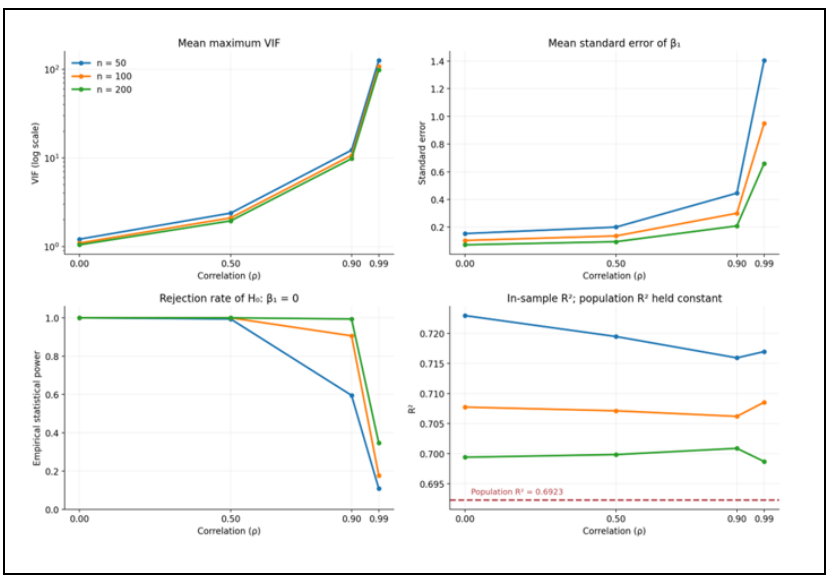
Standard Errors and Test Power

The true value $\beta_1 = 1$ is fixed across all experimental cells. Nevertheless, at $n = 50$ and $\sigma = 1$, increasing ρ from zero to 0.99 raises the mean standard error from 0.153 to 1.405, while the rejection rate for $H_0: \beta_1 = 0$ at the two-sided 5% level falls from

100% to 10.8%. Low power here does not indicate an incorrect model. Rather, extreme collinearity reduces the effective information available about an individual parameter. Even when the mean of the estimates across samples stays near the truth, the coefficient estimate in any one dataset can vary over a wide range.

The same phenomenon can be assessed through a different metric: with $n = 50$ and $\sigma = 1$, the estimated OLS coefficient $\hat{\beta}_1$ is never negative in the 1,000 replications when $\rho = 0$, but has a negative sign in approximately 25% of replications when $\rho = 0.99$. This does not imply that the true effect is negative; it reflects the sampling distribution under severe collinearity and noise. Assigning too much substantive meaning to the sign observed in a single sample can therefore be problematic.

Figure 1. Comparison of VIF, standard error, t-test power, and sample R^2 across sample sizes.



Note. Population R^2 is shown as a dashed red line and is constant across ρ. Power denotes the empirical rejection rate of the two-sided t-test for β_1.

The Mitigating Effect of Sample Size

Larger samples can reduce estimation uncertainty under the same covariance structure. Table 2 compares $n = 50$, 100, and 200 for $\rho = 0.99$ and $\sigma = 1$. Mean standard errors are 1.405, 0.950, and 0.659, respectively, while test power rises from 10.8% to 17.7% and 34.6%. Thus, quadrupling the number of observations does not eliminate the impact of severe collinearity, but it mitigates it to some extent.

Table 2. Effect of sample size under severe collinearity ($\rho = 0.99$, $\sigma = 1$).

n	Mean max. VIF	Mean condition index	SE(β_1)	t-test power	CI coverage
50	126.03	32.91	1.405	10.8%	95.1%
100	108.03	29.76	0.950	17.7%	96.1%
200	98.33	27.84	0.659	34.6%	95.0%

These findings support reporting collinearity diagnostics jointly with sample size. The same VIF or condition index need not produce the same standard error or test power in small and large samples. The sample size required also depends on the true coefficient magnitude and the noise variance.

Confidence Interval Coverage and Length

Under the classical assumptions and with $n = 50$, $\sigma = 1$, empirical coverage of the 95% confidence interval is 96.5%, 95.7%, 95.2%, and 95.1% across the four correlation levels. These figures are near the nominal 95% target; severe collinearity has not automatically destroyed interval coverage. However, the larger mean standard error raises the mean full interval length from approximately 0.618 to 5.668 (using the $t_{0.975,43}$ multiplier). This widening means that the same coefficient can be learned with substantially less precision.

A wide confidence interval for a particular sample does not mean that the analysis is broken; it may mean that the design provides insufficient information to distinguish the parameter precisely. Before searching for an alternative method that happens to produce statistical significance, the data-collection plan, variable definitions, and scientific estimation objective should be reviewed.

Table 3. Confidence interval coverage, interval length, and R^2 ($n = 50$, $\sigma = 1$).

ρ	CI coverage	95% Wilson CI for coverage	Mean CI length	Mean R^2	Population R^2
0.00	96.5%	[95.2%, 97.5%]	0.618	0.723	0.692
0.50	95.7%	[94.3%, 96.8%]	0.806	0.719	0.692
0.90	95.2%	[93.7%, 96.4%]	1.797	0.716	0.692
0.99	95.1%	[93.6%, 96.3%]	5.667	0.717	0.692

Overall F-test, R^2, and Individual T-tests

In every $n = 50$, $\sigma = 1$ experimental cell, the overall F-test rejects in 100% of replications. Yet at $\rho = 0.99$ the rejection rate of the β_1 t-test is only 10.8%. The reason is that the overall F-test examines whether all six slopes are jointly zero, whereas the t-test examines whether the first slope is zero conditional on the other predictors. This does not imply that the overall F-test must always be significant in every dataset; the finding is conditional on the shared signal strength and sample size in this experiment.

Across the four correlation levels, population R^2 remains fixed at 0.6923, while mean in-sample R^2 is approximately 0.723, 0.719, 0.716, and 0.717. These small fluctuations do not constitute evidence that correlation by itself increases explanatory power. Adjusted R^2 also remains near 0.68. The defensible conclusion is

that "a high R^2 does not guarantee that every coefficient is well identified," not that "R^2 necessarily increases with collinearity."

What Happens When the Noise Level Increases?

When the error standard deviation rises from 1 to 2, the error variance is quadrupled and population R^2 falls from 0.6923 to 0.3600. With $n = 50$ and $\rho = 0.99$, the mean standard error increases from 1.405 to 2.792, while power for the β_1 test falls from 10.8% to 6.4%. Even at $\rho = 0$, with no population collinearity, higher noise reduces power from 100% to 89.7%. Collinearity should therefore be interpreted alongside effect magnitude and noise level, not in isolation.

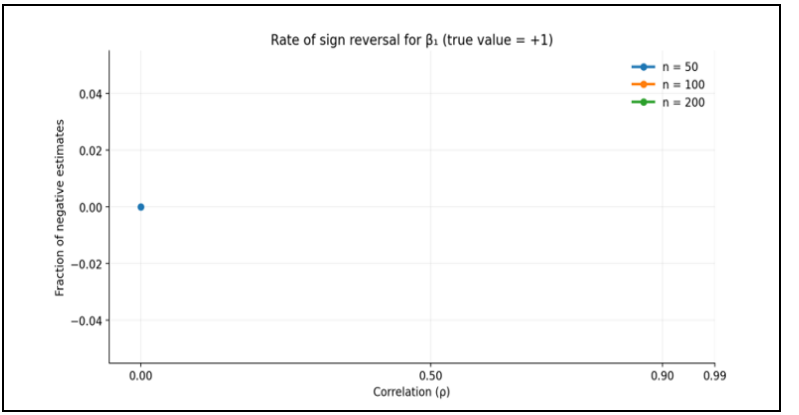
Table 4. Effects of the noise standard deviation on inference ($n = 50$).

ρ	SE(β_1), $\sigma = 1$	SE(β_1), $\sigma = 2$	t-power, $\sigma = 1$	t-power, $\sigma = 2$
0.00	0.153	0.305	100.0%	89.7%
0.50	0.200	0.396	99.4%	69.0%
0.90	0.446	0.886	59.5%	20.3%
0.99	1.405	2.792	10.8%	6.4%

Stability of Coefficient Signs

Because OLS estimates vary across samples, an additional descriptive measure is the frequency with which a coefficient estimate has the opposite sign to its nonzero true value. At $n = 50$, $\sigma = 1$, and $\rho = 0.90$, $\hat{\beta}_1$ is negative in 1.5% of replications; at $\rho = 0.99$, the rate rises to 25%. This observation concerns only the selected first coefficient with known true value $\beta_1 = 1$. It would not be correct to attribute every sign reversal to collinearity alone, because sampling noise also plays a role.

Figure 2. Frequency of sign reversal for the first coefficient by correlation level and sample size.



Note. Rates indicate the replications in which the estimated first slope is negative despite a positive true β_1.

REAL-DATA APPLICATION: LONGLEY (1947–1962)

Data and purpose

In the simulation, the true coefficients and the data-generating mechanism are known. Such information is unavailable in real data. Accordingly, the second application is not intended to provide a causal answer to the question of what multicollinearity causes; rather, it examines what diagnostic tools reveal about an actual design matrix. The Longley (1967) dataset contains 16 annual US macroeconomic observations covering 1947–1962 and six explanatory variables (Longley, 1967). Total employment

(TOTEMP) is the response variable. The explanatory variables are the gross national product deflator (GNPDEFL), gross national product (GNP), number of unemployed persons (UNEMP), size of the armed forces (ARMED), population (POP), and year (YEAR). This dataset is also one of the NIST reference problems used to test numerical least-squares algorithms (Longley, 1967).

The model is an ordinary least-squares linear regression of $TOTEMP$ on the six explanatory variables and an intercept. Estimates were checked using numerically stable SVD/QR solution methods and their agreement with the certified coefficients published by NIST was verified. With $n = 16$ and $p = 6$, there are only nine residual degrees of freedom. This is an independent constraint that increases the sensitivity of statistical inference.

Correlation Matrix and VIF

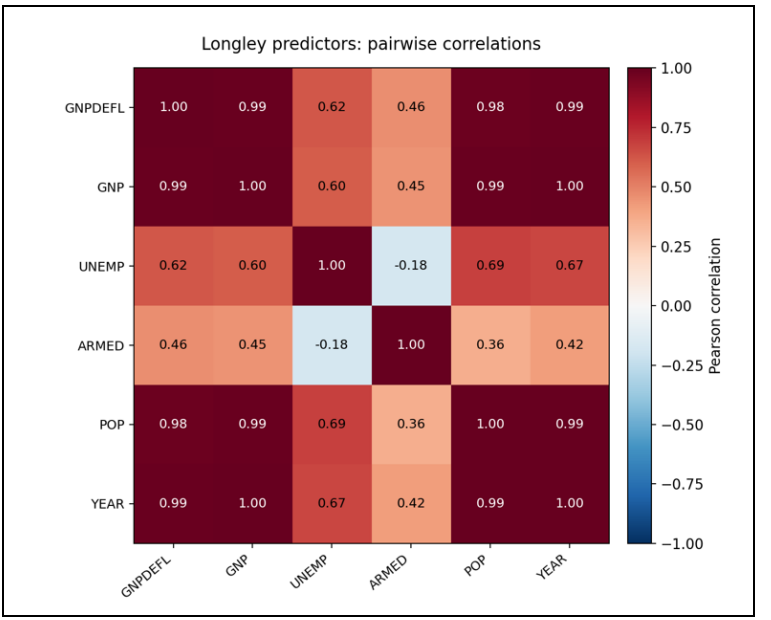
In the correlation matrix of the Longley explanatory variables, several macroeconomic series move strongly in the same direction. For example, GNP, YEAR, and POP can exhibit very similar trends. Yet the pairwise correlation matrix alone does not reveal all higher-order linear dependencies. The maximum VIF in the fitted model is 1,788.51 (GNP); the VIFs for POP and YEAR are approximately 399.15 and 758.98, respectively. The condition index of the centered and standardized X matrix is 110.54. These values provide strong diagnostic evidence of severe collinearity in the regression design.

OLS Coefficients and Significance Tests

For the Longley regression, the in-sample R^2 is 0.995479, the adjusted R^2 is approximately 0.992465, and the conventional overall F-statistic is approximately 330.29. Such a high model fit does not imply that every individual coefficient has a narrow confidence interval. Table 5 reports the six slope coefficients, their conventional

OLS standard errors, two-sided p-values, and corresponding VIFs. For instance, the estimated coefficient on GNP is approximately -0.0358, with a standard error of 0.0335, a p-value of 0.313, and a VIF of 1,788.51. The coefficient on POP is likewise not significant at the conventional 5% level in this model ($p = 0.826$).

Figure 3. Pairwise correlation matrix of the Longley explanatory variables.



Note. The correlation matrix displays pairwise relationships only; VIFs and the condition index also reflect more general linear dependencies.

Table 5. Multiple-regression OLS coefficients and VIFs for the Longley data.

Variable	Estimated coefficient $\widehat{\boldsymbol{\beta}}$	Conventional SE	p-value	VIF
GNPDEFL	15.062	84.915	0.863	135.53
GNP	-0.036	0.033	0.313	1788.51
UNEMP	-2.020	0.488	0.003	33.62
ARMED	-1.033	0.214	<0.001	3.59
POP	-0.051	0.226	0.826	399.15
YEAR	1829.151	455.478	0.003	758.98

The sensitivity of coefficient signs and magnitudes for variables such as GNPDEFL and POP precludes strong conclusions about causal signs or the underlying economic effects. Small p-values for other variables likewise do not establish macroeconomic cause-and-effect relationships. These are observational data; common time trends, simultaneity, omitted variables, and possible structural changes may challenge the classical model. The Longley application should therefore be interpreted as a *demonstration of diagnostics and numerical stability*.

Leave-One-Observation-Out Sensitivity

As an additional real-data analysis, each of the 16 observations was removed in turn, and the same regression model was re-estimated using the remaining 15 observations. This is not an evaluation of predictive performance; it is a simple diagnostic of coefficient stability. In the full-data model, the GNPDEFL coefficient is approximately +15.06, but after deleting one observation at a time it ranges from approximately −52.36 to +40.78; the sign reverses in two of the 16 refits. The POP coefficient is −0.051 in the full dataset; its deletion-based range is approximately −0.192 to +0.166, with sign reversals in three refits. Table 6 presents the results for all explanatory variables.

Table 6. Leave-one-observation-out coefficient sensitivity in the Longley data.

Variable	Full-data coefficient	Leave-one-out range	Sign reversals
GNPDEFL	15.062	[−52.357, 40.781]	2/16
GNP	-0.036	[−0.084, −0.021]	0/16
UNEMP	-2.020	[−2.698, −1.747]	0/16
ARMED	-1.033	[−1.256, −0.904]	0/16
POP	-0.051	[−0.192, 0.166]	3/16
YEAR	1829.151	[1432.125, 2583.579]	0/16

This sensitivity illustrates that coefficient interpretation can be fragile when strong collinearity, a small sample, and potentially influential observations coexist. However, a coefficient's response to deleting one observation is not, by itself, proof that multicollinearity caused the change: observation leverage and model specification may also matter. The sensitivity analysis does not separate these influences.

Numerical Validation and Model Assumptions

In NIST's Longley reference output, the certified GNPDEFL coefficient is 15.06187227, the GNP coefficient is −0.03581918, and model R^2 is 0.99547900. The calculations reported here agree with those numerical references to the stated rounding precision. This agreement means that the algorithm gives the correct numerical solution to this problem; it does not establish that the regression model is correctly specified from an economic perspective.

Conventional standard errors and p-values rely on assumptions such as independence and constant variance of the errors. Given the time-series nature of the Longley data, those assumptions require separate scrutiny. The calculated Durbin–Watson statistic is approximately 2.56; that single statistic does not prove the absence of autocorrelation in a small, highly collinear

model. Statistical significance findings should therefore be interpreted cautiously unless the data are re-examined using appropriate time-series methods.

INTEGRATED DISCUSSION OF THE FINDINGS

The simulation and the Longley application provide different kinds of information. In the simulation, the true β and the error model are known; consequently, increasing standard errors and declining t-test power can be linked to changes in the degree of collinearity under controlled conditions. Such a controlled experiment cannot be carried out on the real dataset. Instead, the real-data analysis yields diagnostics of strong collinearity and indications of sensitivity to sample perturbations. It would therefore be methodologically incorrect to present the causal mechanism demonstrated in the simulation as a definitive explanation of all coefficient behavior in the Longley data.

The central lesson of the study is the need to distinguish three performance criteria. The model's R^2 measures overall in-sample fit; a coefficient's t-statistic and confidence interval indicate how much information is available about an individual effect; and prediction error measures how accurately responses can be predicted for new observations. These measures can complement one another, but they do not answer the same statistical question. Under multicollinearity, high overall fit, low power of individual tests, and preserved predictive performance under certain distributions can coexist.

Purpose and context are equally important when interpreting diagnostics. A VIF greater than 10 may warrant attention, but it does not automatically mandate deleting a variable. Deletion may introduce omitted-variable bias and may change the scientific question being investigated. Physical or theoretical restrictions, new measurements, alternative experimental designs, theoretically meaningful composite variables, principal-component summaries, or prediction-oriented methods such as ridge should be considered according to the research objective. Whatever strategy is chosen, interpretability and external validity must be reassessed (O'Brien, 2007) (Dormann, 2013).

The Gauss–Markov property of the classical linear model cannot be reduced to the claim that OLS is unusable whenever predictors are correlated. Equally, the view that an unbiased OLS estimator presents no problem is incomplete. When the research interest concerns the sign, magnitude, or significance of an individual coefficient, high variance is a consequential practical limitation. Conversely, dropping or penalizing variables that predict jointly can be useful for forecasting but may change the original scientific interpretation of the coefficients.

Limitations

The first limitation is that the principal simulation assumes Gaussian predictors and disturbances, homoskedasticity, a linear conditional mean, exogeneity, and full column rank. Conditions involving heteroskedastic or correlated errors, the high-dimensional case $p \geq n$, and measurement error in explanatory variables are not covered. In addition, only one principal β vector and one family of equicorrelation matrices are used. Although alternative coefficient and covariance structures were considered for robustness in the companion study, the controlled comparison in this report is restricted to the specified primary data-generating process.

The second limitation is the finite number of Monte Carlo replications and the fact that some rejection rates are close to zero or one. Wilson confidence intervals can be constructed for such rates; for example, rejection in all 1,000 replications does not prove that the true statistical power is exactly one. Coverage rates close to the nominal level likewise do not automatically imply that a model is properly specified for real-world data.

The third limitation concerns the Longley dataset's 16 observations, common time trends, and only nine residual degrees of freedom. Leave-one-out sensitivity does not systematically encompass other sampling models. High VIFs and a large condition index do not, by themselves, identify an economic explanation or a causal effect. The study collects no new personal data and relies on publicly available or synthetic datasets.

CONCLUSION

The most conspicuous effect of multicollinearity in this study is that it weakens the information provided by classical OLS estimates *about separate regression coefficients*. In the controlled primary experiment, as ρ rises to 0.99, the standard error of β_1 increases by approximately a factor of nine and the t-test's power to detect the true positive effect decreases sharply. Nevertheless, coverage of the 95% confidence interval stays close to the nominal level, and no mechanically necessary rise in R^2 is observed when the population signal-to-noise ratio is held fixed. The Longley analysis shows that high VIFs and a high condition index can coexist with

sensitivity of individual coefficients to the deletion of just one observation.

A defensible sequence for applied work is as follows. First, determine whether the research question concerns coefficient interpretation or prediction. Then assess the correlation matrix, VIFs, and condition indices in their diagnostic context. Next examine the substantive meaning of the variables, the observation plan, and model assumptions. If necessary, consider collecting new data, combining variables on theoretical grounds, or using a regularization method. Finally, report coefficient uncertainty, independent validation, and exposure to distribution shift explicitly. This procedure is more robust than mechanical choices based solely on a high R^2 or one VIF threshold.

References

Belsley, D. A. (1980). *Regression Diagnostics: Identifying Influent Data and Sources of Collinearity.* John Wiley & Sons. doi:10.1002/0471725153

Dormann, C. F. (2013). *Collinearity: A review of methods to deal with it and a simulation study evaluating their performance* (Cilt 36). Ecography. doi:10.1111/j.1600-0587.2012.07348.x

Longley. (1967). Statistical Reference Datasets (StRD). *Certified regression statistics*. NIST/SEMATECH. July 27, 2026 tarihinde alındı

Montgomery, D. C. (2021). *Introduction to Linear Regression Analysis (6th ed.)*. Wiley.

O'Brien, R. M. (2007). *A caution regarding rules of thumb for variance inflation factors* (Cilt 41). Quality & Quantity. doi:10.1007/s11135-006-9018-6

CHAPTER 2

BAYESIAN HIERARCHICAL REGRESSION FOR MULTIGROUP DATA: SHARED AND GROUP-SPECIFIC EFFECTS THROUGH PARTIAL POOLING

FATİH SAĞLAM[1] BURÇİN ŞEYDA ÇORBA[2]

Introduction

Multigroup data raise a question that a single regression equation cannot resolve on its own: how much information should be shared across groups? Observations collected within demographic categories, institutions, or clinical subpopulations may support both common associations and departures from those associations. The central modelling decision is therefore whether regression parameters should be shared, estimated separately, or linked through a distribution that permits variation. Multilevel modelling addresses this decision by modelling the regression coefficients themselves (Gelman, 2006a).

[1] Dr., Ondokuz Mayıs University, Department of Statistics, ORCID: 0000-0002-2084-2008
[2] Dr., Ondokuz Mayıs University, Department of Statistics, ORCID: 0000-0002-0085-4351

Complete pooling assigns a common regression relationship to all groups, whereas no pooling estimates a separate relationship for each group. These strategies impose different assumptions about how informative one group is about another. A common relationship can obscure meaningful differences; separate estimation can produce unstable estimates when a group contributes little information. Partial pooling offers an intermediate approach in which group-specific coefficients are estimated jointly and drawn towards a shared distribution. Gelman's (2006a) radon example illustrates why this balance matters when group sample sizes differ substantially.

Predictive evaluation provides a complementary basis for comparing pooling strategies. Leave-one-out cross-validation evaluates predictive accuracy for held-out observations, and Pareto smoothed importance sampling can approximate this evaluation from posterior draws (Vehtari et al., 2017). Its Pareto-k diagnostic assesses the reliability of the importance-sampling approximation rather than establishing that the regression model is substantively adequate (Vehtari et al., 2024). Posterior predictive checks address a different question: whether data replicated under the fitted model reproduce relevant features of the observations. Bayesian multilevel software supports both posterior prediction and pointwise likelihood calculations, allowing these assessments to be considered together (Bürkner, 2017).

The vocabulary of shared and group-specific effects also appears in joint and individual component methods. Kara et al. (2026), for example, develop Bayesian joint and individual component regression for multigroup physiological data. That framework represents joint and individual component structure; the hierarchical regression considered here instead links regression coefficients through a group-level distribution. Their common interest in heterogeneity does not make them equivalent methods. The present chapter therefore uses Bayesian hierarchical regression

as its methodological designation and reserves the component-method literature for a distinct discussion of related approaches.

This chapter develops the conceptual and methodological basis for comparing complete pooling, no pooling, and partial pooling, followed by an illustration of systolic blood pressure in 3,885 adults from the NHANES R package. Five race/ethnicity groups and seven predictors are considered. Gaussian pooling strategies are compared with a Student-t hierarchical extension, with attention to predictive accuracy, coefficient shrinkage, prior sensitivity, and computational limitations. The illustration estimates conditional associations in an educational survey extract; it does not establish causal effects or nationally representative population quantities.

Heterogeneity in Multigroup Data

Multigroup heterogeneity concerns differences in the processes relating predictors to outcomes, not simply differences in observed group averages. A group with a higher average outcome may contain individuals with different covariate values, have a different conditional outcome distribution, or display both features. For this chapter, heterogeneity is considered at three distinct levels: the composition of the groups, their conditional regression relationships, and the residual variation left after conditioning on the predictors. This distinction provides a conceptual framework for specifying what a hierarchical regression should allow to vary; it does not constitute an empirical finding about the application.

Differences in group composition arise when predictor distributions vary across groups. Consequently, marginal outcome differences need not imply different intercepts or slopes once covariates are taken into account. A comparison of regression relationships must therefore specify the covariates on which it conditions and the range over which groups can reasonably be

compared. Where covariate ranges overlap little, comparisons may depend heavily on extrapolation. The distinction between prediction within familiar predictor ranges and prediction outside them is also relevant when choosing a validation design (Roberts et al., 2017).

Conditional heterogeneity can be represented through varying intercepts, varying slopes, or both. An intercept difference describes a difference in the expected response at the chosen predictor reference point. Its interpretation therefore depends on how predictors are coded and centred. A slope difference describes variation across groups in a conditional predictor–response association. In a varying-intercept and varying-slope model, the intercept and slope deviations may also be correlated. These features correspond to group-level terms in a multilevel regression, while shared coefficients describe the common level around which group-specific relationships vary (Bürkner, 2017).

Residual heterogeneity is conceptually distinct from heterogeneity in the conditional mean. Groups may share similar regression relationships while differing in their unexplained dispersion, or may have different mean relationships but comparable residual dispersion. Accordingly, allowing intercepts and slopes to vary does not automatically account for group-specific residual scales or heavy tails. The likelihood and its distributional parameters must be considered separately from the group-level coefficient structure. Bayesian multilevel frameworks permit different response distributions and, where specified, distributional extensions; their suitability depends on the outcome and the scientific question (Bürkner, 2017). The application below considers Student-t errors while retaining a common residual scale across groups.

The amount of information available within each group is equally important. Group sample size, covariate variability, and the outcome distribution affect how precisely a group-specific relationship can be estimated. Apparent differences between

independently estimated coefficients may therefore reflect unequal uncertainty as well as substantive heterogeneity. In their discussion of multilevel inference, Gelman et al. (2012) explain how partial pooling moves estimates towards a common model and adapts to the information available. Thus, a small or imprecisely estimated group should not be treated as providing the same evidence about heterogeneity as a group with a well-determined relationship.

Information sharing nevertheless requires a defensible grouping structure. A hierarchical distribution expresses a model of similarity among group-specific coefficients conditional on the included predictors and any group-level information. It does not establish that the groups are interchangeable in every substantive respect. Group categories should therefore be defined by the research question and the data-generating context rather than selected solely because they yield a convenient regression formula. For demographic categories, estimated differences describe conditional associations under that categorisation; they do not, by themselves, identify biological mechanisms or causal effects. The potential for causal misinterpretation in multilevel models is illustrated by Gelman (2006a).

A further distinction concerns the target of inference. Describing conditional relationships in the observed sample differs from estimating quantities for a survey target population. Gelman (2007) discusses why regression modelling and survey weighting must be connected to inclusion probabilities, nonresponse, and population information rather than treated as unrelated choices. Demographic modelling groups must also be distinguished from the clusters and strata used in the sampling design. Defining a demographic hierarchy does not itself resolve survey weighting or establish population representativeness. The application below uses an educational extract without design-based weighting, and its interpretation is restricted accordingly.

These considerations motivate a structured comparison of pooling strategies. The comparison should distinguish covariate composition from conditional parameter variation, recognise unequal information across groups, and make explicit which parameters are shared. It should also preserve the distinction between estimating relationships for represented groups and predicting outcomes in an unobserved group. The next section develops complete pooling, no pooling, and partial pooling as alternative assumptions about the sharing of information, providing the basis for the subsequent hierarchical model formulation.

Complete Pooling, No Pooling, and Partial Pooling

Pooling describes how information is shared in estimating a specified set of parameters. It is not simply a decision about whether observations are stored in one dataset or several. A single fitted model may share some coefficients while allowing others to vary, and separate files do not necessarily imply independent statistical estimation. The three strategies considered below concern the intercepts and slopes of group-specific regression relationships. Their contrasting assumptions provide the basis for the model comparison, rather than a presumption that one strategy will always perform best (Gelman, 2006a; Gelman et al., 2012).

For illustration, let y_{ij} denote the response for individual i in group j, x_{ij} a single predictor, n_j the number of observations in group j, and J the number of groups. The index j runs from 1 to J, and i runs from 1 to n_j. A Gaussian response model is used to explain the pooling structures; it is not a commitment to the likelihood for the empirical application. In the notation below, N(m, v) denotes a normal distribution with mean m and variance v. Residual-scale modelling is a separate decision from coefficient pooling. Conditional on the predictors and model parameters, observations

are assumed independent in these illustrative likelihoods; all residual standard deviations are positive.

Complete Pooling

$$y_{ij} \mid x_{ij}, \alpha, \beta, \sigma \sim N(\alpha + \beta x_{ij}, \sigma^2). \quad (1)$$

In Equation (1), α is the common intercept, β is the common slope, and σ is the residual standard deviation. Every group has the same conditional mean relationship. Thus, after conditioning on x, the group label makes no additional contribution to the predicted mean. All observations inform the shared coefficients, and groups containing more information contribute more strongly to their estimation. Complete pooling can be adequate when relevant group differences are negligible or captured by other covariates; it can be restrictive when systematic intercept or slope differences remain (Gelman, 2006a).

Including a group indicator in an otherwise ordinary regression already changes this specification. A model with group indicators and a common slope permits intercept differences while pooling the slope. Accordingly, the complete-pooling benchmark in this chapter refers to a common intercept and common slopes for the coefficients under comparison. This definition is applied consistently to the seven predictors in the application.

No Pooling

$$y_{ij} \mid x_{ij}, \alpha_j, \beta_j, \sigma_j \sim N(\alpha_j + \beta_j x_{ij}, {\sigma_j}^2). \quad (2)$$

In Equation (2), the group-specific intercept α_j, slope β_j, and residual standard deviation σ_j are estimated from the observations in group j. In a Bayesian no-pooling formulation, independent, prespecified priors can regularise each group's coefficients without linking them through learned common hyperparameters. Such within-group regularisation is not partial pooling. Separate estimation avoids forcing groups to share a relationship, but may

yield weakly identified or highly uncertain slopes when a group has few observations or little predictor variation (Gelman et al., 2012).

No pooling defines an estimated relationship for each represented group. It does not supply a rule for an entirely new group unless an additional model is introduced. If all groups instead share an estimated residual scale, some information is pooled through that nuisance parameter even when the regression coefficients are unpooled. This distinction matters when describing the fitted models: coefficient pooling and residual-scale pooling should be reported separately, and nuisance specifications should be aligned where appropriate for the intended comparison.

Partial Pooling

$$y_{ij} \mid x_{ij}, \alpha_j, \beta, \sigma \sim N(\alpha_j + \beta x_{ij}, \sigma^2). \quad (3)$$

$$\alpha_j \mid \mu_\alpha, \tau_\alpha \sim N(\mu_\alpha, {\tau_\alpha}^2). \quad (4)$$

The varying-intercept model in Equations (3) and (4) shares the slope β while allowing the intercepts to vary. The hyperparameter μ_α is the mean of the group-intercept distribution, and τ_α is its between-group standard deviation. The intercepts are conditionally independent draws from that distribution given the hyperparameters. Because the hyperparameters are inferred jointly using all groups, the posterior for one group can be informed by other groups. This is the mechanism of partial pooling. Varying slopes extend the same principle to β_j, with a joint intercept–slope distribution where appropriate (Bürkner, 2017).

The shared distribution defines conditional exchangeability at the coefficient level: after accounting for the modelled information, group labels do not change the assumed distribution of the deviations. Exchangeability therefore concerns the model, not a claim that demographic groups are substantively identical. When group-level predictors are included, pooling can occur towards

different fitted group-level means rather than a single overall mean (Gelman et al., 2012).

A Mathematical Illustration of Shrinkage

The following conjugate calculation is an analytical illustration developed for this chapter, not an analysis of NHANES. Suppose an approximately normal estimate $\hat{a}_j$ of a scalar group coefficient a_j has known sampling variance v_j, and a_j has a normal prior with mean μ and variance τ^2. Treat μ and τ as fixed in this calculation. Multiplying the normal likelihood and prior gives the conditional posterior moments below. Equations (5) and (6) define the illustrative likelihood and prior; Equations (7)–(9) give the conditional posterior mean, shrinkage weight, and variance. The sampling variance v_j is positive and treated as known.

$$\hat{a}_j \mid a_j, v_j \sim N(a_j, v_j). \tag{5}$$

$$a_j \mid \mu, \tau \sim N(\mu, \tau^2). \tag{6}$$

$$E\left(a_j \mid \hat{a}_j, \mu, \tau, v_j\right) = w_j \hat{a}_j + \left(1 - w_j\right)\mu. \tag{7}$$

$$w_j = \frac{\tau^2}{\tau^2 + v_j}. \tag{8}$$

$$Var(a_j \mid \hat{a}_j, \mu, \tau, v_j) = \frac{\tau^2 v_j}{\tau^2 + v_j}. \tag{9}$$

The weight w_j lies between zero and one for positive finite variances. Larger sampling variance v_j gives less weight to the individual group estimate and more weight to μ. Larger between-group variance τ^2 permits more separation between groups. For fixed v_j, the limit $\tau^2 \to 0$ collapses the coefficient to μ; the limit $\tau^2 \to \infty$ recovers the normal-likelihood estimate $\hat{a}_j$. These limits explain the pooling continuum in this scalar example. They are not a general assertion that an arbitrarily diffuse hierarchical prior is equivalent to every possible no-pooling regression.

In a full hierarchical analysis, μ and τ are uncertain and must be integrated over their posterior distribution. The conditional expression above therefore does not provide the full posterior variance, and its apparent gain in precision should not be interpreted without that integration. Likewise, correlated intercepts and slopes, unequal designs, and non-Gaussian likelihoods generally prevent a simple componentwise weighted-average formula. Shrinkage is best understood as joint probabilistic estimation rather than an adjustment applied after fitting independent regressions.

Implications for Model Comparison

A fair comparison should preserve the outcome definition, analysis sample, predictor coding, and prediction target while making the intended differences in parameter sharing explicit. Proper priors for the Bayesian benchmarks and suitable likelihood assumptions remain necessary even when a model does not pool coefficients. Predictive accuracy, its uncertainty, and posterior predictive behaviour should be evaluated rather than ranking models by their in-sample fit alone (Vehtari et al., 2017). Partial pooling can stabilise estimation, but its practical value depends on whether the assumed hierarchy adequately represents the groups. The next section develops varying-intercept and varying-slope models in a unified regression formulation.

Bayesian Hierarchical Regression: Varying Intercepts and Slopes

A varying-intercept and varying-slope regression allows the conditional mean relationship to differ across groups while linking those relationships through shared hyperparameters. The formulation below extends the varying-intercept illustration in Equations (3) and (4). It uses a single predictor for transparency; additional predictors can have shared or varying coefficients according to the scientific question. A Gaussian likelihood is

retained as a methodological example rather than an application-specific choice (Bürkner, 2017).

$$y_{ij} \mid x_{ij}, \alpha_j, \beta_j, \sigma \sim N(\alpha_j + \beta_j x_{ij}, \sigma^2). \quad (10)$$

$$\alpha_j = \mu_\alpha + u_{\alpha j}, \quad \beta_j = \mu_\beta + u_{\beta j}. \quad (11)$$

$$\boldsymbol{u}_j = \begin{pmatrix} u_{\alpha j} \\ u_{\beta j} \end{pmatrix}, \quad \boldsymbol{u}_j \mid \sum \sim N_2(0, \sum). \quad (12)$$

$$\sum = \begin{pmatrix} {\tau_\alpha}^2 & \rho\tau_\alpha\tau_\beta \\ \rho\tau_\alpha\tau_\beta & {\tau_\beta}^2 \end{pmatrix}. \quad (13)$$

In Equation (10), α_j and β_j are the intercept and slope for group j, and $\sigma > 0$ is the common residual standard deviation. Equation (11) separates each coefficient into a shared mean and a group deviation: μ_α and μ_β are the means of the coefficient distribution, while $u_{\alpha j}$ and $u_{\beta j}$ are its zero-mean deviations. N_2 denotes the bivariate normal distribution. In Equations (12) and (13), Σ is the covariance matrix of the group deviations, τ_α and τ_β are their positive standard deviations, and ρ is their correlation, with $-1 < \rho < 1$ for a positive-definite covariance matrix. Conditional independence across groups is assumed for these deviations given Σ; outcomes are conditionally independent given the group coefficients, predictors, and residual scale.

The shared slope μ_β describes the mean of the specified group-coefficient distribution. It is not automatically the slope obtained by fitting one ordinary regression to all individuals, nor an individual-weighted or survey-weighted average slope. Those quantities depend on covariate distributions, group sizes, the sampling design, and the definition of the estimand. Similarly, a group deviation is an estimated departure from the shared model, not an error term to be discarded. Its uncertainty is part of the posterior inference (Bürkner, 2017).

Interpreting Intercept–Slope Dependence

A positive value of ρ indicates that groups with higher intercept deviations tend to have higher slope deviations under the model; a negative value indicates the reverse. The interpretation is tied to the predictor reference point. An intercept is the expected response at x = 0, so a reference point outside the observed range can make intercept comparisons difficult to interpret. Centring at a common, substantively meaningful value can improve interpretation. However, centring does not eliminate the need to assess slope information within each group.

$$x'_{ij} = x'_{ij} - c, \quad \alpha'_j = \alpha_j + c\beta_j. \tag{14}$$

$$Cov(\alpha', \beta) = \rho\tau_\alpha\tau_\beta + c\tau_\beta^2. \tag{15}$$

Equations (14) and (15) are direct algebraic consequences of changing the predictor origin by a fixed constant c; they are not empirical results. The slope is unchanged, but the intercept and its covariance with the slope change. Thus, the intercept–slope correlation is not invariant to centring. Group-mean centring is a different transformation: it can separate within-group predictor variation from between-group differences when the group means are also modelled. It changes the interpretation of coefficients and should not be introduced without defining the intended association. A change of coordinates preserves the original probabilistic model only when its prior is transformed consistently; assigning new independent priors after centring can change that model.

Model Structure and Prediction Targets

The illustrative R formula y ~ x + (1 + x | group) represents a shared intercept and slope together with correlated group deviations in both coefficients. The formula describes a structure; it does not specify a complete Bayesian analysis. A likelihood, priors, coding decisions, and computational settings must also be supplied.

A varying-intercept model omits the varying slope, while a model with uncorrelated deviations fixes the corresponding correlation to zero. Those restrictions should be justified and evaluated rather than adopted solely to obtain easier computation (Bürkner, 2017).

For a represented group, posterior predictions use that group's coefficient distribution conditional on its observed outcomes. For an unobserved group, its coefficients must instead be generated from the inferred group-level distribution. The latter prediction integrates uncertainty in a new group deviation as well as uncertainty in the shared parameters and residual variation. Setting the new deviation to zero predicts at the shared mean relationship, but does not reproduce the full predictive distribution for a new group. A small number of groups can leave between-group scales and correlations weakly informed even if the total number of individuals is large, making prior assessment especially important (Gelman, 2006b).

Prior Specification and Posterior Computation

A Bayesian hierarchical regression requires priors for the shared coefficients, group-level scales and correlations, and likelihood parameters. The distribution of group deviations in Equation (12) is one layer of the hierarchy; its unknown covariance parameters require their own priors. Prior specification must therefore address the hierarchy as a whole. It should reflect the response scale, predictor coding, plausible conditional associations, and plausible between-group variation rather than copying numerical values from a different dataset (Gelman, 2006b).

Shared Coefficients and Scale Parameters

$$\mu_\alpha \sim N(m_\alpha, s_\alpha^2). \tag{16}$$

$$\mu_\beta \sim N\left(m_\beta, s_\beta^2\right). \tag{17}$$

Equations (16) and (17) illustrate proper normal priors for the shared intercept and slope. The quantities m_α and m_β are prior locations, and $s_\alpha > 0$ and $s_\beta > 0$ are prior standard deviations fixed before fitting under a stated prior specification. They are placeholders for methodological discussion, not chosen NHANES settings. Their units differ: an intercept prior is expressed on the outcome scale, while a slope prior is expressed in outcome units per predictor unit. Standardisation can aid calibration, but binary, ordinal, and continuous predictors require appropriate coding rather than a universal scaling rule.

Positive scale parameters such as σ, τα, and τβ can be assigned proper half-normal or half-Student-t priors with scientifically defensible scales. These are distributions restricted to positive values; they are not priors on unrestricted variances. Gelman's analysis of hierarchical variance priors demonstrates that apparently noninformative inverse-gamma choices can have important consequences, and discusses the half-t family when weak information is desired, including settings with few groups (Gelman, 2006b). The numerical scales used in the application are documented separately in Table 1.

Correlation Priors and Prior Predictive Checks

$$\sum = D\Omega D, \quad D = diag(\tau_\alpha, \tau_\beta). \tag{18}$$

$$\Omega \sim LKJ(\eta), \quad \eta > 0. \tag{19}$$

Equation (18) separates covariance into positive scales and a correlation matrix Ω, whose off-diagonal element is ρ in the two-coefficient case. Equation (19) gives an illustrative LKJ prior. Its shape parameter η is fixed by the prior specification. $\eta = 1$ gives a uniform density over correlation matrices of the specified dimension; $\eta > 1$ favours matrices closer to the identity. A uniform matrix prior should not be confused with independent uniform priors

for all pairwise correlations in higher dimensions. This separation permits scales and correlations to be assessed individually (Bürkner, 2017).

Prior predictive checking generates parameters from the complete prior and then generates replicated outcomes conditional on relevant predictor values and group structure. The resulting distributions should be examined for implausible outcome ranges, unreasonable slopes, excessive group differences, and unrealistic tail behaviour. A prior that looks broad on an isolated parameter can produce highly implausible predictions when combined with other parameters. Gabry et al. (2019) demonstrate how visualisation supports this assessment throughout the Bayesian workflow. Prior sensitivity should compare reasonable alternatives and examine whether substantive conclusions or predictions depend strongly on them, particularly for poorly informed group-level scales and correlations (Gabry et al., 2019).

Posterior Factorisation and Computation

$$p(\theta, \boldsymbol{u} \mid \boldsymbol{y}, X) \propto p(\boldsymbol{y} \mid X, \theta, \boldsymbol{u}) p(\boldsymbol{u} \mid \sum) p(\theta). \qquad (20)$$

In Equation (20), **y** denotes all observed responses, X the predictor and group-design information, **u** the collection of group deviations, and θ the remaining unknown parameters, including shared coefficients, covariance parameters, and residual scale. The likelihood factors over observations under conditional independence, and p(**u** | Σ) factors over groups under the specified hierarchy. The prior p(θ) need not factor into independent components unless that assumption is explicitly made. Posterior inference propagates uncertainty across these layers rather than substituting point estimates for the hyperparameters.

Hamiltonian Monte Carlo uses gradient information to explore continuous posterior distributions. The No-U-Turn Sampler adapts trajectory length, reducing a major tuning requirement of

conventional HMC (Hoffman & Gelman, 2014). Bayesian multilevel models can be fitted through software such as brms using Stan, but the implementation, software versions, chains, warm-up, retained draws, and tuning settings must be reported from the completed analysis rather than specified retrospectively in the conceptual discussion (Bürkner, 2017).

$$\boldsymbol{u}_j = L_\Sigma z_j, \quad z_j \sim N_2(0, I_2). \tag{21}$$

Equation (21) expresses a non-centred representation, where L_Σ is a Cholesky factor satisfying $L_\Sigma L_\Sigma^T = \Sigma$, and I_2 is the two-dimensional identity matrix. This is a reparameterisation of the same Gaussian group-deviation model when the priors and transformations are handled consistently. It can alter computational geometry without changing the intended statistical model. Whether a centred or non-centred form is more efficient depends on the information in the groups and the posterior geometry; diagnostic evidence must guide any change. Divergent trajectories should prompt investigation of parameterisation, scaling, priors, and model structure, rather than simply extending a problematic run (Gabry et al., 2019).

Predictive Evaluation and Model Diagnostics

Model assessment should distinguish three questions: whether posterior computation is reliable, whether predictive evaluation is reliable for the intended target, and whether the model reproduces scientifically relevant features of the data. A model can pass a computational diagnostic yet fit the data poorly, and an apparently favourable cross-validation estimate can be unreliable if importance weights are unstable. These assessments provide complementary evidence rather than substitutes for one another (Gabry et al., 2019).

Computational Diagnostics

Rank-normalised split $\hat{R}$, its folded counterpart, bulk effective sample size, tail effective sample size, and Monte Carlo standard errors should be assessed for the parameters and derived quantities relevant to the conclusions. Rank plots and other chain displays can reveal unequal exploration. Vehtari et al. (2021) recommend the stricter $\hat{R} < 1.01$ threshold and discuss effective sample size requirements, including a rank-normalised ESS exceeding 400 as a practical diagnostic recommendation. Such thresholds are screening criteria, not proofs of convergence; precision requirements depend on the summaries being reported (Vehtari et al., 2021).

For HMC-based sampling, divergent transitions and trajectory-depth warnings also require examination. Increasing retained draws reduces Monte Carlo error only when the sampler explores the target adequately. It cannot repair a systematically missed region of the posterior. Diagnostics should include group-level standard deviations and correlations, whose behaviour may differ from that of well-informed shared coefficients. Any adjustments to tuning, priors, or parameterisation should be documented together with the final diagnostics (Gabry et al., 2019).

Leave-One-Out Predictive Evaluation

$$elpd_{LOO} = \sum_{i=1}^{N} log\, p(y_i \mid x_i, D_{-i}). \quad (22)$$

In Equation (22), the sum runs over all N observations, D_{-i} is the analysis dataset excluding observation i's outcome, x_i includes its predictor and group information, and $elpd_{LOO}$ denotes the sum of held-out log predictive densities. A single observation index replaces the earlier pair of indices for compactness. The conditioning retains predictors and group membership while excluding the held-out outcome. Larger $elpd_{LOO}$ indicates better predictive performance on this log-score scale; LOOIC = −2 $elpd_{LOO}$ reverses the direction.

Model differences must be accompanied by uncertainty rather than interpreted through a ranking alone (Vehtari et al., 2017).

PSIS-LOO uses posterior draws from the full-data fit and smoothed importance weights to approximate the leave-one-out predictive quantities. This avoids fitting the model separately for every observation when the approximation is stable. The pointwise likelihood must match the intended predictive unit and the likelihood actually fitted. Estimated Pareto-k values assess importance-weight tail behaviour and therefore the reliability of the approximation; they do not measure substantive adequacy or certify that predictions are calibrated (Vehtari et al., 2024).

$$k_{crit} = min\left(1 - \frac{1}{log_{10}(S)}, 0.7\right). \quad (23)$$

Equation (23) gives the sample-size-dependent warning threshold discussed by Vehtari et al. (2024), with S denoting the number of posterior draws used in the importance-sampling calculation. Diagnostics for MCMC efficiency remain relevant; simply increasing nominal draws is not a substitute for effective sampling. Problematic cases should be investigated, with exact leave-one-out refitting or an appropriate K-fold design considered when the approximation is unreliable. The final report should state the diagnostic rule and how any problematic cases were handled (Vehtari et al., 2017).

Represented Groups and Unobserved Groups

Observation-level LOO usually evaluates a new individual in a represented group when other individuals from that group remain in training. The group coefficients can then be informed by those retained outcomes. If a group contains only one individual, removing that outcome leaves no within-group information and changes the nature of that prediction. Thus, a mixed collection of group sizes can produce different information conditions across held-out

observations. The target and group-size distribution should be examined before interpreting an aggregate predictive score.

Predicting an entirely unobserved group requires withholding all outcomes from that group and integrating over its group deviations. Retaining coefficient estimates fitted with the held-out group's outcomes would leak information into this evaluation. Leave-one-group-out or grouped K-fold validation can address the new-group target, but it involves different predictive units and potentially different computational requirements from ordinary observation-level PSIS-LOO. The distinction follows the principle that the dependence structure and intended prediction task should determine the validation design (Roberts et al., 2017).

An independent no-pooling model does not define a predictive distribution for an unseen group without an additional assumption. Consequently, all three pooling strategies can be compared directly for represented-group prediction, but a new-group comparison requires explicitly defined predictive rules for every model. Scores for individuals and scores for jointly predicted groups are not interchangeable. Uncertainty calculations must also respect the validation unit, rather than automatically treating observations within a held-out group as independent replications.

Posterior Predictive Checks and Reporting

Posterior predictive checking generates replicated outcomes from posterior draws while preserving the covariate and group structure appropriate to the check. Useful comparisons include overall and group-specific location and dispersion, outcome tails, residual patterns across predictors, and the range of fitted group relationships. Checks within represented groups condition on their inferred deviations; checks for new groups additionally generate fresh deviations. Aggregated plots alone can conceal subgroup

misfit, while very small groups require care because their empirical distributions are themselves uncertain (Gabry et al., 2019).

Predictive intervals and intervals for expected responses answer different questions: the former incorporate residual outcome variation, whereas the latter quantify uncertainty in the conditional mean. Coverage or calibration assessments intended to reflect held-out performance should use the chosen cross-validation design or external validation data. In-sample posterior predictive checks are valuable for model criticism but should not be labelled external validation. Bayesian parameter intervals should be reported as credible intervals (CrI), with their probability level specified.

The application below documents its data version, variables, grouping, complete-case restriction, and prediction target. Its prior sensitivity analysis and PSIS-LOO comparison provide complementary information, but neither replaces prior or posterior predictive checking. No prior predictive checks or graphical posterior predictive checks were supplied with the completed analysis. Accordingly, model adequacy beyond the reported predictive comparisons remains an open diagnostic question (Gabry et al., 2019).

Relationship to Joint and Individual Component Methods

Shared and group-specific structure can be represented at different levels. In hierarchical regression, the primary objects are regression coefficients linked by a probability distribution. In component methods, the objects include latent scores or low-rank structures extracted from predictor matrices. The terminology overlaps, but the statistical constructions and inferential targets differ.

JIVE was developed for multiple data types measured on a common set of objects. It decomposes each data block into joint low-rank structure, individual low-rank structure, and residual noise,

with rank and orthogonality constraints supporting the separation. Here, individual refers to structure specific to a data type, rather than automatically to a demographic group or a varying regression coefficient. The original method is an exploratory data-integration approach and does not itself define the hierarchical response model considered in this chapter (Lock et al., 2013).

JICO addresses multigroup regression in which the same predictors are measured in different sets of individuals. Shared and group-specific contributions to the response are driven by low-rank joint and individual structures in the predictors. Its latent-component representation, rank choices, and continuum-regression construction distinguish it from a conventional varying-coefficient hierarchy. Its common motivation of balancing global and group-specific information does not imply an identical mechanism of partial pooling (Wang et al., 2024).

Kara et al. (2026) develop a Bayesian formulation of joint and individual component regression for multigroup physiological data, incorporating posterior uncertainty into the component framework. This is relevant background for the broad problem of separating common and group-specific information. However, introducing Bayesian estimation does not remove the distinction between a component decomposition and a hierarchical distribution for coefficients. A formula such as y ~ x + (1 + x | group) specifies varying intercepts and slopes; it does not, by itself, estimate joint and individual latent predictor subspaces (Kara et al., 2026).

The methodological choice should follow the data structure and objective. A moderate set of interpretable predictors with interest in group-specific conditional associations is naturally addressed by hierarchical regression. High-dimensional predictor blocks with a substantive need to identify shared and individual latent structure may motivate component methods. This distinction concerns suitability, not a claim that one family dominates the other. The

application is Bayesian hierarchical regression, with complete-pooling and no-pooling benchmarks. It is not designated Bayesian-JICO, and no performance result from a component-method study is presented as evidence for the empirical superiority of the hierarchy used here.

Further Perspectives on Shrinkage and Heterogeneity

Bayesian subgroup analysis offers a closely related setting in which exchangeable subgroup effects are estimated jointly rather than subjected to a sequence of independent tests. Jones et al. (2011) discuss shrinkage approaches in medical subgroup analysis. This connection concerns the mechanism of information sharing; it does not convert the observational associations considered here into treatment effects. Random-effects meta-analysis provides another useful analogy: when only a few studies or groups are available, uncertainty about heterogeneity can remain substantial and inference can depend materially on its prior (Röver et al., 2021). Neither analogy determines the grouping or prior scale for the present application.

Application Model and Estimation

Let y_{ij} be centred systolic blood pressure for individual i in group j, and let x_{ij} contain seven predictors. Models M1, M2, and M3 use Gaussian errors and respectively share all coefficients, estimate independent group coefficients, or link coefficients through a hierarchy. Unlike the fully unpooled illustration in Equation (2), M2 shares one residual scale σ across groups; no pooling here refers specifically to its regression coefficients. M3 uses a shared vector μ and independent deviations u_j:

$$b_j = \mu + u_j,\ u_j \sim N_8(0, diag(\tau_0^2, \ldots, \tau_7^2))) \qquad (24)$$

Here $b_j = (\alpha_j, \beta_{j1}, \ldots, \beta_{j7})^T$ comprises the intercept and slopes, μ contains their shared means, and τ_k is the between-group standard

deviation for coefficient k. Independence is a parsimonious specification with five groups, rather than evidence that intercept and slope deviations are uncorrelated. The correlated hierarchy in Equations (10)–(15) remains a general model option; it was not fitted in this application. M4 retains this coefficient hierarchy and uses a Student-t likelihood with location $\alpha_j + x_{ij}^T\beta_j$, scale $\sigma > 0$, and degrees of freedom $\nu > 1$. All four models use a residual scale common to the groups.

Priors and computation

The response was centred at its sample mean, so that all effects are in mmHg, and the six continuous predictors were standardized. Table 1 lists the priors, which are intended to be weakly informative on this scale. A $N(0, 10^2)$ prior places 95% of its mass within about ±20 mmHg per SD, and the half-Student-t prior with scale 5 for the group-level SDs favours between-group differences of a few mmHg while keeping a heavy tail. Models were fitted with brms 2.23.0 (Bürkner, 2017) in R 4.5.1 using the No-U-Turn sampler (Hoffman & Gelman, 2014), with four chains of 2,000 iterations (1,000 warm-up), adapt_delta = 0.95, and seed 2026.

Table 1. Priors (shared by all models unless noted).

Parameter	Prior	Remark
Intercept (M1, M3, M4)	Normal(0, 10^2)	mmHg, relative to the sample mean
Coefficients (b; M2: group means and slopes)	Normal(0, 10^2)	per SD for continuous predictors; sex: male vs female
Residual scale σ	Student-t(3, 0, 15), $\sigma > 0$	half-Student-t; common to all groups
Group-level SDs τ_k (M3, M4)	Student-t(3, 0, 5), $\tau > 0$	half-Student-t; scales 2.5 and 10 in sensitivity runs

Parameter	Prior	Remark
Degrees of freedom ν (M4)	Gamma(2, 0.1), ν > 1	shape–rate; truncated to ν > 1

The fitting workflow screened for R-hat < 1.01, bulk and tail ESS above 400, and no more than 4 divergent transitions among 4,000 post-warm-up draws. Fits exceeding this workflow threshold were refitted once with a different seed and adapt_delta = 0.965. These are reported screening rules, not sufficient conditions for reliable posterior exploration. In particular, even a small number of unresolved divergences requires investigation; satisfactory R-hat and ESS do not eliminate this concern (Vehtari et al., 2021).

Model assessment

Models were compared with PSIS-LOO, reporting ELPD differences and their standard errors (Vehtari et al., 2017). Pareto-k values above 0.7 were used as a warning threshold (Vehtari et al., 2024). Uncertainty estimates for LOO comparisons can themselves be unreliable, especially for similar predictions or misspecified models; small differences should therefore be interpreted cautiously (Sivula et al., 2025). Observation-level LOO targets a new individual from a group that remains represented in the training data. Coefficient shrinkage was summarized using posterior means:

$$r_{jk} = \frac{\bar{b}_{jk}^{PP} - \bar{\mu}_k}{\bar{b}_{jk}^{NP} - \bar{\mu}_k} \tag{25}$$

In Equation (25), $\bar{b}^{PP}_{jk}$ and $\bar{b}^{NP}_{jk}$ are coefficient posterior means under partial and no pooling, and $\bar{\mu}_k$ is the shared-effect posterior mean under partial pooling. A ratio of 1 denotes unchanged distance from that shared estimate and 0 denotes equality with it. Ratios outside [0, 1] can occur because these are estimates from different fitted models; this descriptive statistic is not the exact conjugate weight in Equation (8). Median ratios were calculated

within each group over coefficients with $|\bar{b}^{NP}_{jk} - \bar{\mu}_k| > 0.1$ mmHg to avoid unstable denominators. Group-level SD prior scales of 2.5, 5, and 10 were compared.

Application to NHANES Blood Pressure Data

Data

We used the NHANES R package, a resampled version of the 2009–2012 survey cycles in which a participant can appear in more than one row. For adults aged 20 years or older, the data contain 7,235 rows from 4,654 distinct participants. We kept one row per participant and excluded 769 participants (16.5%) with missing values on any analysis variable, leaving n = 3,885. The outcome was average systolic blood pressure (mean about 121 mmHg, SD 17.2). Survey weights were not used, although the National Center for Health Statistics recommends weights and design variables for NHANES analyses (Johnson et al., 2014), so the results describe conditional associations in this sample and not population quantities (Gelman, 2007). The package documentation describes NHANES as an educational resampling of the original survey, rather than a research database (Project MOSAIC, n.d.). Removing repeated identifiers and restricting to complete cases does not restore the original survey design or establish a simple random national sample.

The groups were the survey-defined race/ethnicity categories White (n = 2,393; mean SBP 121.4), Black (531; 123.7), Mexican (390; 118.9), Other (313; 117.1), and Hispanic (258; 120.0). These categories are a grouping device, they do not identify biological mechanisms, and the Mexican and Hispanic categories both refer to Hispanic origin. The predictors were age (SD 17.3 years), body mass index (6.8 kg/m^2), pulse (11.9 beats per minute), total cholesterol (1.06 mmol/L), HDL cholesterol (0.41 mmol/L), poverty income ratio (1.67), and male sex. Blood pressure medication information was not included in the supplied educational extract used for this

analysis; this does not imply that it is absent from the original NHANES survey.

Computation and model comparison

All fits satisfied R-hat < 1.01 (Table 2). The partial-pooling model had the lowest ESS (888 bulk, 676 tail), which concerned the group-level SDs, and 4 divergent transitions (0.1%). Its Student-t version needed a second attempt and then showed 2 divergent transitions. For the SD-prior sensitivity fits, the scale-2.5 model had 2 divergent transitions and the scale-10 model needed a second attempt and then showed 3 divergent transitions. No Pareto-k value exceeded 0.7; one observation in the no-pooling model exceeded 0.5. These remaining divergences limit confidence in the reported posterior summaries and predictive comparisons until their locations are examined and satisfactory refits are obtained.

Table 2. Computation and predictive comparison. ΔELPD is relative to the best model (M4).

Model	Max R-hat	Min ESS bulk / tail	Div.	$ELPD_{LOO}$ (SE)	ΔELPD (SE)	p_{loo}
M1 Complete pooling	1.002	1,844 / 2,696	0	−16,067.9 (68.5)	−171.4 (29.4)	11.1
M2 No pooling	1.003	1,696 / 2,491	0	−16,062.1 (67.4)	−165.6 (28.4)	43.9
M3 Partial pooling	1.004	888 / 676	4	−16,055.0 (68.0)	−158.5 (28.6)	30.7
M4 Partial pooling, Student-t	1.007	1,058 / 1,315	2	−15,896.6 (58.2)	0	28.1

With Gaussian errors, partial pooling had a higher ELPD than no pooling (by 7.1, SE 3.8) and complete pooling (by 12.9, SE 6.9), about 1.9 standard errors in both cases, which is modest evidence. Its effective number of parameters (ploo = 30.7) was lower than that

of the no-pooling model (43.9), which indicates regularization, and higher than that of complete pooling (11.1), which indicates that some group-level structure is used. However, these differences are small next to the gain from Student-t errors, which improved the ELPD of the partial-pooling model by 158.5 (SE 28.6), about 5.5 standard errors. The distribution of the residuals mattered far more here than the pooling structure.

Shared effects

Table 3. Shared effects under partial pooling (M3) and the pooled effects under complete pooling (M1); posterior means (95% CrI), mmHg per SD (sex: male vs female).

Predictor	M3 shared effect	M1 pooled effect	CrI width M3 / M1
Intercept	−2.19 (−5.04, 0.65)	−2.73 (−3.40, −2.07)	4.3
Age	8.41 (6.46, 10.55)	7.54 (7.04, 8.02)	4.2
BMI	1.37 (0.41, 2.58)	1.53 (1.01, 2.03)	2.1
Pulse	1.27 (−0.36, 3.15)	0.73 (0.25, 1.23)	3.6
Total cholesterol	1.12 (0.19, 1.96)	1.13 (0.61, 1.62)	1.8
HDL cholesterol	0.37 (−0.87, 1.50)	0.55 (0.01, 1.08)	2.2
Poverty ratio	−1.22 (−2.23, 0.01)	−1.61 (−2.09, −1.13)	2.3
Male sex	5.71 (4.20, 7.53)	5.55 (4.57, 6.53)	1.7
Residual scale σ	15.02	15.11	–

Age, BMI, total cholesterol, and male sex had 95% credible intervals excluding zero under both models, with broadly similar posterior means. M3 shared-effect intervals were 1.7 to 4.3 times wider than M1 pooled-effect intervals. Pulse changed from 0.73 (0.25, 1.23) under M1 to 1.27 (−0.36, 3.15) under M3; the HDL and poverty-ratio intervals also included zero under M3. These

differences show sensitivity to the coefficient structure and estimand, rather than establishing which intervals are better calibrated. The shared age mean of 8.41 differs from the pooled mean of 7.54. The hierarchical shared mean is a parameter of the group-coefficient distribution inferred jointly with the deviations and their scales; it is not an arithmetic average of observed group slopes with equal weights.

Group-level heterogeneity and shrinkage

Table 4. Shared-effect means and between-group standard deviations (M3), with 95% CrI for the standard deviations, mmHg.

Coefficient	Shared effect	Group-level SD	95% CrI of SD
Intercept	−2.19	2.59	0.99 – 5.97
Age	8.41	1.86	0.51 – 4.73
BMI	1.37	0.66	0.02 – 2.40
Pulse	1.27	1.40	0.14 – 3.93
Total cholesterol	1.12	0.55	0.02 – 2.03
HDL cholesterol	0.37	0.77	0.03 – 2.75
Poverty ratio	−1.22	0.81	0.03 – 2.68
Male sex	5.71	1.00	0.03 – 3.48

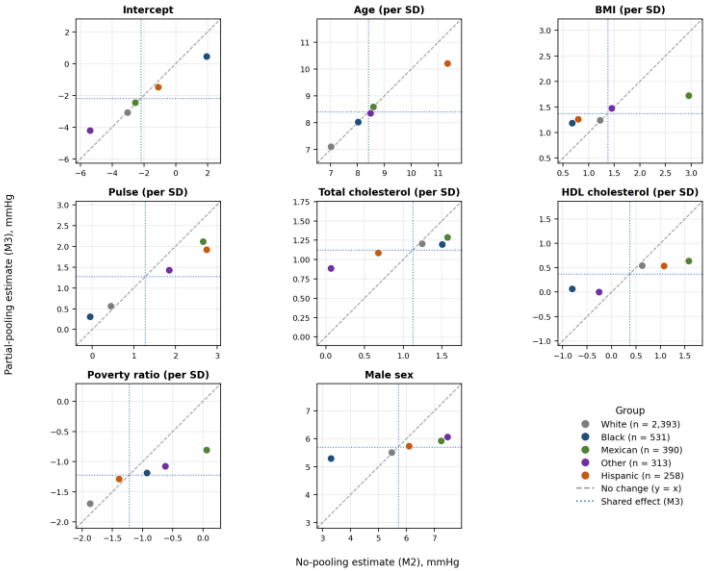
Group-level variation was modest and imprecisely estimated (Table 4). The intercept SD was 2.6 mmHg (0.99–5.97), and the share of residual-plus-group variance attributable to group intercepts was small (0.022; 95% CrI 0.004–0.138). Age and pulse showed the largest slope SDs, while the intervals of the other slope SDs reached close to zero. None of the individual deviation intervals excluded zero: all 40 group deviations (five groups, eight coefficients) had 95% credible intervals that included zero. A notable unpooled age slope was that of the Hispanic group (11.4 per SD, against 7.0 for the

White group), which partial pooling reduced to 10.2. The reported variance fraction of 0.022 pertains to the predictor reference x = 0, where the random-slope terms vanish, and is based on $\tau_0^2/(\tau_0^2 + \sigma^2)$. With varying slopes, between-group variance generally depends on x, so this fraction is not a universal intraclass correlation across covariate values. Interval inclusion of zero does not establish identical group relationships.

Shrinkage was strong for the four smaller groups (Figure 1). The median shrinkage ratio was 0.31 for Hispanic (n = 258), 0.25 for Other (n = 313), 0.35 for Mexican (n = 390), and 0.26 for Black (n = 531), so partial pooling moved about two thirds to three quarters of the distance from the unpooled estimate to the shared effect. The ratio was 0.84 for the White group (n = 2,393). The ratio is not strictly ordered by group size because it also depends on predictor variation and on the coefficient.

Figure 1. Group-specific coefficients under no pooling (x-axis) and partial pooling (y-axis). Points on the dashed line are unchanged;

dotted lines mark the shared effect. Points for smaller groups lie closer to the shared effect than to the diagonal.



Prior sensitivity

Table 5. Posterior means of the group-level SDs (mmHg) under three prior scales.

Coefficient	Scale 2.5	Scale 5 (primary)	Scale 10
Intercept	2.19	2.59	2.82
Age	1.60	1.86	1.95
BMI	0.59	0.66	0.68
Pulse	1.25	1.40	1.50
Total cholesterol	0.51	0.55	0.59
HDL cholesterol	0.69	0.77	0.88
Poverty ratio	0.73	0.81	0.87

Coefficient	Scale 2.5	Scale 5 (primary)	Scale 10
Male sex	0.82	1.00	1.14

The group-level SDs increased monotonically with the prior scale (for example the intercept SD from 2.19 to 2.82, the age SD from 1.60 to 1.95), but predictive performance was practically unchanged: the ELPD differences relative to the scale-2.5 model were −0.3 (SE 0.3) for the primary prior and −0.4 (SE 0.4) for scale 10. The monotone pattern shows that the data inform the SDs only weakly, as expected with five groups, and the SD estimates should be read as prior-dependent.

Student-t errors

The Student-t model had substantially higher reported ELPD than the Gaussian models, motivating examination of its estimates (Table 6). Its degrees-of-freedom posterior mean was 4.85 (95% CrI 4.17–5.68), supporting heavier tails under this likelihood. Its scale mean of 11.61 is not a residual standard deviation. For $\nu > 2$, the conditional standard deviation is given by Equation (26); inserting the posterior means yields approximately 15 mmHg, rather than the posterior mean of that transformed quantity. For $1 < \nu \leq 2$, allowed by the prior, the variance is infinite.

$$SD(y_{ij} \mid b_j, \sigma, \nu) = \sigma\sqrt{\frac{\nu}{\nu - 2}}, \quad \nu > 2 \tag{26}$$

Table 6. Partial pooling with Student-t errors (M4): shared effects and group-level SDs (posterior means, 95% CrI), mmHg.

Parameter	Shared effect	Group-level SD
Intercept	−3.58 (−6.03, −1.21)	2.18 (0.75, 5.32)
Age	7.63 (6.07, 9.29)	1.47 (0.38, 3.69)
BMI	1.59 (0.53, 2.63)	0.77 (0.04, 2.52)

Parameter	Shared effect	Group-level SD
Pulse	0.95 (−0.20, 2.29)	0.90 (0.05, 2.90)
Total cholesterol	1.02 (0.24, 1.80)	0.47 (0.02, 1.75)
HDL cholesterol	0.34 (−0.56, 1.17)	0.49 (0.02, 1.93)
Poverty ratio	−1.06 (−1.85, −0.07)	0.62 (0.03, 1.94)
Male sex	6.23 (4.74, 7.86)	0.90 (0.03, 3.20)
Residual scale σ	11.61 (11.15, 12.07)	–
Degrees of freedom ν	4.85 (4.17, 5.68)	–

The shared effects were close to those of the Gaussian model (for example age 7.63 versus 8.41, male sex 6.23 versus 5.71), and their intervals were narrower for every coefficient (by 3% to 29%). Pulse and HDL cholesterol still included zero, while the poverty effect (−1.06; −1.85, −0.07) now narrowly excluded it, so its status is borderline. The group-level SDs were smaller than under Gaussian errors (for example age 1.47 versus 1.86, pulse 0.90 versus 1.40) and remained imprecise, with lower limits near zero for the slope SDs. Part of the between-group variation estimated under Gaussian errors may therefore reflect heavy-tailed noise. As under Gaussian errors, none of the 40 group deviations had a 95% credible interval excluding zero; the closest were the intercept of the Black group (2.19; −0.25, 5.11) and the age slope of the Hispanic group (1.30; −0.41, 3.65).

Discussion

With Gaussian errors, partial pooling predicted slightly better than both extremes: it regularized the unstable no-pooling estimates (lower effective number of parameters) and used more group-level structure than complete pooling. Its main practical effect was on inference rather than prediction. The shared effects had much wider intervals than the pooled effects, and the conclusions about pulse,

HDL cholesterol, and the poverty ratio were no longer supported under the hierarchical model. This illustrates why the shared effect of a hierarchical model should not be read as the coefficient of a single pooled regression.

Estimated heterogeneity was modest and uncertain. Under both likelihoods, all 40 individual deviation intervals included zero, and group-level SDs were imprecise and sensitive to their prior. This does not establish identical groups or absence of heterogeneity. Shrinkage ratios show that the four smaller groups moved considerably towards the shared regression. The Black group had the highest observed mean SBP (123.7 mmHg) and an M3 intercept deviation of 2.65 mmHg, but the deviation interval included zero. National hypertension prevalence estimates concern a different outcome, population definition, and survey period, and do not validate this conditional coefficient (Ostchega et al., 2020).

The Student-t likelihood improved the ELPD by 158.5, far more than any pooling structure, and ν was well identified (posterior mean 4.85, 95% CrI 4.17–5.68), so the data clearly favour heavy tails. A symmetric heavy-tailed likelihood can improve predictive scores without reproducing every feature of the response distribution; this comparison alone does not establish overall model adequacy. Alternatives such as a log transformation or skewed distributions were not compared. The comparison does show that the likelihood and the coefficient structure are separate decisions, which should each be examined.

Several limitations apply. Observation-level LOO assesses represented groups; prediction in an unobserved group was not evaluated. Only five groups inform the SD parameters, deviation correlations were excluded, and unresolved divergent transitions remain. The educational data were analysed without survey weights or design variables and after identifier deduplication. Complete-case exclusion removed 16.5% of distinct adult participants; the

missingness mechanism was not assessed, so ignorability or missing at random cannot be asserted, and selection bias remains possible. Medication information was unavailable in the analysed extract. Race/ethnicity categories are social and survey classifications, not biological explanations. A common residual scale was assumed across groups. Neither prior predictive checks nor graphical posterior predictive checks were reported. Finally, one outcome and grouping cannot establish how pooling strategies perform in other settings.

Conclusion

In 3,885 NHANES adults grouped into five race/ethnicity categories, Bayesian partial pooling gave a modest predictive gain over complete and no pooling with Gaussian errors, strong shrinkage of the four smaller groups, and wider intervals for the shared effects. None of the group-deviation intervals excluded zero, and the group-level SDs depended on the prior. A Student-t likelihood improved prediction far more than the pooling structure. We recommend reporting pooled and hierarchical effects side by side, checking the residual distribution as a separate step, and describing group-level SDs together with their prior sensitivity. These conclusions remain conditional on the analysed extract, model specification, and the unresolved computational diagnostics.

References

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 182*(2), 389–402. https://doi.org/10.1111/rssa.12378

Gelman, A. (2006a). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics, 48*(3), 432–435. https://doi.org/10.1198/004017005000000661

Gelman, A. (2006b). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1*(3), 515–533. https://doi.org/10.1214/06-BA117A

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science, 22*(2), 153–164. https://doi.org/10.1214/088342306000000691

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*(2), 189–211. https://doi.org/10.1080/19345747.2011.618213

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15*(47), 1593–1623. https://jmlr.org/papers/v15/hoffman14a.html

Johnson, C. L., Dohrmann, S. M., Burt, V. L., & Mohadjer, L. K. (2014). *National Health and Nutrition Examination Survey: Sample design, 2011–2014*. Vital and Health Statistics, Series 2(162). National Center for Health Statistics. https://stacks.cdc.gov/view/cdc/22155

Jones, H. E., Ohlssen, D. I., Neuenschwander, B., Racine, A., & Branson, M. (2011). Bayesian models for subgroup analysis in clinical trials. *Clinical Trials, 8*(2), 129–143. https://doi.org/10.1177/1740774510396933

Kara, M., Cengiz, M. A., Dünder, E., & Şenel, T. (2026). Bayesian joint and individual component regression for multigroup physiological data. *Scientific Reports, 16*, Article 16274. https://doi.org/10.1038/s41598-026-52063-z

Lock, E. F., Hoadley, K. A., Marron, J. S., & Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics, 7*(1), 523–542. https://doi.org/10.1214/12-AOAS597

Ostchega, Y., Fryar, C. D., Nwankwo, T., & Nguyen, D. T. (2020). *Hypertension prevalence among adults aged 18 and over: United States, 2017–2018* (NCHS Data Brief No. 364). National Center for Health Statistics. https://www.cdc.gov/nchs/products/databriefs/db364.htm

Project MOSAIC. (n.d.). *NHANES: NHANES data*. In NHANES R package documentation. https://search.r-project.org/CRAN/refmans/NHANES/html/NHANES.html

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography, 40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Röver, C., Bender, R., Dias, S., Schmid, C. H., Schmidli, H., Sturtz, S., Weber, S., & Friede, T. (2021). On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. *Research Synthesis Methods, 12*(4), 448–474. https://doi.org/10.1002/jrsm.1475

Sivula, T., Magnusson, M., Alonzo Matamoros, A., & Vehtari, A. (2025). Uncertainty in Bayesian leave-one-out cross-validation based model comparison. *Bayesian Analysis*. Advance online publication. https://doi.org/10.1214/25-BA1569

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis, 16*(2), 667–718. https://doi.org/10.1214/20-BA1221

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., & Gabry, J. (2024). Pareto smoothed importance sampling. *Journal of Machine Learning Research, 25*(72), 1–58. https://jmlr.org/papers/v25/19-556.html

Wang, P., Wang, H., Li, Q., Shen, D., & Liu, Y. (2024). Joint and individual component regression. *Journal of Computational and Graphical Statistics*, *33*(3), 763–773. https://doi.org/10.1080/10618600.2023.2284227

CHAPTER 3

BAYESIAN ROBUST REGRESSION FOR CENSORED SURVIVAL DATA: A COMPARISON OF NORMAL, STUDENT-T, AND CAUCHY ACCELERATED FAILURE TIME MODELS USING PBC DATA

BURÇİN ŞEYDA ÇORBA[1] FATİH SAĞLAM[2]

Introduction

Clinical time-to-event data are routinely right-censored: for a censored patient the event time is known only to exceed the follow-up time. The likelihood must therefore use the survival function for these patients instead of treating follow-up as an observed event time (Rubio & Steel, 2018; Hamura et al., 2025).[1]

The accelerated failure time (AFT) model regresses the logarithm of survival time on covariates, so that effects are interpreted as multiplicative changes in survival time. The error distribution on the log scale determines the shape of the lifetime

[1] Dr., Ondokuz Mayıs University, Department of Statistics, ORCID: 0000-0002-0085-4351
[2] Dr., Ondokuz Mayıs University, Department of Statistics, ORCID: 0000-0002-2084-2008

distribution. Normal errors imply a lognormal lifetime, whereas Student-t errors allow heavier tails and reduce the influence of extreme residuals (Lange et al., 1989; Rubio & Hong, 2016). A heavier-tailed model is not automatically better: right-skewness of raw survival times does not imply heavy-tailed errors on the log scale, and the need for robustness should be judged after accounting for covariates and censoring.

Several studies combine censoring with flexible or heavy-tailed errors. Rubio and Steel (2018) studied Bayesian scale-mixture models and posterior propriety under censoring; Cengiz et al. (2025) combined Student-t and Cauchy errors with censored regression for medical data; and the preprint by Hamura et al. (2025) proposed scale mixtures of generalized gamma distributions. The prior for the degrees of freedom ν deserves particular care, because posterior propriety cannot be taken for granted under improper priors. Fonseca et al. (2008) studied objective priors for Student-t regression; their subsequent correction clarified that the independence Jeffreys prior yields a proper posterior under the stated conditions, whereas the Jeffreys-rule prior does not (Fonseca et al., 2014). These results concern the prior structures studied in those papers and should not be interpreted as a direct propriety theorem for the present censored model.

This chapter examines Bayesian robust regression for right-censored survival data by comparing Normal, Student-t, and Cauchy error distributions within an accelerated failure time (AFT) framework. Building on the Bayesian censored regression approach discussed by Cengiz et al. (2025), the models are applied to the primary biliary cirrhosis (PBC) trial data (Dickson et al., 1989). The analysis evaluates predictive performance, sampling diagnostics, and sensitivity to censoring levels, prior specifications, and event definitions. All models are fitted in R using the brms package (Bürkner, 2017).

Bayesian AFT models for right-censored data

Model and likelihood

Let $T_i > 0$ be the event time, $C_i > 0$ the censoring time, $U_i = \min(T_i, C_i)$ the observed time, and $\delta_i = I(T_i \leq C_i)$ the event indicator for patient i = 1, …, n, with covariate vector x_i (including an intercept). The AFT model is

$$\log T_i = \mu_i + \sigma Z_i. \quad (1)$$

Here $\mu_i = x_i^T\beta$, $\sigma > 0$. The errors Z_i are independent and follow a distribution with location zero and unit scale, cumulative distribution function G, and density g. Unit scale does not imply unit variance for Student-t errors. For a symmetric G, the conditional median survival time is $\exp(\mu_i)$, and increasing covariate x_j by one unit multiplies every conditional quantile by $\exp(\beta_j)$. This quantity is a time ratio: values above one indicate longer survival.

Assume that censoring is conditionally independent of the event time given the covariates and that the censoring mechanism has parameters distinct from θ (Rubio & Steel, 2018). A patient with an observed event then contributes a density, whereas a censored patient contributes a survival probability. With $z_i = (\log U_i - \mu_i)/\sigma$, the likelihood for log-follow-up time is

$$L_{log}(\theta) = \prod_{i=1}^{n} [g(z_i)/\sigma]^{\delta_i} [1 - G(z_i)]^{1-\delta_i}. \quad (2)$$

Here $\theta = (\beta, \sigma)$ for the Normal and Cauchy models and $\theta = (\beta, \sigma, \nu)$ for the Student-t model. The likelihood on the original time scale is $L_{time}(\theta) = L_{log}(\theta) \prod_{i=1}^{n} U_i^{-\delta_i}$. The Jacobian factor is independent of θ, so the two likelihoods give the same posterior under identical priors. For LOO and WAIC differences to be valid, all candidate models must use the same response transformation, observations, and censoring coding.

Error distributions

The three models share the covariates and the location-scale structure and differ only in G. For Normal errors, $Z_i \sim N(0, 1)$ and the lifetime is lognormal; this is the benchmark. For Student-t errors, Z_i follows a Student-t distribution with ν degrees of freedom, with density

$$g_\nu(z) = \{\Gamma[(\nu + 1)/2] / [\sqrt{(\nu\pi)}\,\Gamma(\nu/2)]\}\,(1 + z^2/\nu)^{-(\nu+1)/2}. \quad (3)$$

Smaller ν gives heavier tails, and the model approaches the Normal model as $\nu \to \infty$. The scale-mixture representation

$$Z_i \mid \lambda_i \sim N(0, 1/\lambda_i),\ \lambda_i \sim \text{Gamma}(\nu/2, \nu/2). \quad (4)$$

In Equation (4), the second parameter of the Normal distribution is the variance, and the Gamma distribution uses the shape–rate parameterization. This representation explains the robustness: large residuals can be accommodated through smaller latent precisions, reducing their influence on the fitted location relative to a Normal model. The representation is used here for interpretation; brms evaluates the Student-t density directly and does not separately sample λ_i. The parameter σ is a scale, not a standard deviation ($\text{Var}(\log T_i \mid x_i, \beta, \sigma, \nu) = \sigma^2\nu/(\nu - 2)$ for $\nu > 2$). The Cauchy model fixes $\nu = 1$; μ_i is then a median, and neither the mean nor the variance of $\log T_i$ exists. For any finite ν, the lifetime T_i has no finite positive-order moments, so mean survival time is not a suitable summary for the Student-t and Cauchy models. We report medians, time ratios, and survival probabilities instead.

Estimation, model assessment, and sensitivity design

Priors and computation

Survival time was measured in days and modelled on the log scale, and continuous covariates were standardized. Table 1 lists the priors, which are weakly informative on the log-time scale: a N(0, 1) prior for a slope per standard deviation places roughly 95% of prior

probability for the time ratio $\exp(\beta_j)$ between 0.14 and 7.1. The same priors for the intercept, slopes, and σ were used in all three models. Posterior sampling used the No-U-Turn sampler in Stan through brms 2.23.0 (R version 4.5.1), with four chains of 2,000 iterations (1,000 warm-up), adapt_delta = 0.95, and seed 2026. Convergence was judged by $\widehat{R} < 1.01$, bulk and tail effective sample sizes, and divergent transitions. An $\widehat{R} \geq 1.01$ would require investigation of chain mixing and sampler diagnostics before interpretation; changing the random seed alone is not a remedy for poor convergence.

Table 1. Prior distributions (identical in all models unless noted).

Parameter	**Prior**	**Remark**
Intercept	Normal(7.5, 2^2)	Log-days scale; intercept for internally centred predictors.
Slopes βj	Normal(0, 1^2)	per SD for standardized covariates; per unit for sex and edema
Scale σ	Student-t(3, 0, 2.5), σ > 0	half-Student-t: 3 df, location 0, scale 2.5
Degrees of freedom ν	Gamma(2, 0.1), ν > 1	Student-t model only (shape-rate); fixed at ν = 1 for Cauchy

Model comparison

The primary comparison used PSIS-LOO based on the censoring-aware log-likelihood, with WAIC as a complement (Vehtari et al., 2017). We report ELPD differences with their standard errors rather than rankings alone, because a small difference relative to its uncertainty provides limited evidence of a predictive advantage. Pareto-k values were assessed against the sample-size-dependent threshold $\min(1 - \frac{1}{log_{10}S}, 0.7)$, where S is the number of posterior draws; for the reported 4,000 post-warm-up draws, this threshold is 0.7 (Vehtari et al., 2024). The effective parameter count pLOO was reported as a diagnostic summary rather than a direct test

of misspecification. For censored patients, LOO scores the observed censoring record, not their unobserved event time.

Posterior predictive check

For each posterior draw we simulated latent event times at the observed covariate profiles,

$$T_i^{rep} = \exp(\mu_i + \sigma Z_i^{rep}). \quad (5)$$

Each replication used the parameter values from the corresponding posterior draw, with Z_i^{rep} sampled from the specified error distribution. No censoring was applied to the replicated event times. The replicated survival curve was computed as the proportion of simulated event times exceeding *t*. These curves were overlaid on the observed Kaplan–Meier (KM) estimate. The KM estimate provides a valid marginal reference under independent censoring. Because the replicated curves also include sampling variability, the overlay serves as a coarse check of the fitted survival distribution rather than a formal assessment of predictive calibration.

Sensitivity analyses

Three sensitivity analyses were run. (i) Added censoring: follow-up was truncated at the 90th, 75th, and 60th percentiles of the observed death times (3,237, 2,264, and 1,478 days), which raised censoring to 64.2%, 70.0%, and 76.1% (111, 93, and 74 deaths). This mechanism is deterministic, so there are no random realizations, and LOO values are not compared across scenarios because the scored outcomes differ. (ii) Alternative priors for ν: Gamma(2, 0.2) and Exponential(0.05), both truncated at $\nu > 1$. (iii) Event definition: death or transplantation as a composite event. This addresses a different clinical question from the primary analysis and does not show that censoring at transplantation is noninformative.

Application to the PBC data

Data

The PBC data come from the Mayo Clinic trial of D-penicillamine (Dickson et al., 1989) and are available in the R package survival. Of the 312 randomized patients, 2 were excluded because copper was missing, leaving $n = 310$ (274 women; mean age 50.0 years, SD 10.6). During follow-up (median 1,840 days, maximum 4,556), 124 patients died, 19 received a liver transplant, and 167 were alive at last contact. Transplantation was coded as censoring in the primary analysis, so 60.0% of observations were censored.

Covariates were age, sex (male = 1), log bilirubin, albumin (g/dl), log prothrombin time, edema (0 = none, 0.5 = untreated or successfully treated, 1 = persistent despite diuretics; entered as a numeric 0/0.5/1 score), and log copper. Cholesterol, triglycerides, and other laboratory variables were not used because of their missing values. The analysis was restricted to complete records, excluding 2 patients (0.6%). A missing-at-random assumption alone does not guarantee valid complete-case inference; possible selection bias remains a limitation. The standard deviations used for standardization were 10.6 years (age), 1.03 (log bilirubin; one SD is a 2.8-fold change in bilirubin), 0.42 g/dl (albumin), 0.088 (log prothrombin time), and 0.83 (log copper).

Convergence and model comparison

The reported sampling diagnostics were satisfactory: the largest $\hat{R}$ was 1.005, the smallest bulk and tail effective sample sizes were 1,445 and 2,089, there were no divergent transitions, and no refit was needed. No Pareto-k value exceeded 0.5 in any model, supporting the reliability of the PSIS approximation (Table 2).

Table 2. Predictive comparison (higher ELPD is better; ΔELPD is relative to the best model).

Model	ELPD (LOO), SE	ΔELPD (SE)	pLOO	LOOIC	WAIC
Student-t	−221.1 (16.2)	0	9.8	442.2	442.1
Normal	−225.9 (17.2)	−4.8 (3.2)	9.4	451.8	451.7
Cauchy	−242.5 (15.3)	−21.4 (5.0)	15.1	484.9	484.9

Source: Authors' analysis.

The Cauchy model was clearly worst: its ELPD was 21.4 lower than that of the Student-t model (SE 5.0, about four standard errors). Its pLOO was 15.1, compared with 9.8 and 9.4 for the Student-t and Normal models, respectively. This effective parameter count does not by itself establish that the tails are excessively heavy; the predictive comparison provides the relevant evidence against the Cauchy specification. The Student-t model was nominally best, but its advantage of 4.8 over the Normal model (SE 3.2) is only about 1.5 standard errors, and WAIC gave the same ranking. The evidence for Student-t over Normal errors is therefore modest.

Posterior estimates

Table 3. Posterior estimates and time ratios under Student-t and Normal errors.

Parameter	Student-t	Normal	Time ratio (Student-t)
Intercept	8.189 (8.032, 8.355)	8.215 (8.048, 8.406)	–
Age (per SD, 10.6 y)	−0.250 (−0.382, −0.121)	−0.243 (−0.378, −0.108)	0.78 (0.68, 0.89)
Male sex (vs female)	−0.071 (−0.418, 0.285)	0.013 (−0.364, 0.385)	0.93 (0.66, 1.33)
Log bilirubin (per SD, 1.03)	−0.434 (−0.583, −0.285)	−0.435 (−0.605, −0.276)	0.65 (0.56, 0.75)

Parameter	Student-t	Normal	Time ratio (Student-t)
Albumin (per SD, 0.42 g/dl)	0.220 (0.086, 0.359)	0.167 (0.024, 0.311)	1.25 (1.09, 1.43)
Log prothrombin (per SD, 0.088)	−0.236 (−0.362, −0.109)	−0.246 (−0.380, −0.112)	0.79 (0.70, 0.90)
Edema (0 → 1)	−0.722 (−1.143, −0.301)	−0.846 (−1.286, −0.395)	0.49 (0.32, 0.74)
Log copper (per SD, 0.83)	−0.259 (−0.414, −0.113)	−0.273 (−0.438, −0.111)	0.77 (0.66, 0.89)
Scale σ	0.735 (0.593, 0.890)	0.886 (0.778, 1.015)	–
Degrees of freedom ν	9.79 (3.39, 27.65)	–	–

Note. Coefficients are posterior means with 95% credible intervals. Time ratios are posterior medians with 95% credible intervals under Student-t errors. Source: Authors' analysis.

Higher bilirubin, edema, older age, longer prothrombin time, and higher copper were associated with shorter survival, and higher albumin with longer survival; the 95% credible intervals for these time ratios exclude one. A one-SD increase in log bilirubin corresponded to a 35% shorter survival time (time ratio 0.65, 95% CrI 0.56–0.75), and persistent edema to about half the survival time of no edema (0.49, 0.32–0.74), assuming the linear edema score. Sex showed no clear association (0.93, 0.66–1.33). These time ratios are associations, not causal effects.

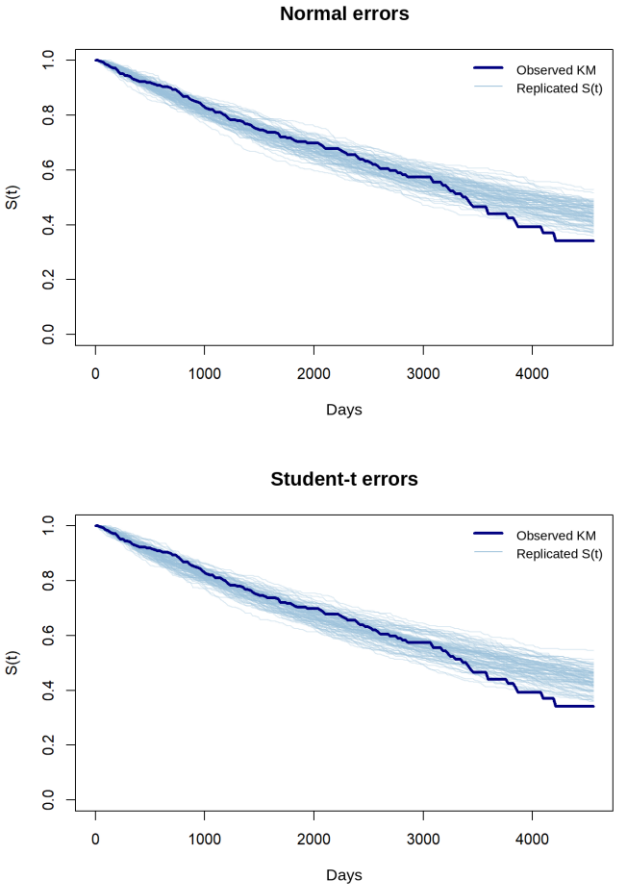
Coefficients under Normal and Student-t errors were close; the largest absolute coefficient difference was for edema (−0.722 versus −0.846), followed by sex (−0.071 versus 0.013) and albumin (0.220 versus 0.167). Under Student-t errors the displayed 95% credible intervals for regression slopes were approximately 3–9% narrower. Narrower intervals alone do not demonstrate improved frequentist coverage or estimation efficiency. The scale σ should not be compared directly as a measure of residual standard deviation across models: it equals the standard deviation under Normal errors,

whereas the Student-t standard deviation also depends on ν and exists only for $\nu > 2$. The posterior of ν was diffuse (median 7.9, 95% CrI 3.4–27.7): the data support moderately heavy tails but cannot distinguish them from near-Normal errors.

Posterior predictive check

Figure 1 overlays the observed KM curve on 100 replicated survival curves. Both models reproduce the KM curve closely up to about 3,000 days. Beyond about 3,500 days, the observed curve (about 0.34 at the end of follow-up) falls below most replicated curves under both models. Few patients remain at risk at those times, so the KM estimate is imprecise there, and censoring at transplantation may also affect it. The Student-t model is visually indistinguishable from the Normal model, in line with their small LOO difference.

Figure 1. Observed Kaplan–Meier and replicated survival curves under Normal and Student-t errors.



Note. Normal errors are shown in the upper panel and Student-t errors in the lower panel. The dark line is the observed Kaplan–Meier estimate; the light lines are 100 replicated latent survival curves generated at the observed covariate profiles without censoring. S(t) denotes the survival probability; time is measured in days. Source: Authors' analysis.

Sensitivity analyses

Table 4. Sensitivity to added censoring and to the event definition (Student-t model; posterior means with 95% credible intervals).

Scenario	Events, n	Log bilirubin	Albumin	Edema	ν
Primary (60.0% censored)	124	−0.434 (−0.583, −0.285)	0.220 (0.086, 0.359)	−0.722 (−1.143, −0.301)	9.8 (3.4, 27.7)
64.2% censored	111	−0.440 (−0.610, −0.281)	0.240 (0.093, 0.393)	−0.739 (−1.196, −0.277)	11.7 (3.7, 35.3)
70.0% censored	93	−0.437 (−0.623, −0.267)	0.227 (0.065, 0.388)	−0.810 (−1.298, −0.335)	10.8 (3.3, 33.1)
76.1% censored	74	−0.429 (−0.641, −0.230)	0.223 (0.044, 0.404)	−0.823 (−1.312, −0.323)	10.2 (3.2, 30.8)
Death or transplant (53.9% censored)	143	−0.426 (−0.581, −0.280)	0.209 (0.083, 0.338)	−0.771 (−1.187, −0.359)	12.5 (4.0, 35.2)

Source: Authors' analysis.

Table 5. Posterior of ν under alternative priors (all truncated at $\nu > 1$).

Prior for ν	Median	95% CrI	Max R-hat
Gamma(2, 0.1) (primary)	7.90	3.39 – 27.65	1.002
Gamma(2, 0.2)	6.47	3.14 – 16.94	1.003
Exponential(0.05)	6.87	3.13 – 29.23	1.002

Source: Authors' analysis.

Added censoring. The bilirubin and albumin coefficients were stable as censoring increased from 60.0% to 76.1%, whereas their intervals widened (the 95% CrI width for log bilirubin rose from 0.30 to 0.41, about 38%). The edema coefficient drifted from −0.72 to −0.82, still within the primary interval. The scale σ ranged from 0.74 to 0.80, and ν remained diffuse (posterior means 9.8–11.7) without a systematic trend.

Prior for ν. The lower end and the median of ν were similar across the three priors (medians 6.5–7.9; lower limits 3.1–3.4), but the upper limit depended on the prior (16.9–29.2). The data therefore inform mainly the lower tail of ν, while the upper end, which corresponds to near-Normal behaviour, is prior-dependent.

Event definition. Counting transplantation as an event (143 events, 53.9% censoring) left the bilirubin and albumin coefficients essentially unchanged and moved the edema coefficient to −0.77, while the posterior mean of ν rose to 12.5. This is a different estimand and should be reported separately from the primary results.

Discussion

In this application the error distribution mattered little for the substantive conclusions, but it did matter at the extreme: the Cauchy model was clearly inferior, while Student-t errors gave at most a small and uncertain predictive gain over Normal errors. Fitting a Student-t model is nevertheless a practical sensitivity check. It left the coefficients nearly unchanged and narrowed the intervals slightly, and the diffuse posterior of ν documents how little the data say about tail thickness. These findings are consistent with the general caution that heavy-tailed errors should not be presumed superior but assessed empirically (Lange et al., 1989; Rubio & Hong, 2016).

Several limitations apply. First, this is a single dataset without a simulation study, so the results do not show when Student-t errors help. Second, the PSIS-LOO comparison scores observed event and censoring records, not unobserved event times, and the KM overlay assumes independent censoring. Censoring at transplantation may be informative if transplantation depends on disease severity, and the composite endpoint changes the question rather than removing this concern. Third, symmetric error

distributions cannot capture skewness of log survival time. The marginal distribution of log death times is left-skewed (sample skewness about −0.96), although the conditional residuals were not examined, so skewed alternatives or other lifetime families could be compared. Fourth, effects were assumed linear, edema was entered as a numeric score, and only a subset of covariates was used in a complete-case analysis. Finally, the added-censoring scenarios are deterministic truncations of follow-up and do not represent random censoring.

Conclusion

Using the PBC data, we compared Normal, Student-t, and Cauchy errors in a Bayesian AFT model with right censoring. Cauchy errors were clearly worse. Student-t errors were nominally best but gave only a modest, uncertain predictive improvement over Normal errors, with a diffuse posterior for ν. Covariate effects were robust to the error distribution, to added censoring up to 76%, to the prior for ν, and to the event definition. We recommend choosing the error distribution by predictive criteria and sensitivity analyses, and reporting priors, diagnostics, and the dependence of ν on its prior, rather than defaulting to a heavy-tailed model.

References

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Cengiz, M. A., Şenel, T., & Kara, M. (2025). Bayesian robust symmetric regression for medical data with heavy-tailed errors and censoring. *PLOS ONE, 20*(8), Article e0329589. https://doi.org/10.1371/journal.pone.0329589

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D., & Langworthy, A. (1989). Prognosis in primary biliary cirrhosis: Model for decision making. *Hepatology, 10*(1), 1–7. https://doi.org/10.1002/hep.1840100102

Fonseca, T. C. O., Ferreira, M. A. R., & Migon, H. S. (2008). Objective Bayesian analysis for the Student-t regression model. *Biometrika, 95*(2), 325–333. https://doi.org/10.1093/biomet/asn001

Fonseca, T. C. O., Ferreira, M. A. R., & Migon, H. S. (2014). Amendments and corrections: "Objective Bayesian analysis for the Student-t regression model." *Biometrika, 101*(1), 252. https://doi.org/10.1093/biomet/asu001

Hamura, Y., Onizuka, T., Hashimoto, S., & Sugasawa, S. (2025). *Robust Bayesian inference for censored survival models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2504.11147

Lange, K. L., Little, R. J. A., & Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association, 84*(408), 881–896. https://doi.org/10.1080/01621459.1989.10478852

Rubio, F. J., & Hong, Y. (2016). Survival and lifetime data analysis with a flexible class of distributions. *Journal of Applied Statistics*, *43*(10), 1794–1813. https://doi.org/10.1080/02664763.2015.1120710

Rubio, F. J., & Steel, M. F. J. (2018). Flexible linear mixed models with improper priors for longitudinal and survival data. *Electronic Journal of Statistics*, *12*(1), 572–598. https://doi.org/10.1214/18-EJS1401

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., & Gabry, J. (2024). Pareto smoothed importance sampling. *Journal of Machine Learning Research*, *25*(72), 1–58. https://www.jmlr.org/papers/v25/19-556.html

CHAPTER 4

DEĞİŞEN AİLE, YAVAŞLAYAN NÜFUS: TÜRKİYE'NİN 16 YILLIK DEMEOGRAFİK EĞİLİMLERİ

CENGİZ GAZELOĞLU[1]

Giriş

Demografik dönüşüm, modern toplumların sosyo-ekonomik ve kültürel yapısını derinden şekillendiren, çok boyutlu ve dinamik bir süreçtir (Inglehart & Baker, 2000; Lesthaeghe, 2010). Bu dönüşüm sürecinin makro düzeydeki en hayati göstergelerinin başında, aile kurumunun oluşumu ve çözülmesini ifade eden evlenme ve boşanma eğilimleri ile bu eğilimlerin zamanlaması gelmektedir (Becker, 1993; Kalmijn, 2007). İkinci Dünya Savaşı sonrasında belirginleşen ve günümüzde "ikinci demografik geçiş" olarak adlandırılan küresel eğilim, ilk evlenme yaşının gecikmesi, kaba evlenme hızlarının düşmesi ve boşanma oranlarının artması gibi temel özelliklerle karakterize edilmektedir (Sobotka & Toulemon, 2008; Thornton, 2001). Bireyselleşme, kadınların iş gücüne ve yükseköğrenime katılım oranlarındaki artış ile sekülerleşme gibi faktörlerin tetiklediği bu değişim, geleneksel aile

[1] Prof. Dr., Süleyman Demirel Üniversitesi, İstatistik Bölümü, Orcid: 0000-0002-8222-3384

yapılarını ve bireysel yaşam döngülerini köklü bir biçimde dönüştürmektedir (Cherlin, 2004; Giddens, 1992).

Türkiye de bu küresel demografik eğilimlerden bağımsız değildir ve son yıllarda belirgin bir değişim ivmesi sergilemektedir (Behar, 1995). Kentleşme hızının artması ve sosyo-kültürel modernleşme süreçlerinin bir yansıması olarak, Türkiye'de hem erkekler hem de kadınlar için ilk evlilik yaşı istikrarlı bir şekilde ileriye kaymaktadır (Ergöçmen & Çağatay, 2017). Literatürde yer alan araştırmalar, evlilik yaşının ilerlemesinin evlenme oranlarında genel bir azalmayla ilişkili olduğunu göstermektedir. Öte yandan, boşanmanın toplumsal açıdan kabulüne ilişkin normların değişmesi ve esnemesi, boşanma hızlarının belirgin biçimde yükselmesine zemin hazırlamaktadır (Koç, Hancıoğlu, & Çavlin, 2010).

Bir ülkenin nüfus artış hızı doğrudan bir bağa sahiptir aile oluşumu (Bongaarts, 1998). Aile oluşumunda kadının doğurganlık hızı ve aktif evlenme süresi nüfus artış hızını olumlu ve olumsuz yönde etkilemektedir. Eğer kadın doğurganlığının yüksek olduğu zamanda evlenirse ve çocuk sahibi olursa nüfus artış hızı olumlu yönde etkilenirken bu durumun tam tersi olumsuz etkilemektedir (Véron, 2002; Raymo, Park, Xie, & Yeung, 2015).

Türkiye'deki demografik göstergeler incelendiğinde, nüfus artış hızında son yıllarda gözlemlenen yavaşlama eğiliminin, evlenme oranlarındaki düşüş ve boşanma oranlarındaki artış örüntüleriyle yüksek düzeyde ilişkili olduğu görülmektedir (Erdem, 2022). Türkiye bağlamında, nüfus artış hızında son yıllarda gözlemlenen hızlı yavaşlama eğilimi, evlilik ve boşanma örüntülerindeki yapısal değişimlerle yüksek düzeyde korelasyon göstermektedir (Güneş, 2025). Buna ek olarak, 2020 yılında ortaya çıkan COVID-19 küresel salgını gibi dışsal şoklar, demografik serilerde geçici ancak son derece sert anomaliler yaratmıştır. Resmi kurumsal hizmetlerin ve toplumsal hayatın kısıtlanmasına bağlı olarak ilk aşamada evlenme ve boşanma hızlarında ani düşüşler

gözlemlenmiş; kısıtlamaların kaldırılmasının ardından ise ertelenen taleplerin devreye girmesiyle ani tepki artışları (patlamalar) yaşanarak zaman serisi verilerinde yapısal kırılmalar meydana gelmiştir (Şahinalp, 2024).

Kavramsal ve ampirik literatür, demografik değişkenlerin tekil olarak incelenmesinin yanı sıra, birbirleriyle olan eşzamanlı ve doğrusal etkileşimlerinin boylamsal (longitudinal) olarak izlenmesinin gerekliliğini vurgulamaktadır. Bu bağlamda, güncel veriler ışığında Türkiye'nin demografik göstergelerinin senkronize hareketini inceleyen analitik çalışmalara duyulan ihtiyaç devam etmektedir.

Bu çalışmanın amacı; Türkiye'de 2010-2025 yıllarını kapsayan on altı yıllık dönemde, kaba evlenme hızı, kaba boşanma hızı, cinsiyete göre ilk evlenme yaşları ve nüfus artış hızı arasındaki doğrusal ilişkileri ve yıllık eğilimleri incelemektir. Türkiye İstatistik Kurumu (TÜİK, 2026) veri tabanından elde edilen makro veriler kullanılarak gerçekleştirilen bu keşifsel korelasyon analizi, değişkenler arası etkileşim yönlerini ve şiddetini istatistiksel olarak ortaya koymayı hedeflemektedir. Bu çalışmadan elde edilecek sonuçlar, ileride yapılması planlanan yapısal eşitlik modeli ve ileri düzey istatistiksel analizler ile Türkiye'nin demografik değişkenler açısından oluşacak dönüşümler için bir zemin oluşturacağı öngörülmektedir.

Materyal ve Yöntemler

Veri Seti & Değişkenler

16 yıllık dönemi kapsayan (2010-2025) yılları arası Türkiye'nin bazı demografik özellikleri incelenmiştir. Araştırma için kullanılan verileri, Türkiye İstatistik Kurumunun resmi web sitesinden elde edilmiştir. Çalışmada kullanılan değişkenler erkek ve kadınların ilk evlendikleri yaşlar, boşanma oranları, kaba evlenme

hızı ve nüfus artış hızı yer almaktadır. Ayrıca evlenme yaşları hariç diğer değişkenlere ait değerler binde bir üzerinden hesaplanmıştır.

Zaman serisi analizlerinde, nüfustaki farklılıklardan kaynaklanabilecek yanıltıcı etkilere (sahte korelasyon) karşı dikkatli olunması büyük önem taşımaktadır (Preston, Heuveline, & Guillot, 2000). Bu durumu önlemek adına, toplam sayıların doğrudan kullanılması yerine, nüfusa oranlanmış kaba hızlar analizlere dahil edilmiştir

Veri Analizi Süreci

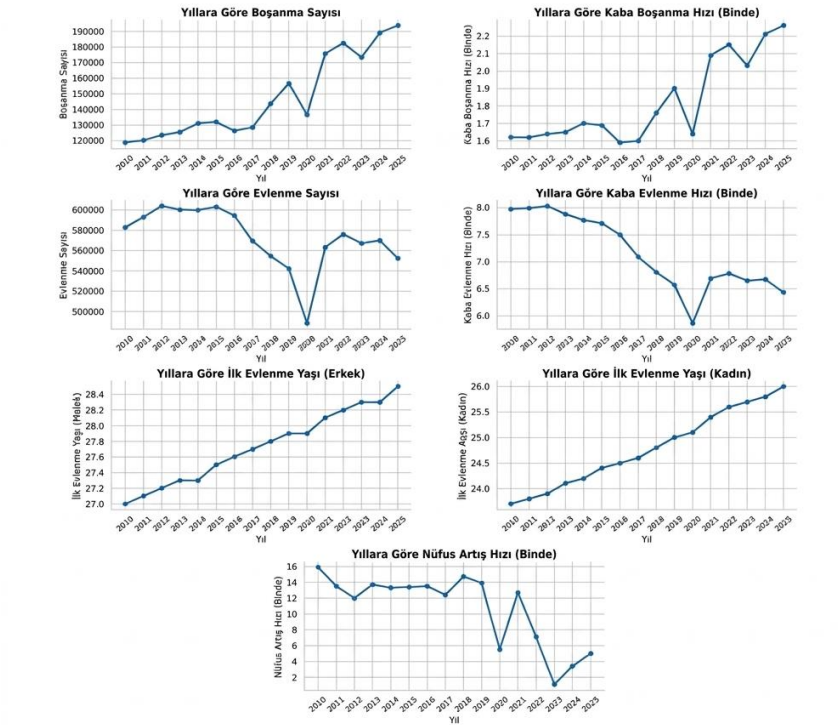
Araştırma verilerinin analizi iki farklı düzeyde yürütülmüştür. Başlangıçta; veri setinin sınırları, dağılım karakteristikleri ve merkezi eğilimleri ortalama, medyan ve standart sapma gibi temel betimsel metriklerle özetlenmiştir (Field, 2018). Bu özetlemelere ek olarak, özellikle salgın dönemi gibi sıra dışı etkilerin yaşandığı kırılma noktalarını saptayabilmek ve genel eğilimi görselleştirebilmek için yıllık oransal değişimler (year-over-year change) hesaplanmıştır.

Değişkenler arasındaki ilişkilerin doğasını test etmeye yönelik ikinci aşamada Pearson Korelasyon Analizi uygulanmıştır. Çalışmadaki sürekli demografik değişkenlerin doğrusal etkileşimlerini ve bunların matris formatını belirlemek amacıyla standart yöntem olan Pearson korelasyon katsayısından faydalanılmıştır (Pallant, 2020). Katsayıların etki büyüklüğü (effect size) değerlendirmesinde Cohen'in (1988) sosyal bilimler literatürüne kazandırdığı aralıklar ölçüt alınırken, istatistiksel anlamlılık (significance) kararlarında %95 ve %99 güvenirlilik düzeyi kullanılmıştır. Bunlara ek olarak ham verilerin düzenlenmesinde ve çalışmanın analiz kısmını oluşturan ısı haritası ile korelasyon analizi için Python programı kullanılmıştır.

Sonuç ve Tartışma

Çalışmanın bu bölümde, verilerin kapsadığı 16 yıllık dönemde demografik özelliklere (kaba evlenme hızı, kaba boşanma hızı ve nüfus artış hızı) ait betimsel istatistikler ile Pearson korelasyon analizi sonuçları verilmiştir. Elde edilen istatistiksel bulgular, ilgili literatür ile desteklenmiştir.

Şekil 1. Türkiye'deki Demografik Eğilimlerin Boylamsal Analizi (2010-2025)



Tablo 1. Değişkenlere Ait Tanımlayıcı İstatistikler

Değişken	En Düşük	En Yüksek	Ortalama	Medyan	Standart Sapma
Boşanma Sayısı	118,568	193,793	147,23	134,2	26,796.6

Kaba Boşanma Hızı (Binde)	1.59	2.26	1.82	1.69	0.24
Evlenme Sayısı	488,335	603,751	572,457	572,937	29,762.5
Kaba Evlenme Hızı (Binde)	5.86	08.03	7.15	6.95	0.68
İlk Evlenme Yaşı (Erkek)	27.0	28.5	27.73	27.75	0.47
İlk Evlenme Yaşı (Kadın)	23.7	26.0	24.79	24.70	0.75
Nüfus Artış Hızı (Binde)	1.10	15.90	10.69	13.00	4.61

Tablo 1'de çalışmada kullanılan değişkenlere ilişkin tanımlayıcı istatistikler verilmiştir. İncelenen yıllarda kaba boşanma hızı binde 1,59 ile 2,26 arasında değişmiş, dönem ortalaması binde 1,82 olmuştur. Evlenme sayısının en düşük olduğu yıl 488.335, en yüksek olduğu yıl ise 603.751 evlilik kaydedilmiştir. Yıllık evlenme sayısı ortalama 572.457'dir. İlk evlenme yaşı erkeklerde kadınlara göre daha yüksektir. Erkeklerde bu değer için hesaplanan ortalama 27,73 iken kadınlarda 24,70'tir.

Tablo 2. Değişkenlerin Yıllık Değişim Oranları

Yıl	**Kaba Boşanma Hızı Değişimi (%)**	**Kaba Evlenme Hızı Değişimi (%)**	**Nüfus Artış Hızı Değişimi (%)**
2011	0.00	0.25	-15.09
2012	1.23	0.50	-11.11
2013	0.61	-1.87	14.17
2014	03.03	-1.40	-2.92
2015	-0.59	-0.77	0.75
2016	-5.92	-2.72	0.75

2017	0.63	-5.47	-8.15
2018	10.00	-3.95	18.55
2019	7.95	-3.52	-5.44
2020	-13.68	-10.81	-60.43
2021	27.44	14.16	130.91
2022	2.87	1.35	-44.09
2023	-5.58	-1.92	-84.51
2024	8.87	0.30	209.09
2025	2.26	-3.60	47.06

Tablo 2, her bir göstergenin bir önceki yıla göre yüzde (%) kaç oranında arttığını veya azaldığını gösterir. Tabloda sadeleşme adına 3 ana değişkene yer verilmiştir: Boşanma Hızı, Evlenme Hızı ve Nüfus Artış Hızı. Yıllık değişim analizinde en çok dikkat çeken yıl 2020 (Pandemi yılı) olarak öne çıkıyor. 2020'de kaba evlenme hızı bir önceki yıla göre %10.81, boşanma hızı ise %13.68 oranında ani bir düşüş yaşamış, nüfus artış hızı ise %60'ın üzerinde sert bir fren yapmıştır. Hemen ardından, "ertelenen taleplerin (evlilik/boşanma) devreye girmesiyle" 2021 yılında boşanma hızı %27, evlenme hızı %14, nüfus artış hızı ise %130 (binde 5.5'ten binde 12.7'ye çıkış) oranında çok sert bir yukarı yönlü düzeltme (tepki artışı) yaşamıştır.

Şekil 2. Demografik Göstergelerin Korelasyon Matrisi



Üç temel demografik gösterge arasındaki doğrusal ilişkileri ortaya koyan Pearson korelasyon analizi, değişkenler arasında dikkat çekici ilişkiler bulunduğunu göstermektedir. En güçlü ilişki, kaba boşanma hızı ile nüfus artış hızı arasında negatif yönde (image) belirlenmiştir. Bu sonuç, nüfus artış hızının düştüğü dönemlerde boşanma hızının yükselme eğilimi gösterebildiğini ortaya koymaktadır. Kaba evlenme hızı ile nüfus artış hızı arasında ise orta-yüksek düzeyde pozitif (image) bir ilişki tespit edilmiştir. Buna göre, evlenme hızının görece yüksek olduğu dönemlerde nüfus artış hızının da yükseldiği görülmektedir. Aile yapısıyla ilişkili iki gösterge arasındaki ilişki incelendiğinde, kaba boşanma hızı ile kaba evlenme hızı arasında orta düzeyde negatif (image) bir korelasyon olduğu belirlenmiştir. Bu bulgu, evlenme hızının azaldığı dönemlerde boşanma hızının artış eğilimi gösterebildiğine işaret etmektedir.

Tablo 3. Değişkenler Arasındaki Korelasyon Katsayılarının Anlamlılık Düzeyi

Değişken 1	**Değişken 2**	**Korelasyon (r)**	**P-Değer**	**Önemlilik Seviyesi**
Crude Divorce Rate	Population Growth Rate	−0.66	0.0054	Highly Significant ($p<0.01$)

Crude Marriage Rate	Population Growth Rate	0.64	0.0073	Highly Significant ($p<0.01$)
Crude Divorce Rate	Crude Marriage Rate	−0.59	0.0152	Significant ($p<0.05$)

Tablo 3, üç temel demografik değişken arasındaki korelasyon katsayılarını ve bu ilişkilerin istatistiksel anlamlılık düzeylerini (p-değerlerini) detaylandırmaktadır.

Analiz sonuçları incelendiğinde, ölçülen tüm ilişkilerin istatistiksel olarak geçerli (anlamlı) olduğu görülmektedir. Kaba boşanma hızı ile nüfus artış hızı arasındaki negatif korelasyon ($r = -0.66$) ve kaba evlenme hızı ile nüfus artış hızı arasındaki pozitif korelasyon ($r = 0.64$), $p < 0.01$ seviyesinde (sırasıyla $p = 0.0054$ ve $p = 0.0073$) yüksek düzeyde anlamlı bulunmuştur. Bu değerler, değişkenler arasındaki söz konusu etkileşimlerin %99 güven düzeyinde şans eseri olmadığını ve istatistiksel açıdan çok güçlü bir şekilde kanıtlandığını göstermektedir. Öte yandan, kaba boşanma hızı ile kaba evlenme hızı arasındaki orta düzeyli negatif ilişki ($r = -0.59$), $p = 0.0152$ değeri ile $p < 0.05$ sınırları içerisinde kalmış ve istatistiksel olarak anlamlı kabul edilmiştir.

Özetle, tablo 3 verileri, incelenen demografik göstergelerin birbirleriyle tesadüfi olmayan, istatistiksel standartlar çerçevesinde doğrulanmış bağlara sahip olduğunu bilimsel olarak ortaya koymaktadır.

Kaynakça

Becker, G. S. (1993). A treatise on the family. Harvard University Press.

Behar, C. (1995). Türkiye'nin demografik dönüşümü. T.C. Başbakanlık Devlet İstatistik Enstitüsü.

Bongaarts, J. (1998). Fertility and reproductive preferences in post-transitional societies. Population and Development Review, 27, 260–281. https://doi.org/10.1111/j.1728-4457.2001.00260.x

Cherlin, A. J. (2004). The deinstitutionalization of American marriage. Journal of Marriage and Family, 66(4), 848–861. https://doi.org/10.1111/j.0022-2445.2004.00058.x

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Ergöçmen, B. A., & Çağatay, P. (2017). Türkiye'de evliliklerin zamanlaması ve evlilik biçimleri. İçinde Türkiye Nüfus ve Sağlık Araştırması 2018 ileri analiz raporu (ss. 45–67). Hacettepe Üniversitesi Nüfus Etütleri Enstitüsü.

Field, A. (2018). Discovering statistics using IBM SPSS statistics (5th ed.). SAGE Publications.

Giddens, A. (1992). The transformation of intimacy: Sexuality, love and eroticism in modern societies. Stanford University Press.

Güneş, E. (2025). Türk Toplumunda Boşanma Olgusu: Sosyoloji ve Sosyal Hizmet Perspektifinden Bir Değerlendirme. Habitus Toplumbilim Dergisi, Toplumsal Tedkikler, 281-304. https://doi.org/10.62156/habitus.1784327

Inglehart, R., & Baker, W. E. (2000). Modernization, Cultural Change, and the Persistence of Traditional Values. *American Sociological Review*, *65*(1), 19–51. https://doi.org/10.2307/2657288

Kalmijn, M. (2007). Explaining cross-national differences in marriage, cohabitation, and divorce in Europe, 1990–2000. *Population Studies*, *61*(3), 243–263. https://doi.org/10.1080/00324720701571806

Koç, İ., Hancıoğlu, A., & Çavlin, A. (2010). Türkiye'nin demografik dönüşümü: Doğurganlık, aile planlaması, anne-çocuk sağlığı ve beş yaş altı ölümlerdeki değişimler. Hacettepe Üniversitesi Nüfus Etütleri Enstitüsü.

Lesthaeghe, R. (2010). The unfolding story of the second demographic transition. Population and Development Review, 36(2), 211–251. https://doi.org/10.1111/j.1728-4457.2010.00328.x

Pallant, J. (2020). SPSS survival manual: A step by step guide to data analysis using IBM SPSS (7th ed.). McGraw-Hill Education.

Raymo, J. M., Park, H., Xie, Y., & Yeung, W. J. J. (2015). Marriage and family in East Asia: Continuity and change. Annual Review of Sociology, 41, 471–492. https://doi.org/10.1146/annurev-soc-073014-112428

Şahinalp, M. S. (2024). COVID-19 pandemisi döneminde Türkiye'de yaşanan boşanmalar üzerine bir değerlendirme. Fırat University Journal of Social Sciences, 34(3), 1115–1130. https://doi.org/10.18069/firatsbed.1499734

Sobotka, T., & Toulemon, L. (2008). Overview chapter 4: Changing family and partnership behaviour: Common trends and persistent diversity across Europe. Demographic Research, S7(6), 85–138. https://doi.org/10.4054/DemRes.2008.19.6

Thornton, A. (2001). The developmental paradigm, reading history sideways, and family change. Demography, 38(4), 449–465. https://doi.org/10.1353/dem.2001.0039

Türkiye İstatistik Kurumu. (2026). Evlenme ve boşanma istatistikleri veri tabanı. https://data.tuik.gov.tr/

Véron, J. (2002). [Review of *Demography, Measuring and Modeling Population Processes*, by S. H. Preston, P. Heuveline, &

M. Guillot]. *Population (French Edition)*, *57*(3), 591–592. https://doi.org/10.2307/1535065


BİDGE