

Partial Least Squares Regression

Test name: Partial Least Squares Regression

Dataset: D:\DATA ANALYSIS\H Regression Tests and Models\Partial Least Squares Regression\dataset.csv

Rows in original dataset: 649

Columns in original dataset: 33

Rows used in PLS model: 649

Outcome variable: G3

Numeric predictors: G1, G2, studytime, failures, absences, age, Medu, Fedu

Categorical predictors: school, sex, address, famsize

Design matrix columns after dummy coding: 12

Selected PLS components: 5

R squared: 0.8515

Adjusted R squared using components: 0.8503

RMSE: 1.2442

MAE: 0.7843

Best cross-validation RMSE: 0.3905

Interpretation: PLS components are supervised latent dimensions chosen to predict the outcome.

Coefficient guide: larger absolute coefficients show stronger directional contribution to the fitted outcome.

VIP guide: values near or above 1 are commonly interpreted as influential predictors.

Reporting guide: state outcome, predictors, selected components, R squared, RMSE, coefficients, VIP scores, and residual diagnostics.

Charts generated: 8

Model fit

model	n	design_matrix	selected_columns	r_squared	adjusted_r_squared	rmse	mae	sse	sst	best_cv_rmse	intercept
Partial Least	649.0000	12.0000	5.0000	0.8515	0.8503	1.2442	0.7843	1004.6828	6763.2666	0.3905	-0.0755

Component selection

n_components	cv_rmse	cv_rmse_sd
1.0000	0.5060	0.0942
2.0000	0.4196	0.1004
3.0000	0.4070	0.0973
4.0000	0.3944	0.0992
5.0000	0.3905	0.0909

Top coefficients

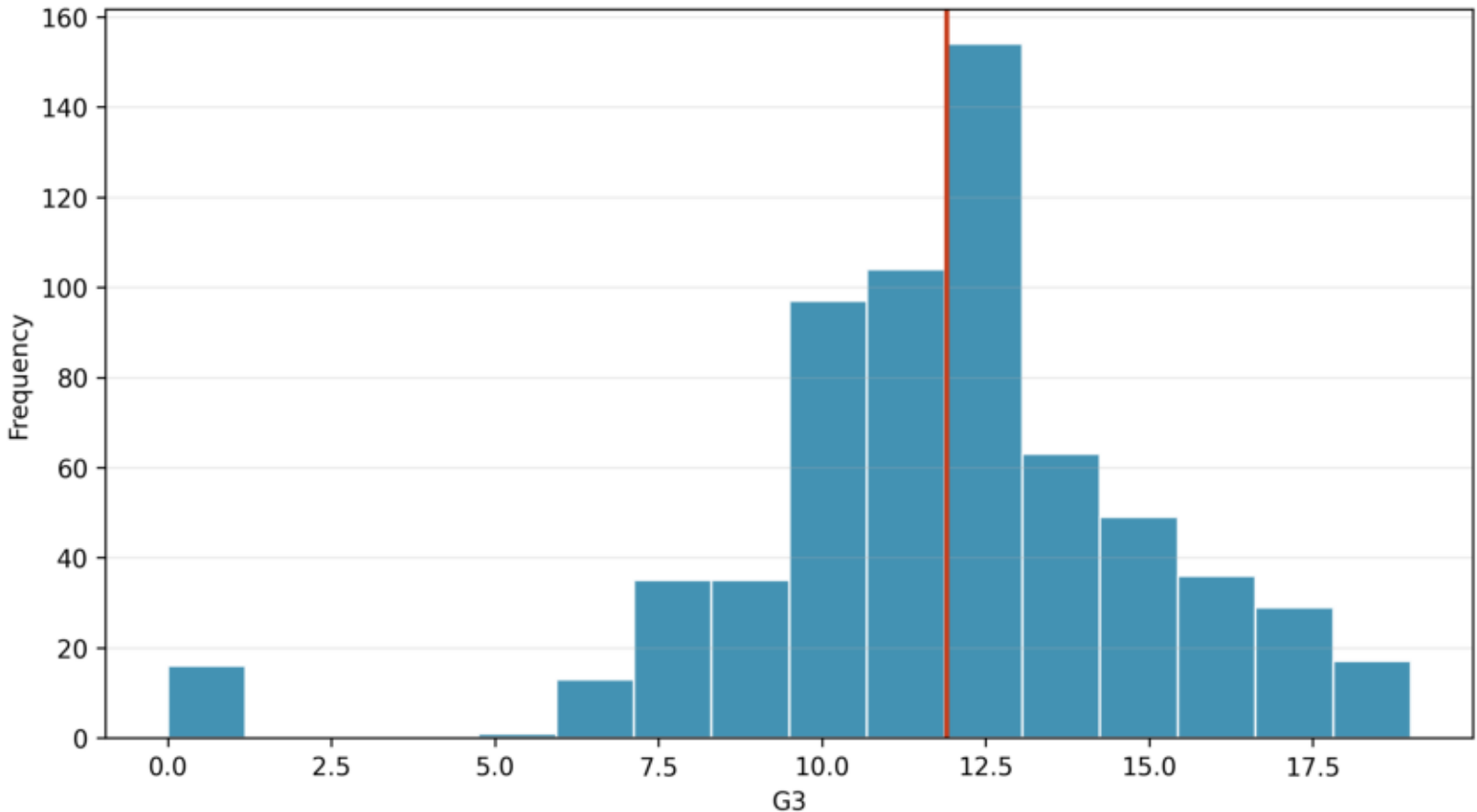
predictor	coefficient	coefficient_abs
G2	0.8592	0.8592
address_U	0.2308	0.2308
sex_M	-0.2033	0.2033
G1	0.1657	0.1657
failures	-0.1330	0.1330
studytime	0.1133	0.1133
school_MS	-0.0800	0.0800
Medu	-0.0760	0.0760
famsize_LE3	0.0547	0.0547
Fedu	0.0394	0.0394
absences	0.0264	0.0264
age	-0.0060	0.0060

Top VIP scores

predictor	vip_score
G2	2.1711
G1	1.9192
failures	0.9546
Medu	0.7508
school_MS	0.7380
Fedu	0.6778
studytime	0.6584
age	0.4770
address_U	0.4705
sex_M	0.3580
absences	0.3077
famsize_LE3	0.1365

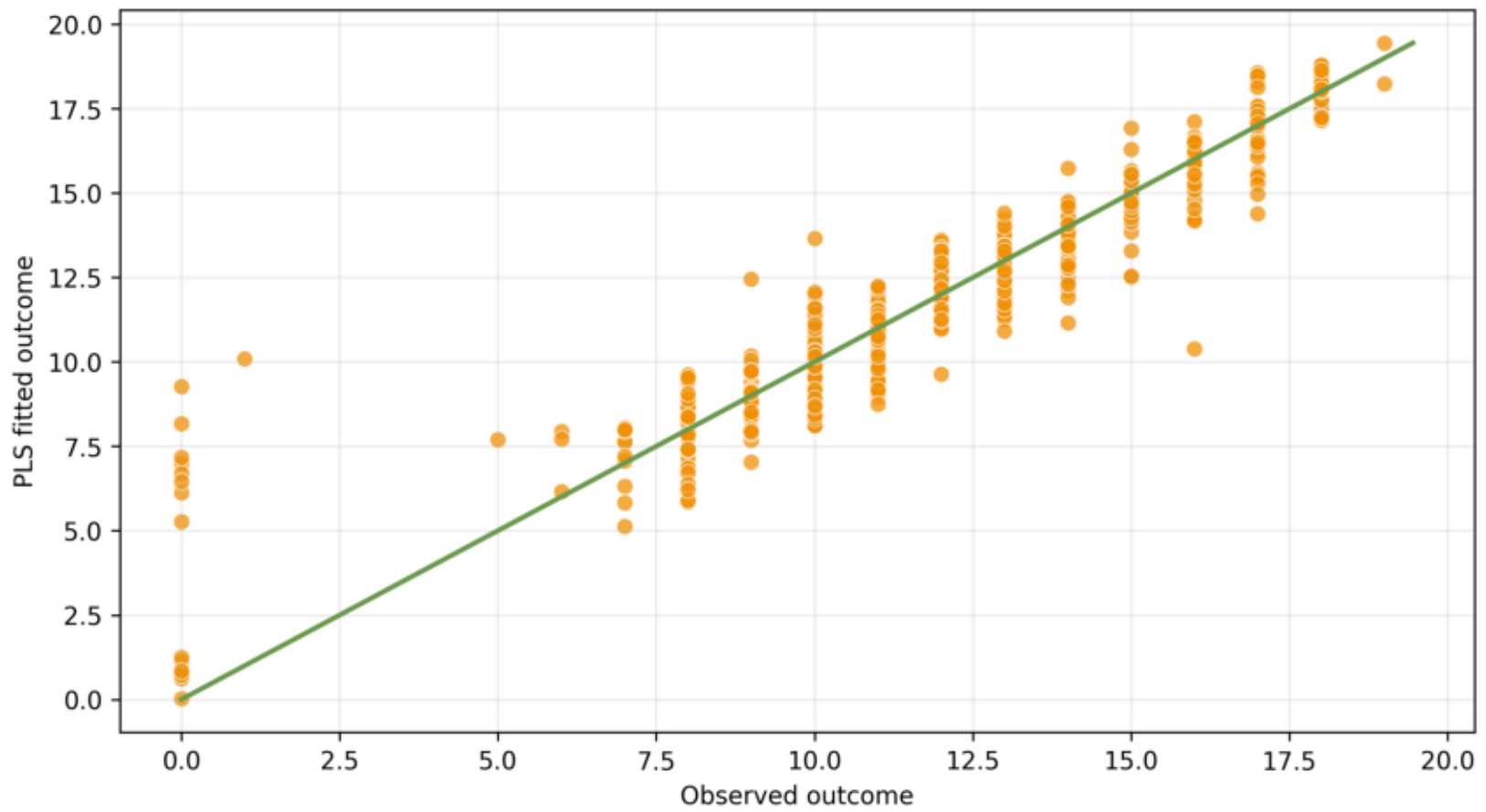
Outcome distribution

Distribution of the dependent variable used in the PLS regression model.



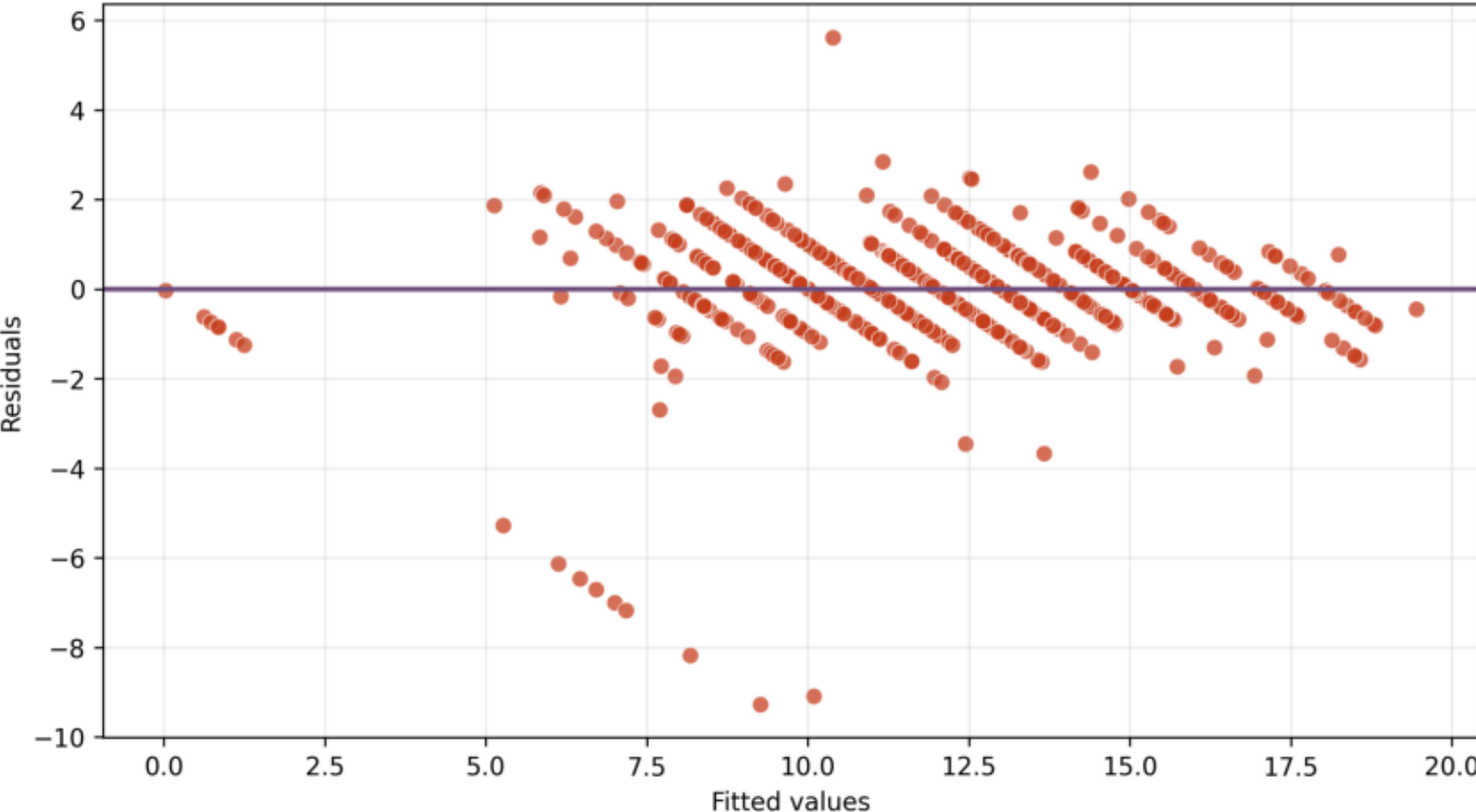
Observed vs fitted values

Points near the diagonal indicate stronger predictive fit.



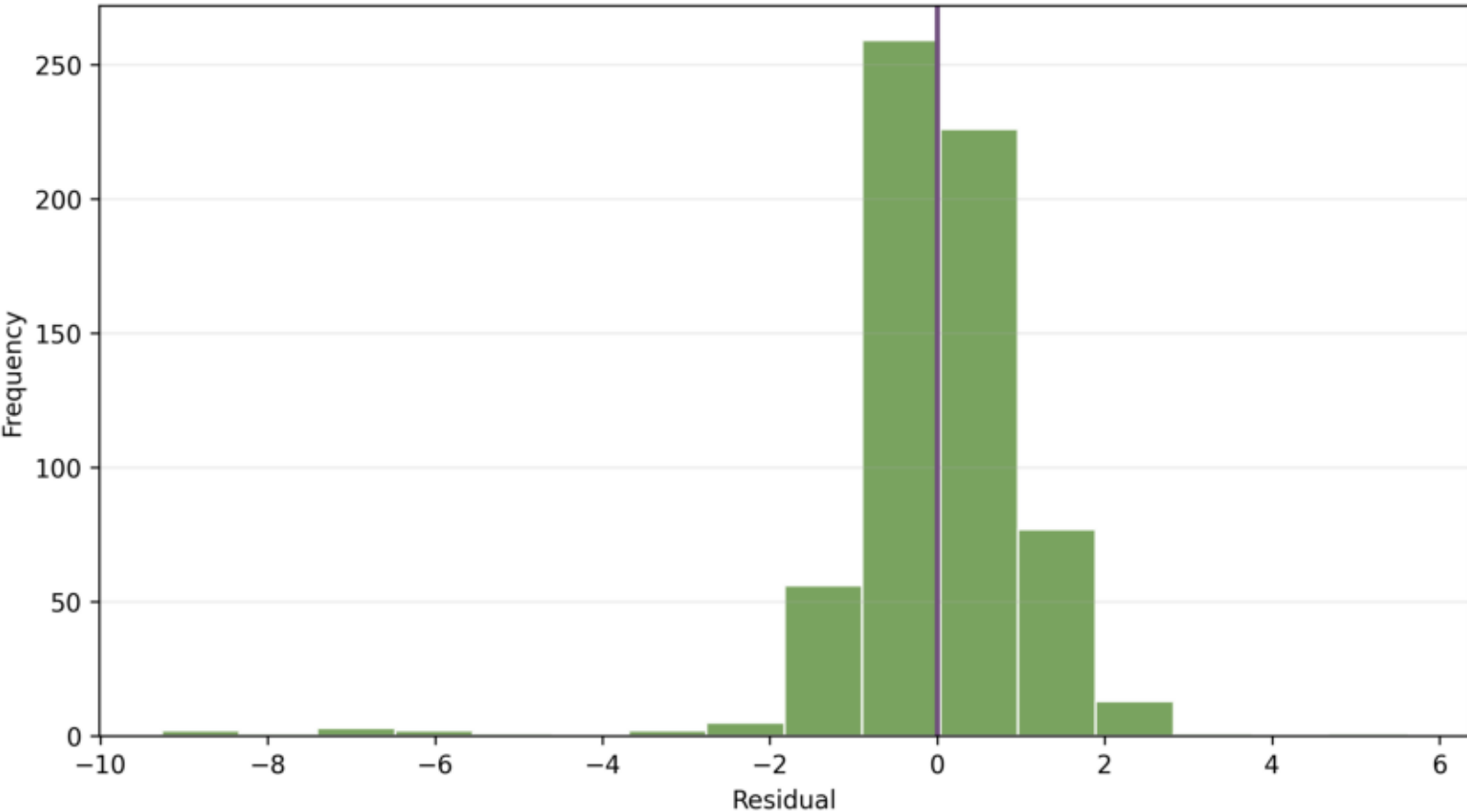
Residuals vs fitted values

Residual patterns help diagnose nonlinearity, heteroscedasticity, and outliers.



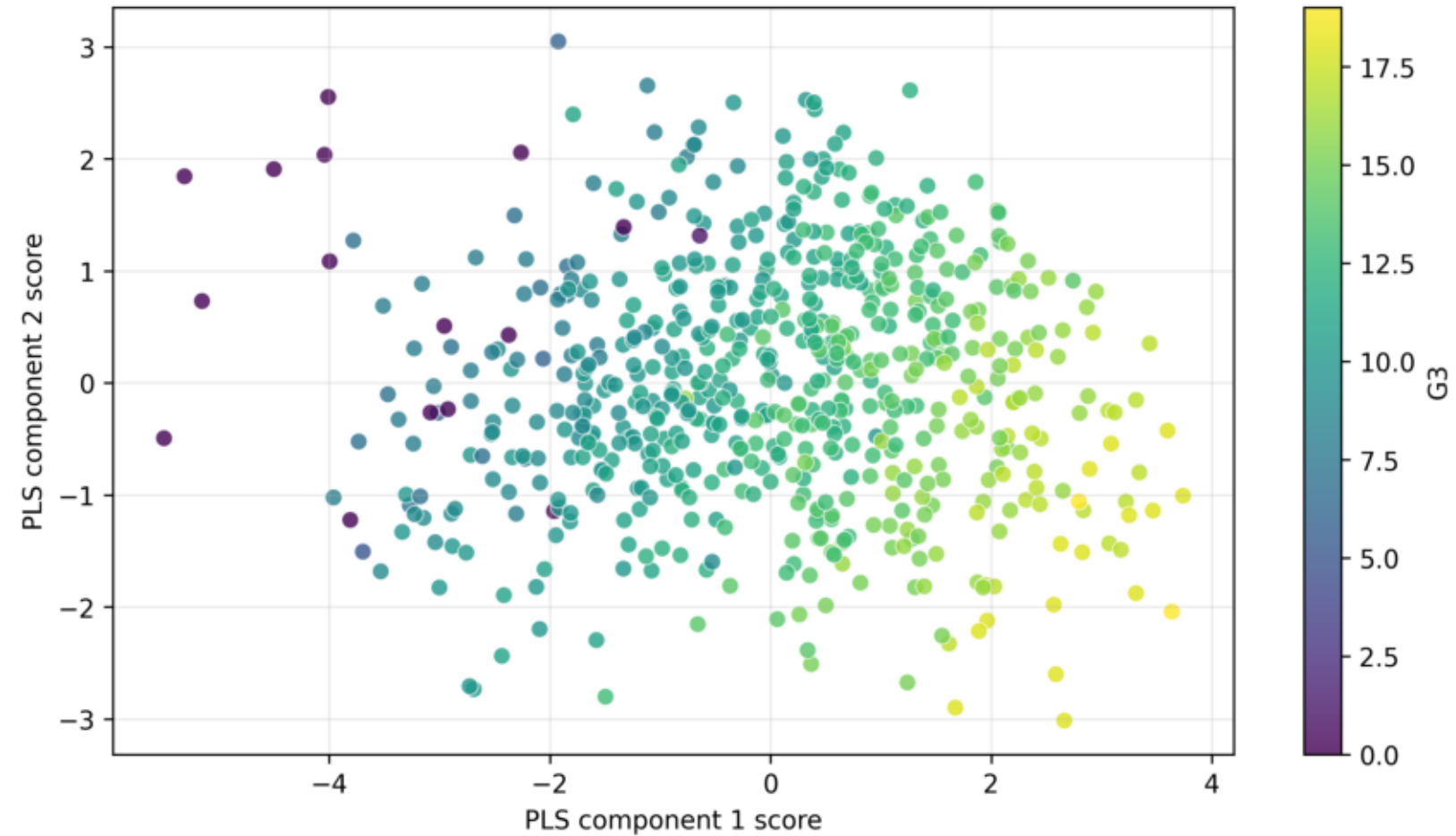
Residual distribution

A centered residual distribution supports stable prediction.



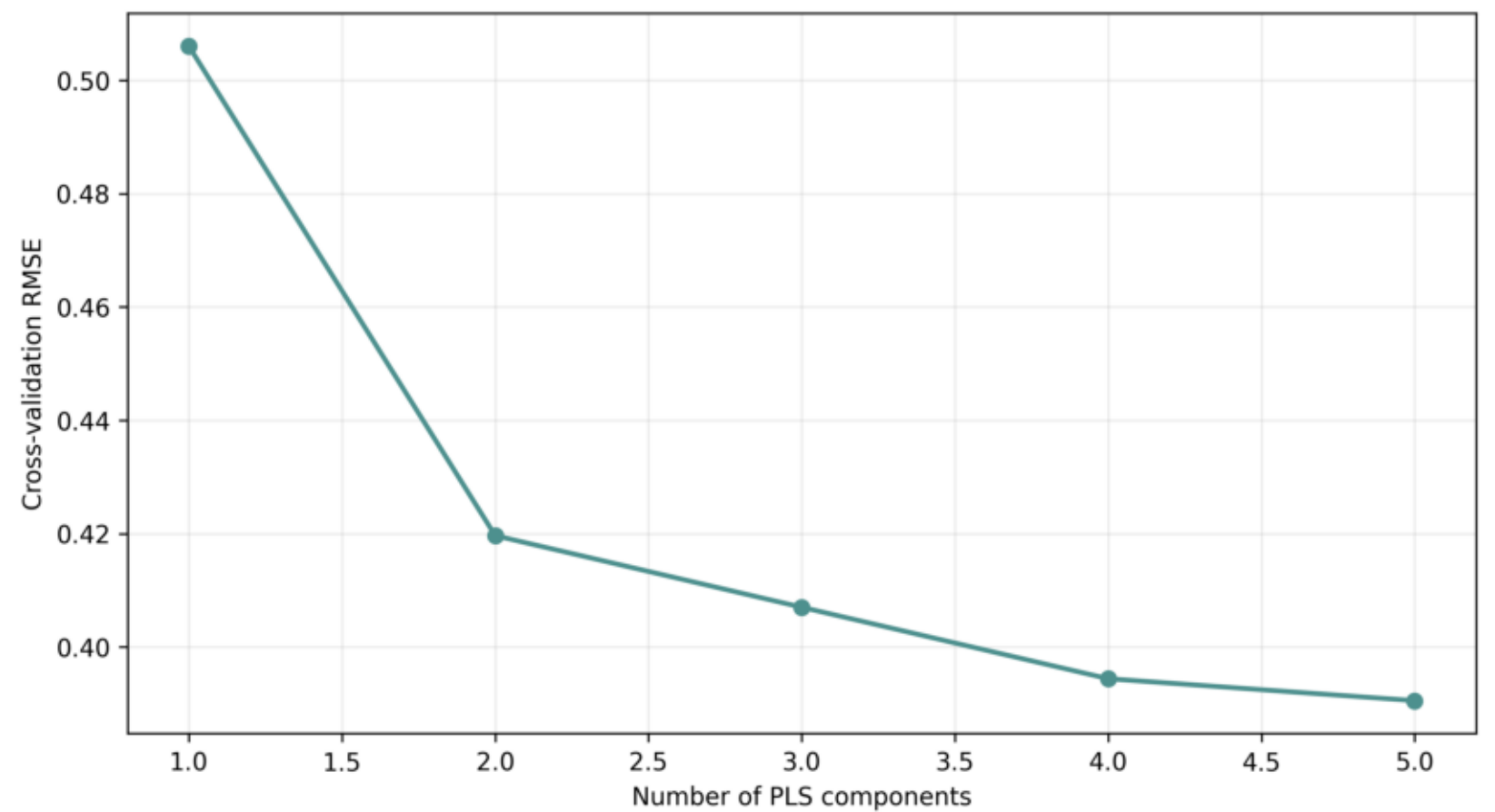
PLS component score map

Scores show how cases spread across the latent PLS components.



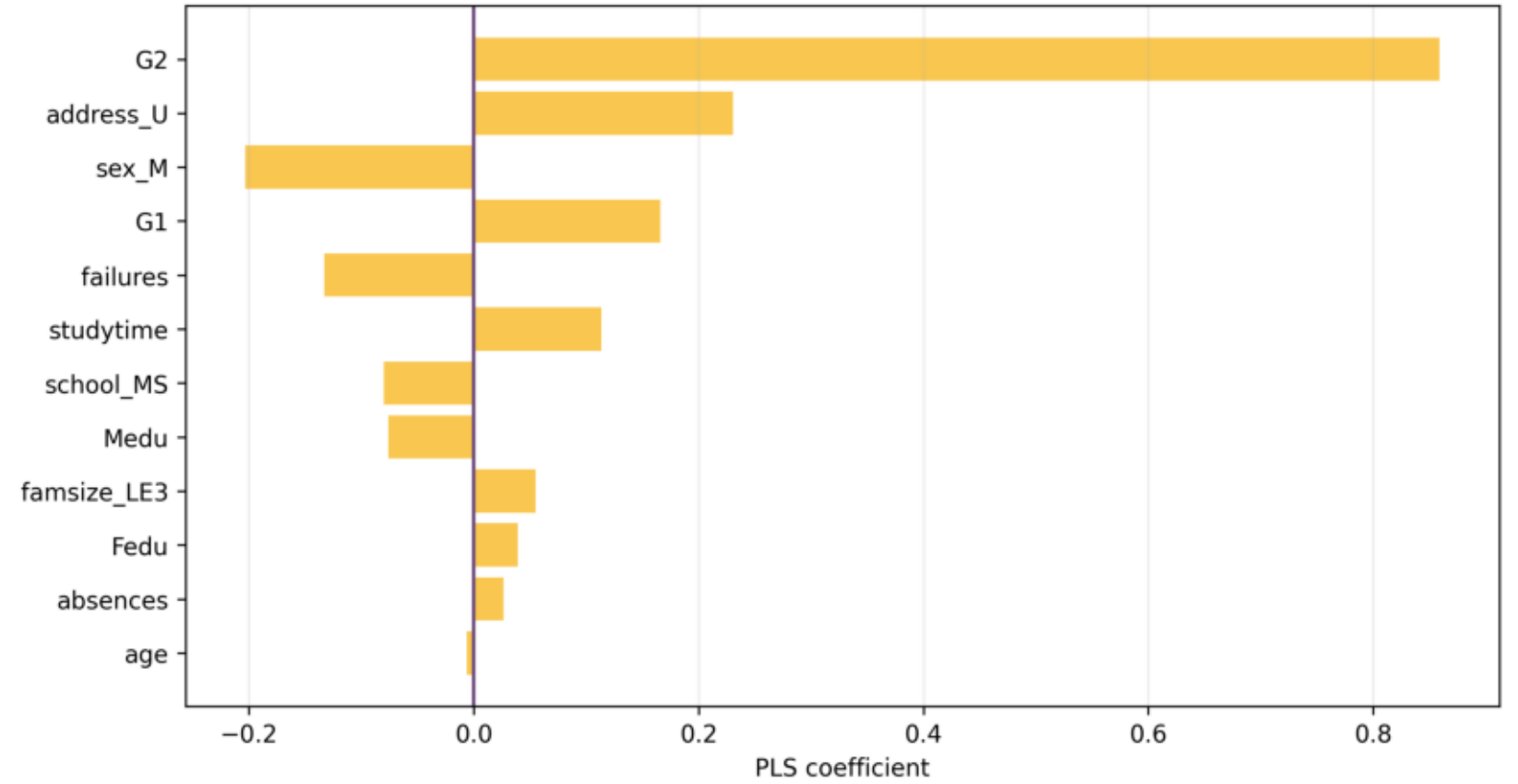
Component selection by CV RMSE

The selected component count is the value with the lowest cross-validation RMSE.



Largest PLS regression coefficients

Positive coefficients raise the predicted outcome; negative coefficients lower it.



Variable importance in projection

VIP values near or above 1 are commonly treated as influential predictors.

