

Principal Component Regression

Test name: Principal Component Regression

Dataset folder used for outputs: D:\DATA ANALYSIS\H Regression Tests and Models\Principal Component Regression

Outcome variable: G3

Rows used in PCR model: 649

Predictor columns after dummy coding: 12

Selected principal components: 8

Component selection rule: lowest cross-validated RMSE among available components

Explained variance retained by selected components: 0.8655

Training R-squared: 0.8122

Training adjusted R-squared: 0.8099

Training RMSE: 1.3989

Training MAE: 0.9058

Formula: $G3 \sim \text{principal components of } G1 + G2 + \text{studytime} + \text{failures} + \text{absences} + \text{age} + \text{Medu} + \text{Fedu} + \text{school_MS} + \text{sex_M} + \text{address_U} + \text{famsize_LE3}$

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

Categorical predictors considered: school, sex, address, famsize

Interpretation: PCA creates orthogonal components from standardized predictors; regression is then fitted on selected components.

Reporting guide: report selected components, retained variance, R-squared, RMSE, coefficients reconstructed to original scale, and residual diagnostics.

Charts generated: 10

Model fit summary

	test_name	outcome	n	predictor_columns_after_dummies	selected_components		selection_rule	variance_rule_suggested_components	explained_variance_selected_components	r_squared_training	adjusted_r_squared_training	rmse_training	mae_training	sse	mse	intercept	original_scale
Principal Component Regression	G3	649		12	8	lowest	cross-validated RMSE among available components	8	0.86548	0.812203	0.809856	1.398943	0.905775	1270.119833	1.984562		-3.020182

Component selection by cross-validation

components	cv_rmse	cv_mae	cv_r_squared	cumulative_explained_variance
1	2.073035	1.521733	0.587616	0.242768
2	1.882171	1.338983	0.660057	0.375921
3	1.825239	1.319489	0.680311	0.481555
4	1.800505	1.300932	0.688917	0.575192
5	1.748174	1.241825	0.706737	0.656790
6	1.613727	1.118430	0.750110	0.730465
7	1.421893	0.921224	0.805991	0.802781
8	1.415663	0.919743	0.807687	0.865480

Component explained variance

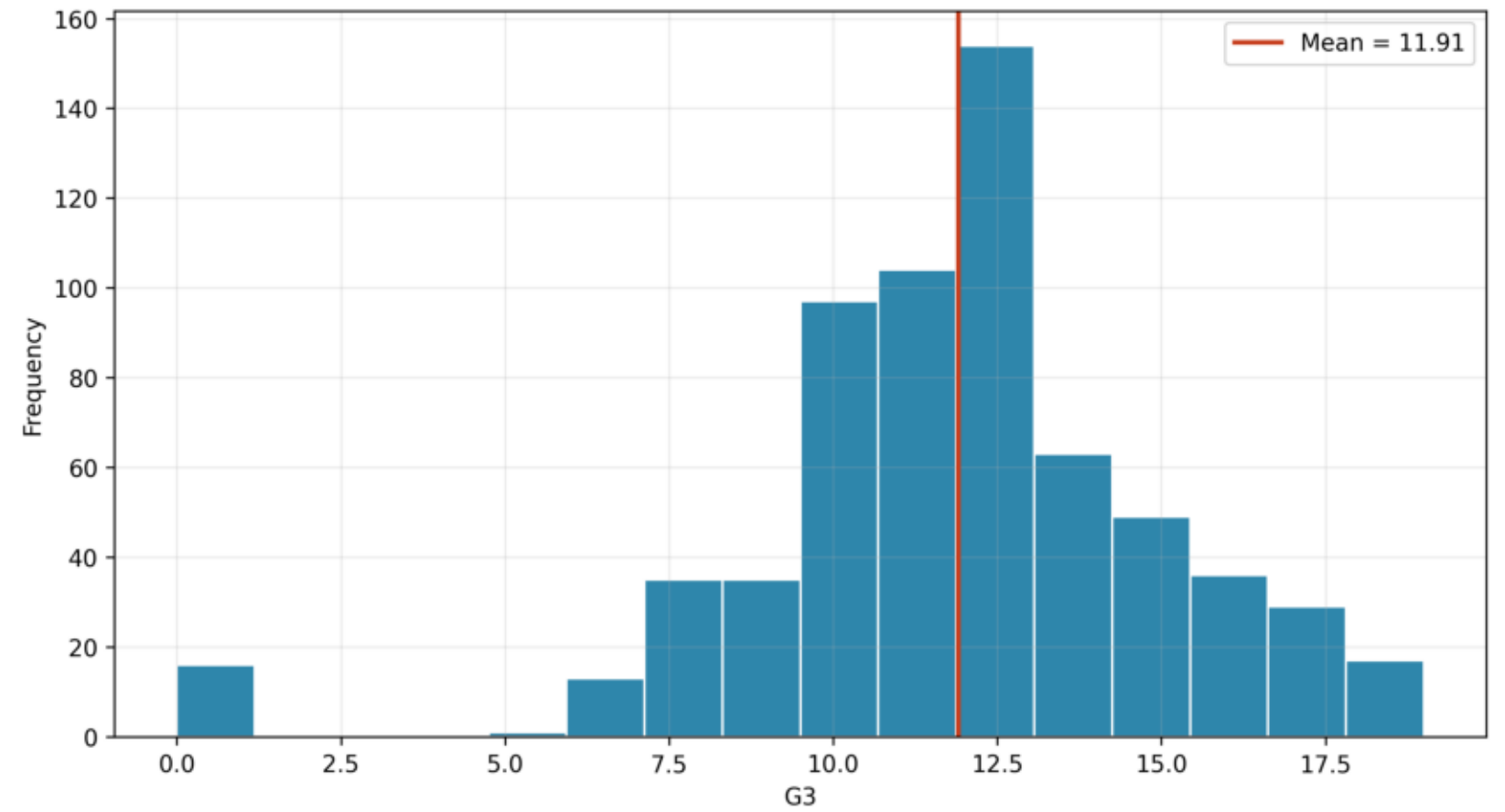
component	component_number	explained_variance_ratio	cumulative_explained_variance	selected_in_final_model
PC1	1	0.242768	0.242768	True
PC2	2	0.133153	0.375921	True
PC3	3	0.105634	0.481555	True
PC4	4	0.093637	0.575192	True
PC5	5	0.081598	0.656790	True
PC6	6	0.073675	0.730465	True
PC7	7	0.072316	0.802781	True
PC8	8	0.062699	0.865480	True

Original-scale coefficient summary

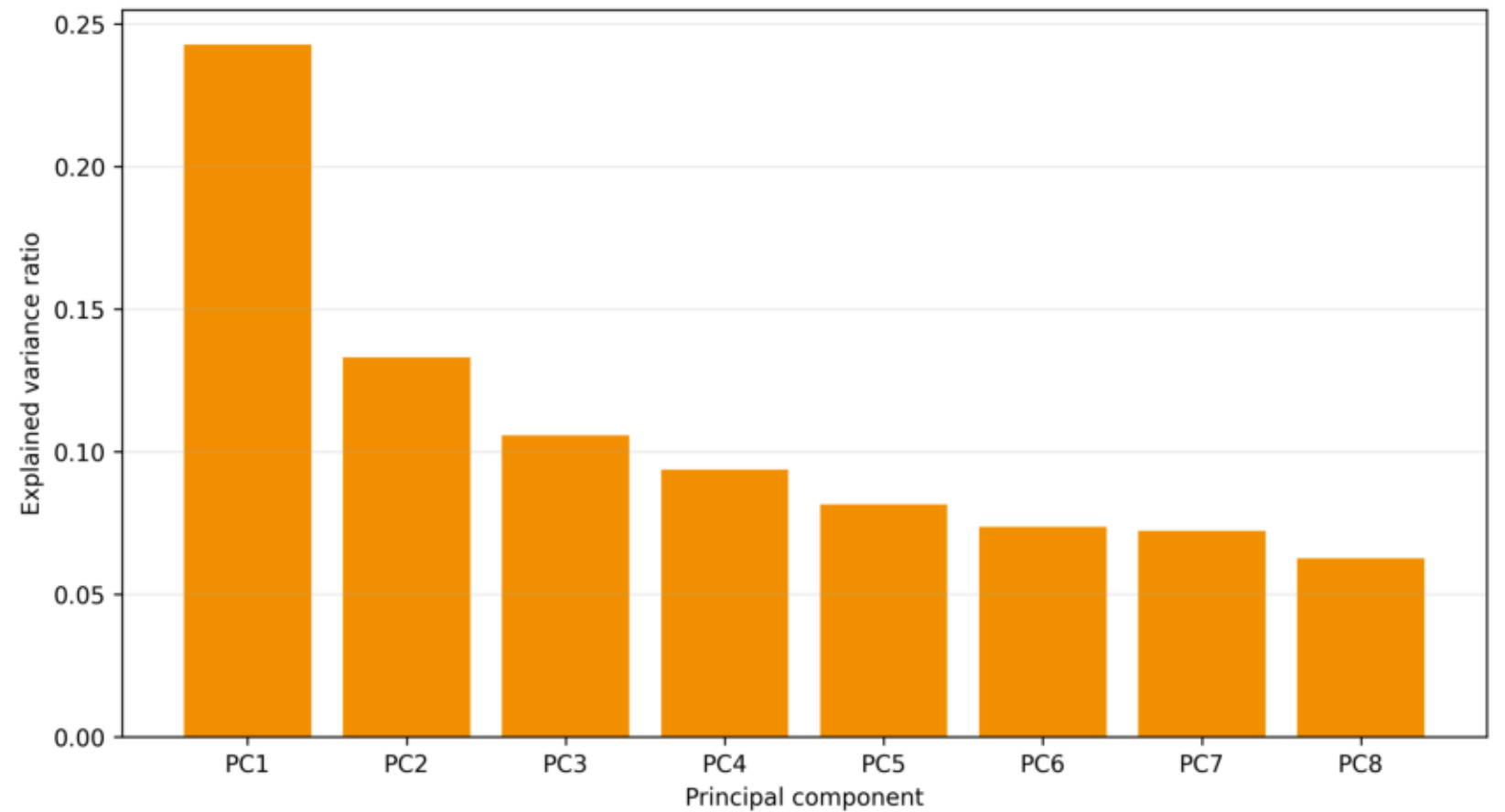
term	original_scale_coefficient	standardized_predictor_coefficient	interpretation						
Intercept	-3.020182	NaN	Predicted outcome when all original predictors are zero						
G1	0.494354	1.356088	Approximate	PCR	coefficient	reconstructed	from	selected	principal components
G2	0.484717	1.411202	Approximate	PCR	coefficient	reconstructed	from	selected	principal components
studytime	-0.069432	-0.057550	Approximate	PCR	coefficient	reconstructed	from	selected	principal components
failures	-0.707365	-0.419310	Approximate	PCR	coefficient	reconstructed	from	selected	principal components
absences	-0.005446	-0.025253	Approximate	PCR	coefficient	reconstructed	from	selected	principal components
age	0.262102	0.319030	Approximate	PCR	coefficient	reconstructed	from	selected	principal components
Medu	0.015300	0.017345	Approximate	PCR	coefficient	reconstructed	from	selected	principal components
Fedu	-0.028179	-0.030971	Approximate	PCR	coefficient	reconstructed	from	selected	principal components
school_MS	-0.573307	-0.273128	Approximate	PCR	coefficient	reconstructed	from	selected	principal components
sex_M	-0.275694	-0.135589	Approximate	PCR	coefficient	reconstructed	from	selected	principal components
address_U	-0.095369	-0.043850	Approximate	PCR	coefficient	reconstructed	from	selected	principal components
famsize_LE3	0.035845	0.016360	Approximate	PCR	coefficient	reconstructed	from	selected	principal components

Outcome distribution

Distribution of the outcome variable before principal component regression.

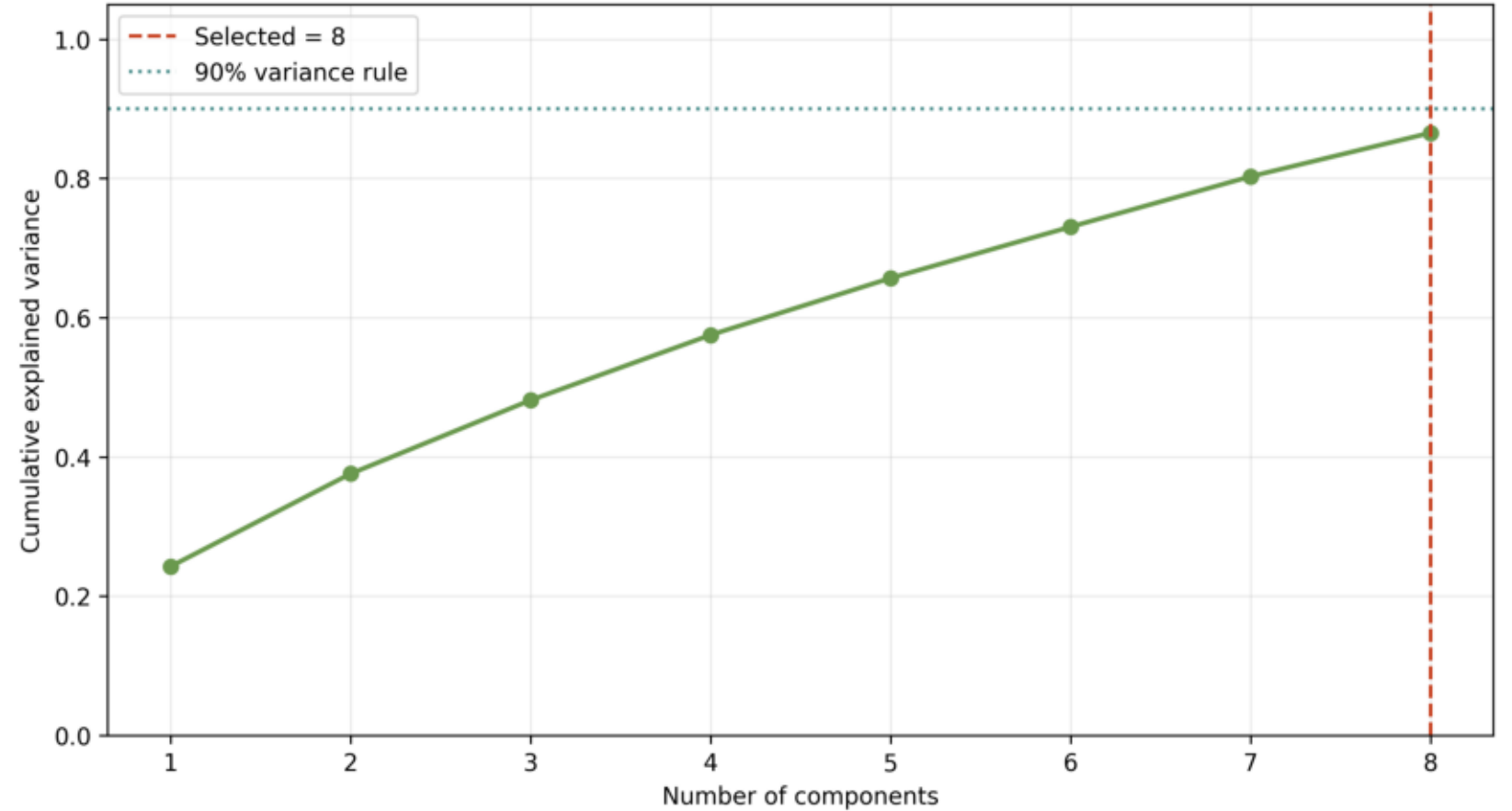


PCR scree chart
Explained variance contributed by each principal component.



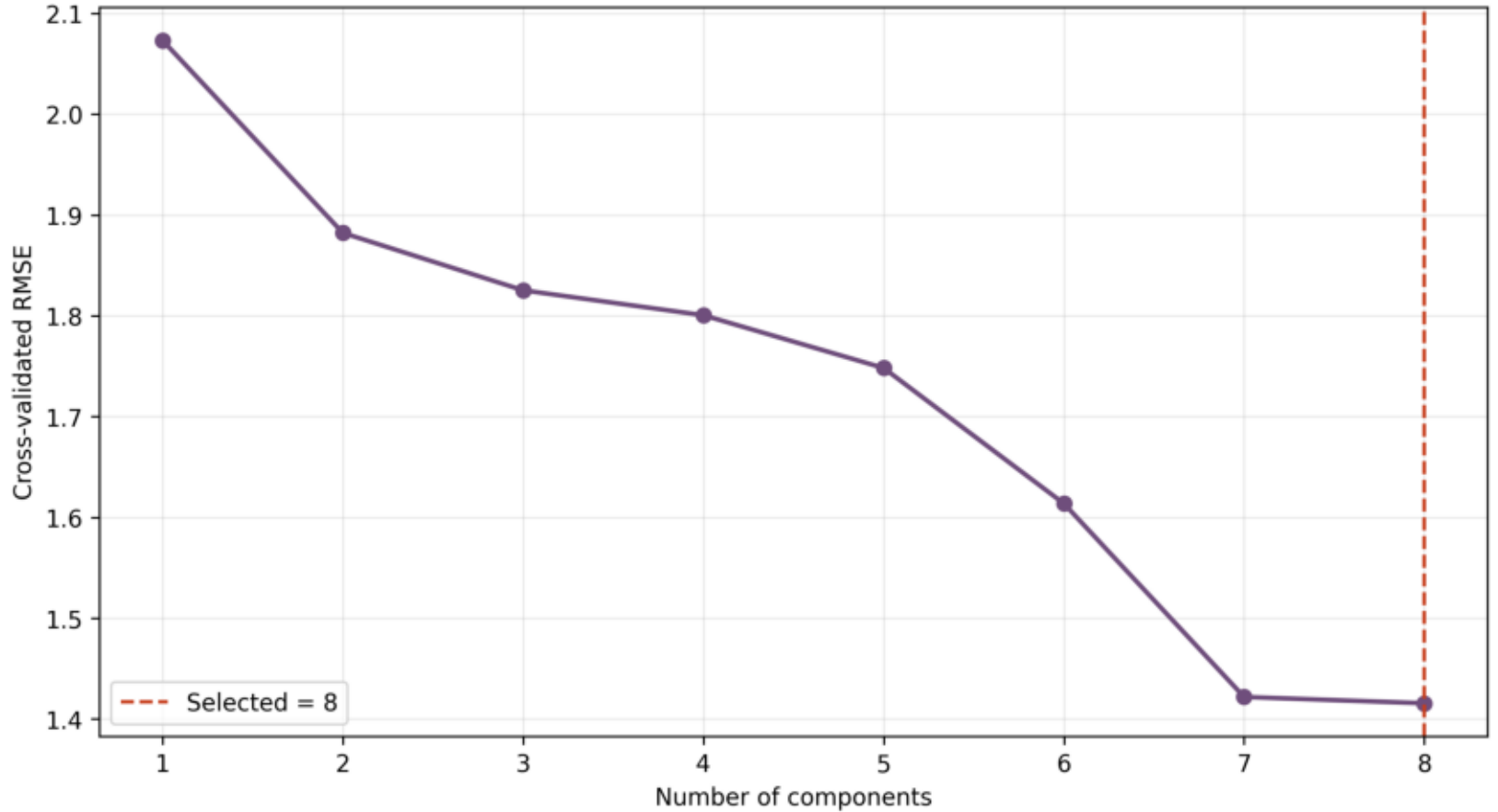
Cumulative explained variance

How much predictor information is retained as components are added.



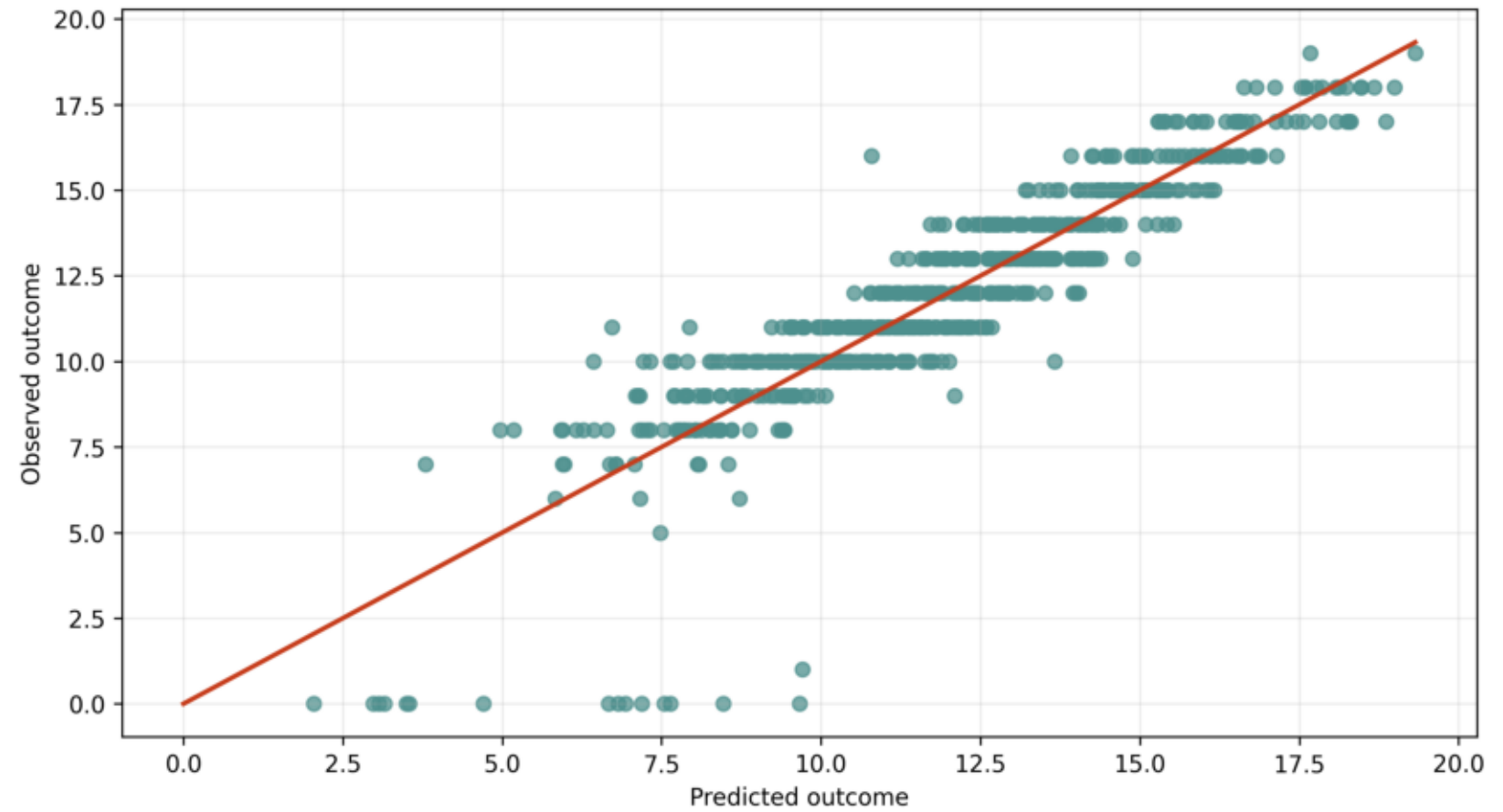
Component selection by CV RMSE

Lower cross-validated RMSE indicates better predictive component count.



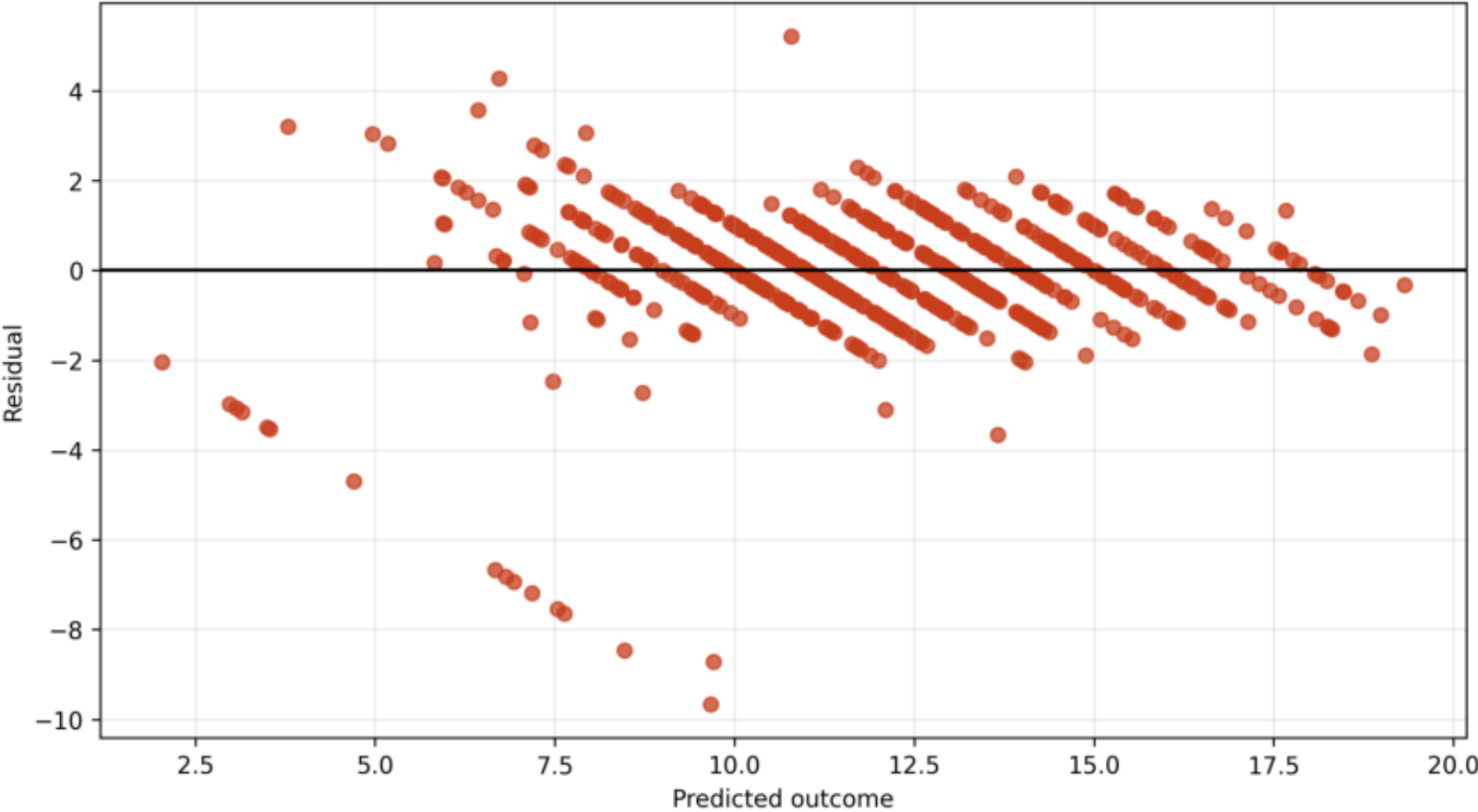
Observed vs predicted values

Points near the diagonal line indicate stronger PCR prediction accuracy.



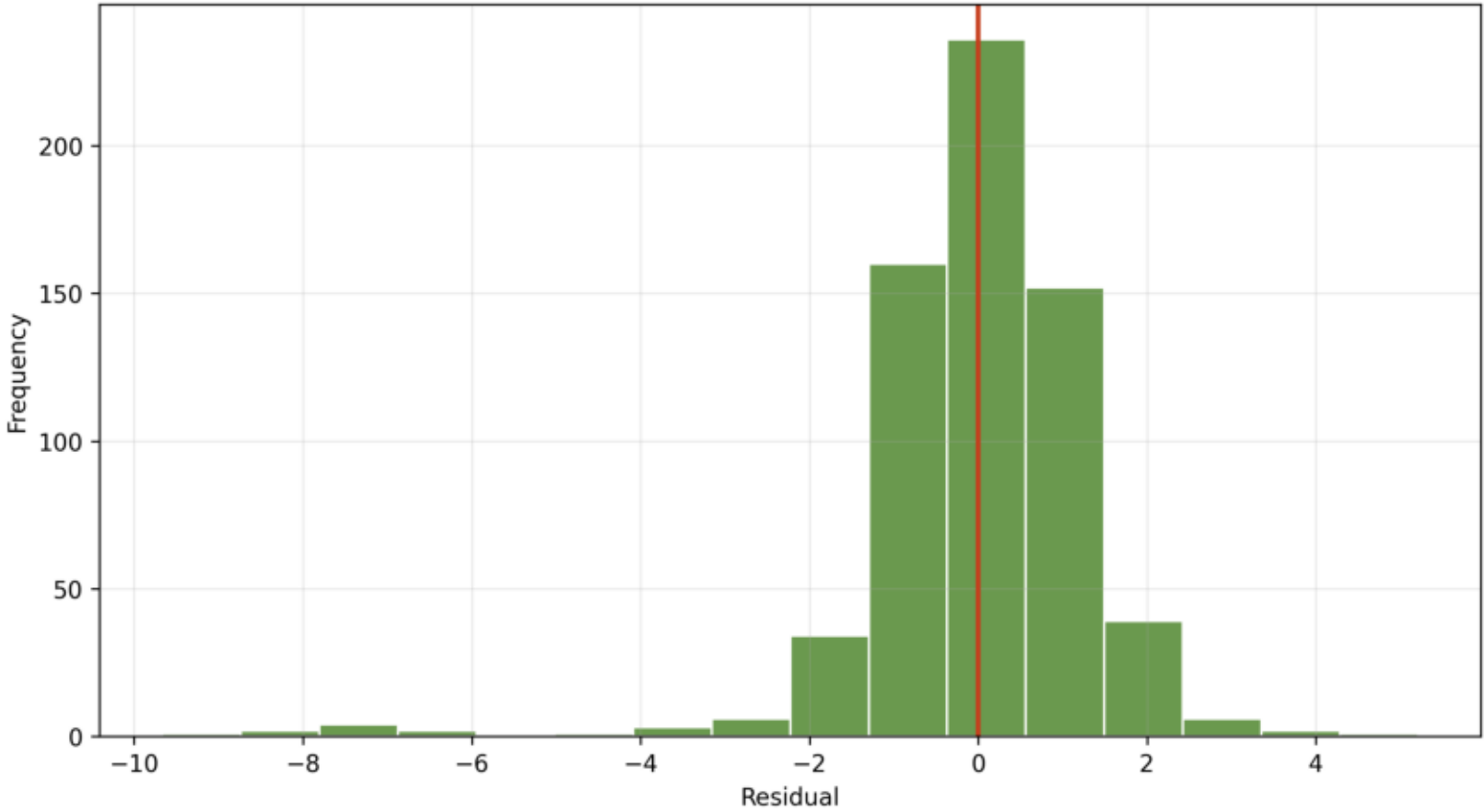
Residuals vs fitted values

Residuals should be centered around zero without a strong pattern.



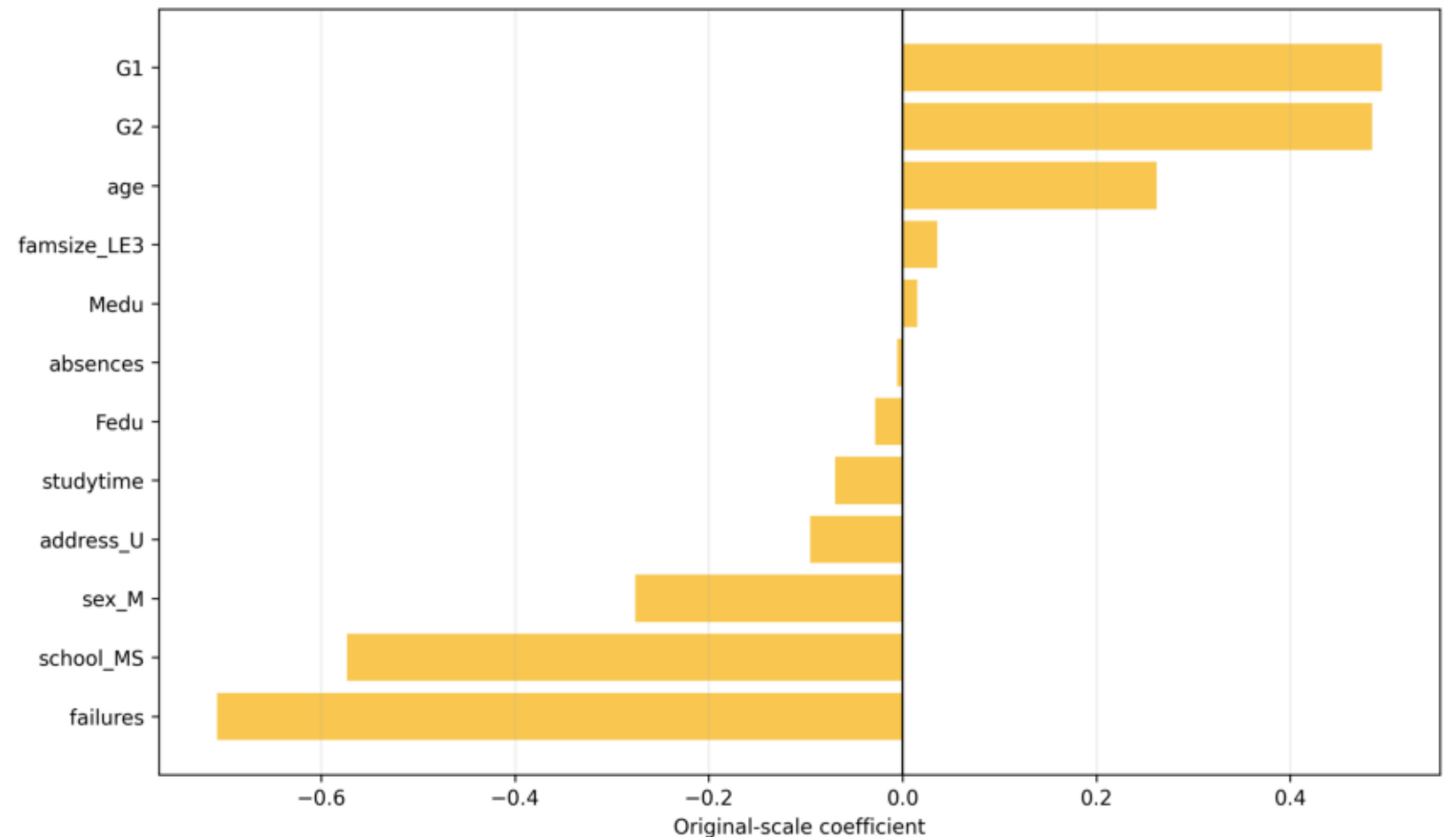
Residual distribution

The residual histogram checks the error distribution after PCR fitting.



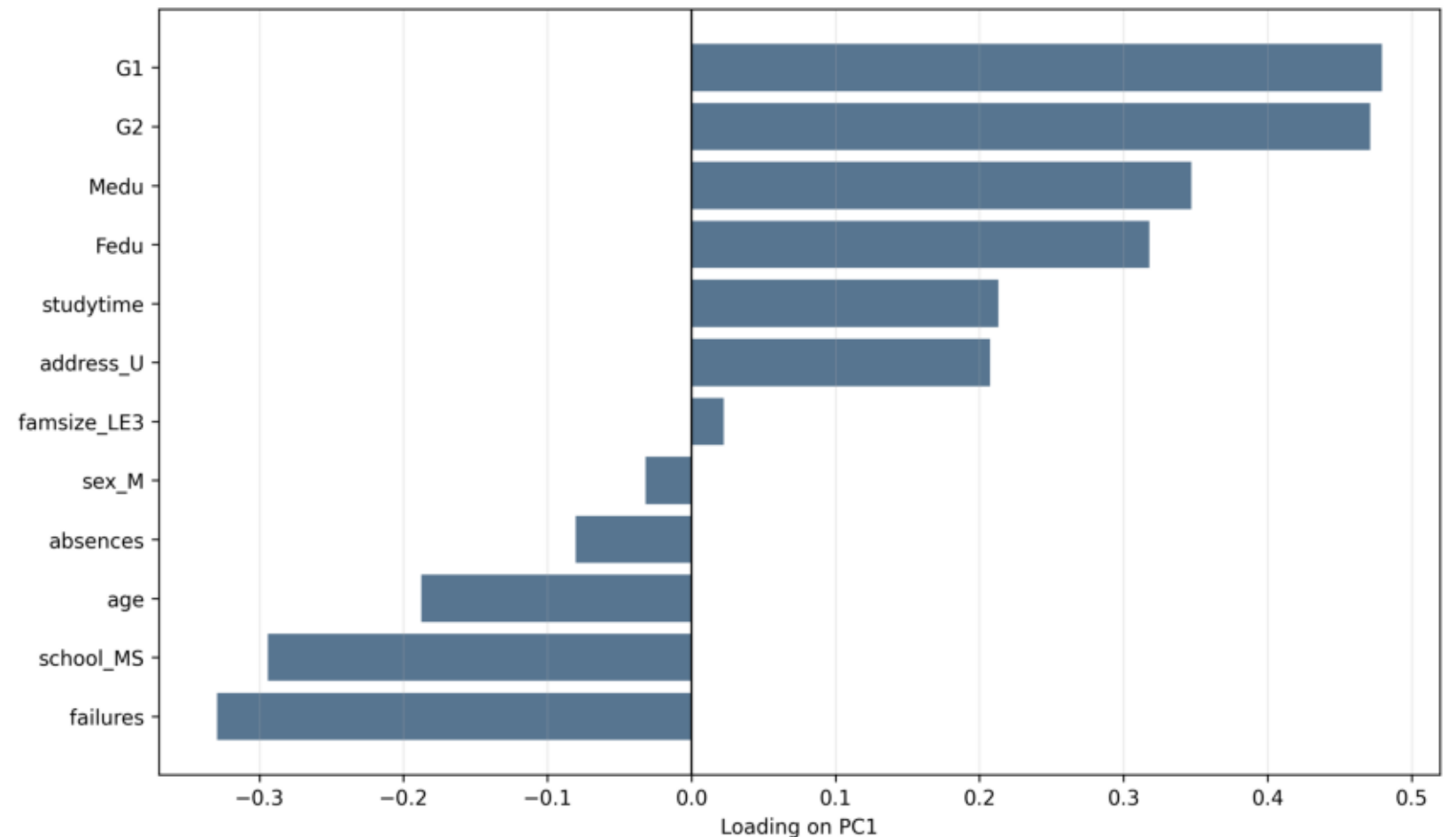
Largest PCR reconstructed coefficients

Original-scale coefficients reconstructed from selected principal components.



Dominant loadings on first component

Predictors with larger absolute loadings contribute more to the first component.



PC1 and PC2 score map

Observations positioned by the first two principal component scores.

