

Multiple Linear Regression

Test name: Multiple Linear Regression

Dataset: D:\DATA ANALYSIS\H Regression Tests and Models\Multiple Linear Regression\dataset.csv

Rows in original dataset: 649

Columns in original dataset: 33

Rows used in multiple linear regression model: 649

Outcome variable: G3

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

Categorical predictors: school, sex, address

Formula: $G3 \sim G1 + G2 + studytime + failures + absences + age + C(school) + C(sex) + C(address)$

R-squared: 0.8522

Adjusted R-squared: 0.8501

F statistic: 409.4514

Model p-value: 0.000000

RMSE: 1.2410

MAE: 0.7799

Residual Shapiro-Wilk p-value: 0.000000

Interpretation: each coefficient estimates the unique expected outcome change when that predictor increases by one unit while other predictors are held constant.

Decision guide: $p < 0.05$ indicates statistically significant evidence for a predictor effect.

Model fit guide: adjusted R-squared is preferred when comparing models with different numbers of predictors.

Assumption guide: inspect residual-vs-fitted, residual histogram, Q-Q plot, and VIF diagnostics.

Charts generated: 8

Model fit statistics

model	outcome	n	df_model	df_residual	r_squared	adjusted_r_s	f_statistic	f_p_value	aic	bic	rmse	mae
Multiple line	G3	649.0000	9.0000	639.0000	0.8522	0.8501	409.4514	0.0000	2142.0090	2186.7633	1.2410	0.7799

ANOVA table

source	sum_of_squares	df	mean_square	f_value	p_value
Regression	5763.8066	9.0000	640.4230	409.4514	0.0000
Residual	999.4600	639.0000	1.5641		
Total	6763.2666	648.0000			

Regression coefficients

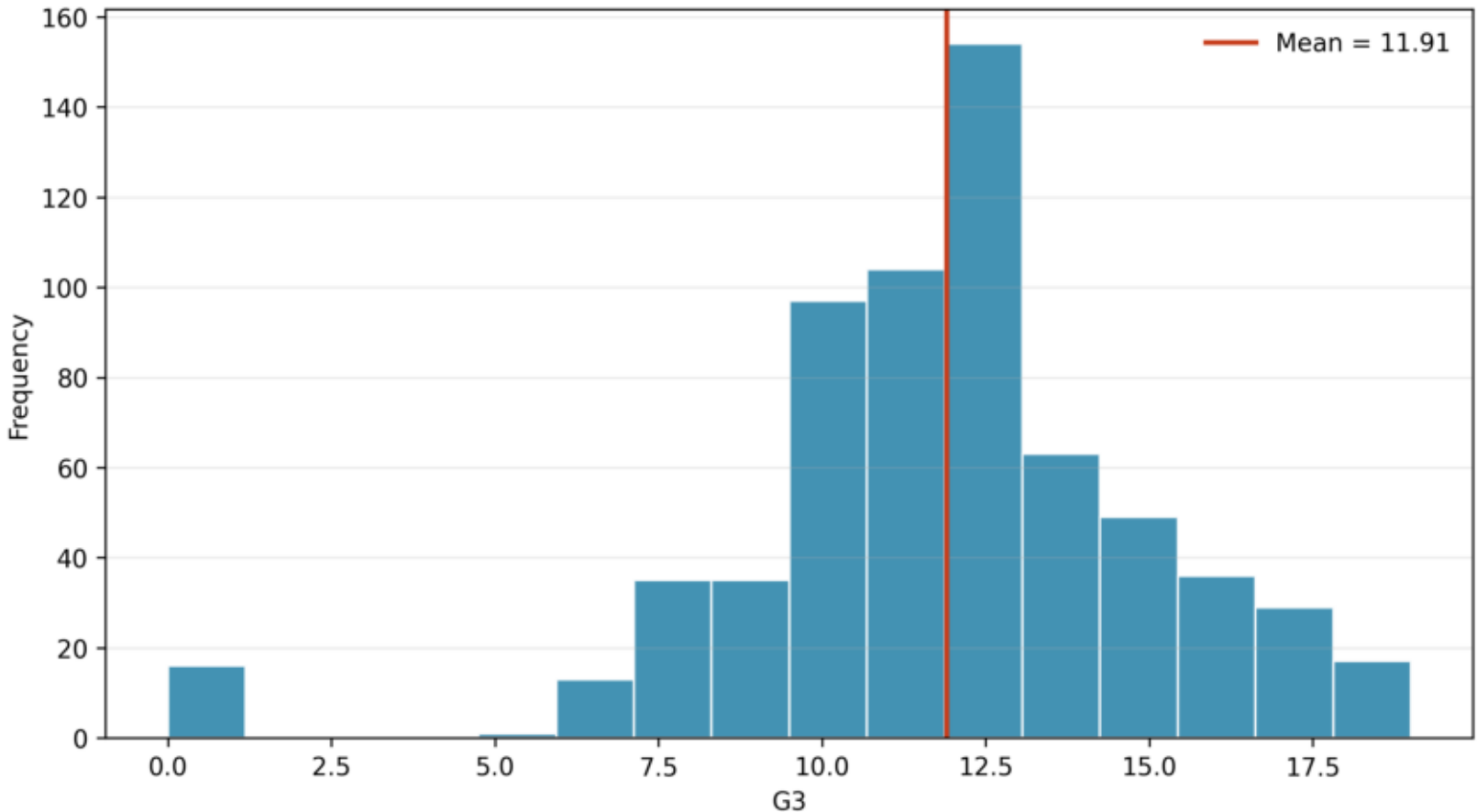
term	coefficient_b	standard_error	t_value	p_value	ci95_lower	ci95_upper	decision_alpha_0_05
Intercept	-0.2405	0.7779	-0.3091	0.7573	-1.7679	1.2870	Not statistically significant
C(school)[T.MS]	-0.1665	0.1181	-1.4099	0.1590	-0.3983	0.0654	Not statistically significant
C(sex)[T.M]	-0.2037	0.1034	-1.9701	0.0493	-0.4068	-0.0007	Statistically significant
C(address)[T.U]	0.0989	0.1145	0.8636	0.3881	-0.1260	0.3237	Not statistically significant
G1	0.1328	0.0368	3.6121	0.0003	0.0606	0.2050	Statistically significant
G2	0.8794	0.0342	25.7517	0.0000	0.8124	0.9465	Statistically significant
studytime	0.0619	0.0632	0.9783	0.3283	-0.0623	0.1860	Not statistically significant
failures	-0.2266	0.0947	-2.3922	0.0170	-0.4127	-0.0406	Statistically significant
absences	0.0177	0.0112	1.5856	0.1133	-0.0042	0.0396	Not statistically significant
age	0.0236	0.0435	0.5432	0.5872	-0.0618	0.1091	Not statistically significant

VIF diagnostics

term	vif	tolerance	diagnostic
C(school)[T.MS]	1.3128	0.7617	Acceptable
C(sex)[T.M]	1.0733	0.9317	Acceptable
C(address)[T.U]	1.1499	0.8696	Acceptable
G1	4.2214	0.2369	Acceptable
G2	4.1019	0.2438	Acceptable
studytime	1.1399	0.8773	Acceptable
failures	1.3087	0.7641	Acceptable
absences	1.1113	0.8999	Acceptable
age	1.1653	0.8582	Acceptable

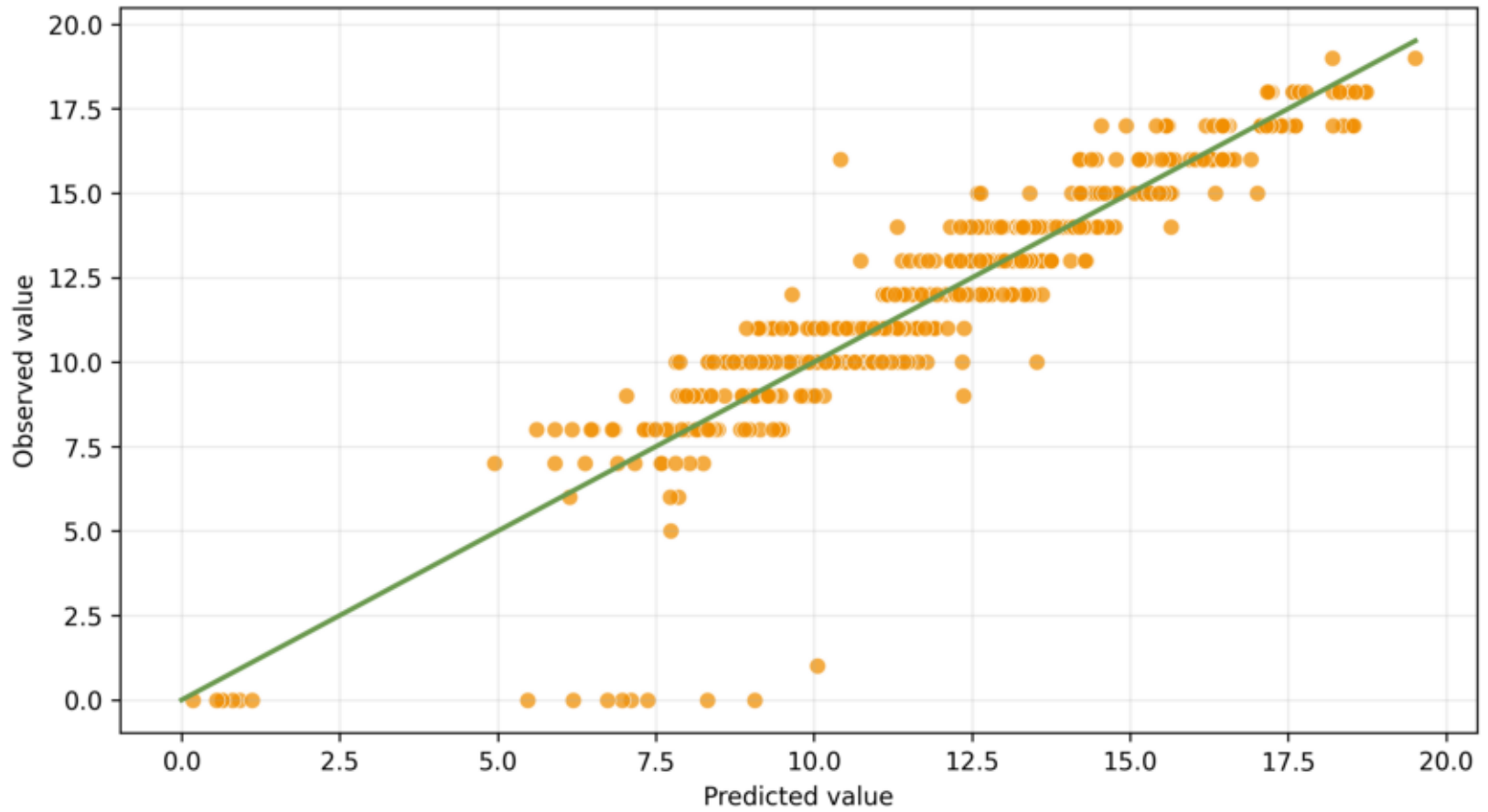
Outcome distribution

Distribution of the dependent variable used in the multiple linear regression model.



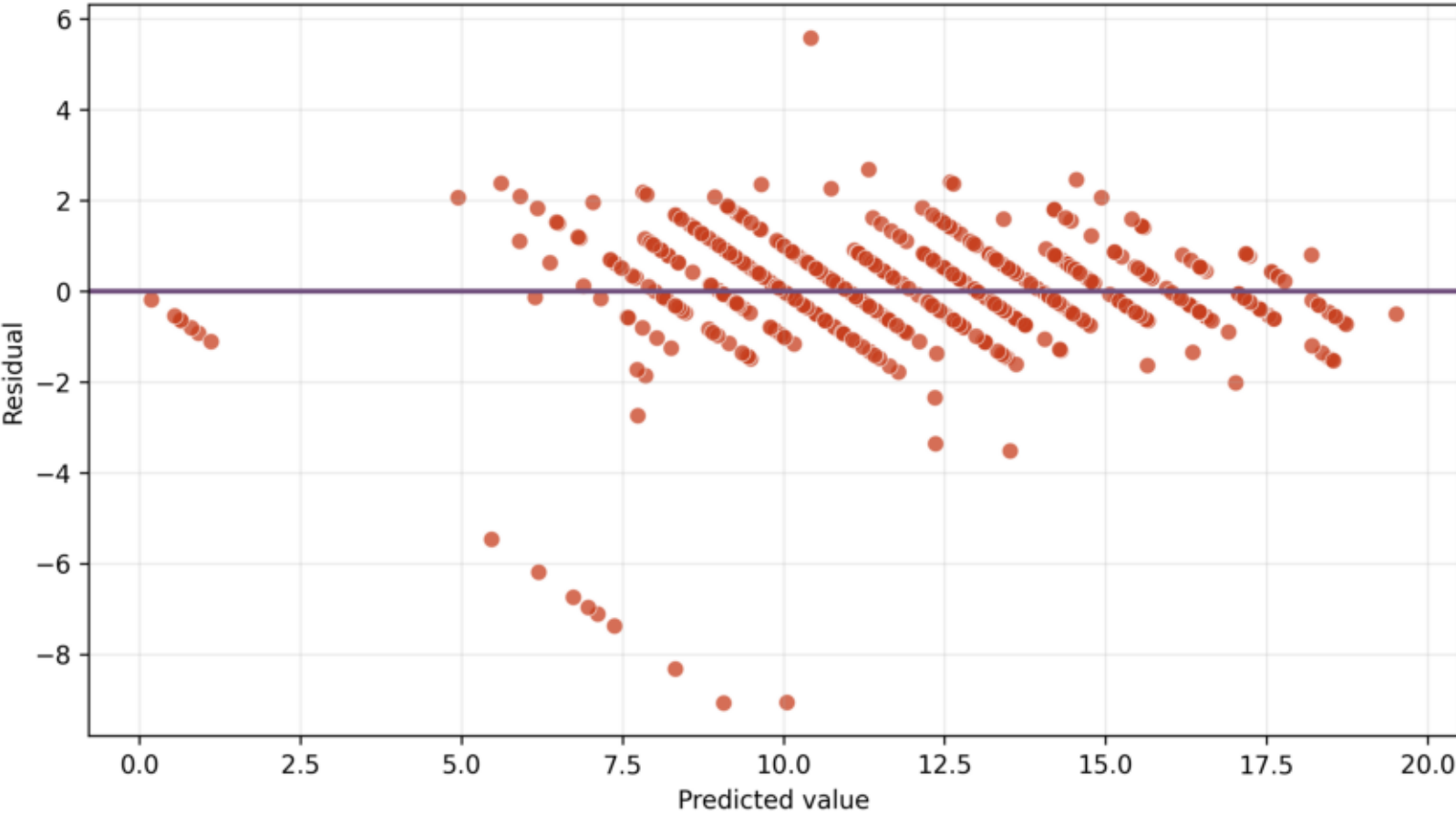
Observed vs predicted values

Points close to the diagonal indicate stronger prediction accuracy.



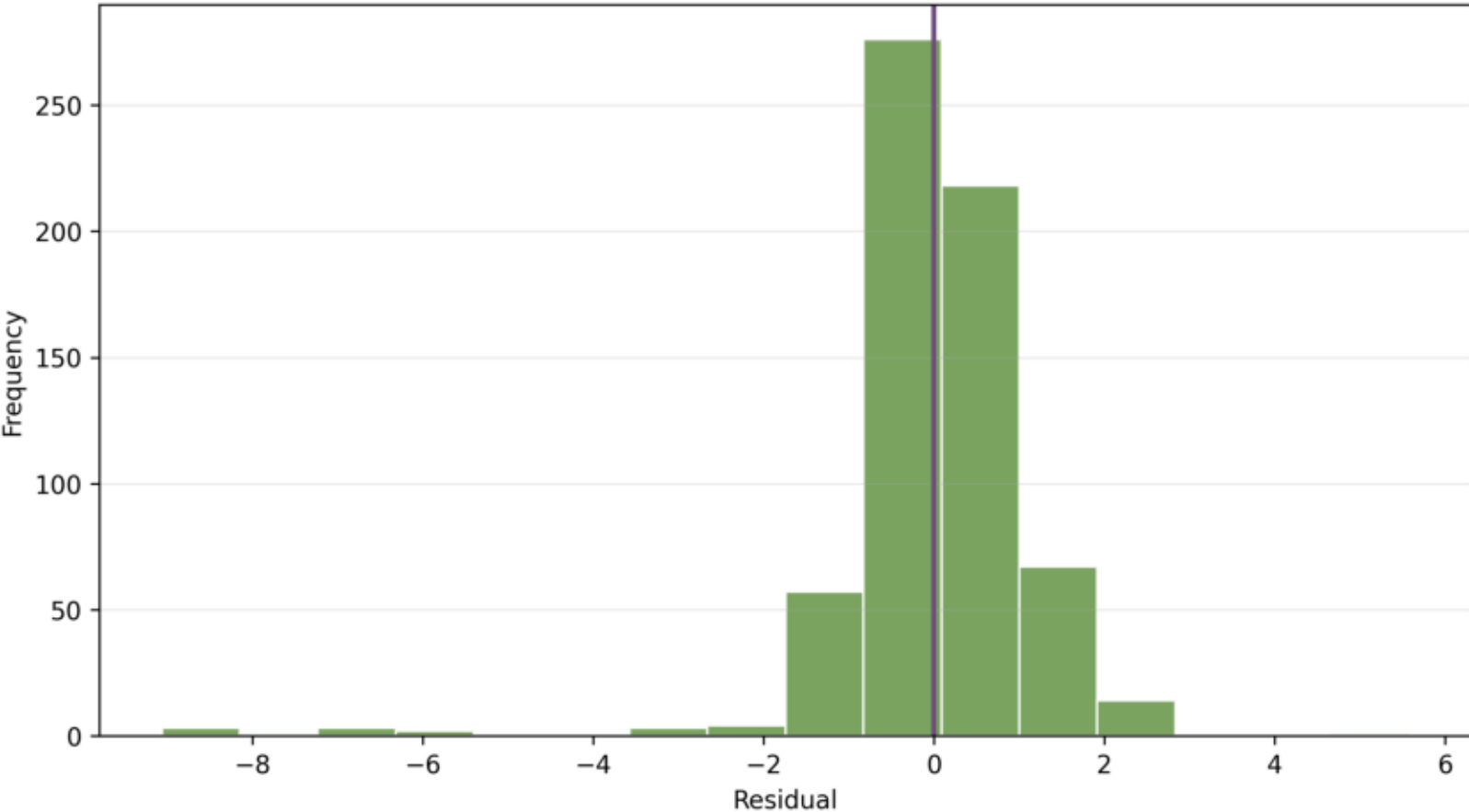
Residuals vs predicted values

A random cloud around zero supports the linearity and equal-variance assumptions.



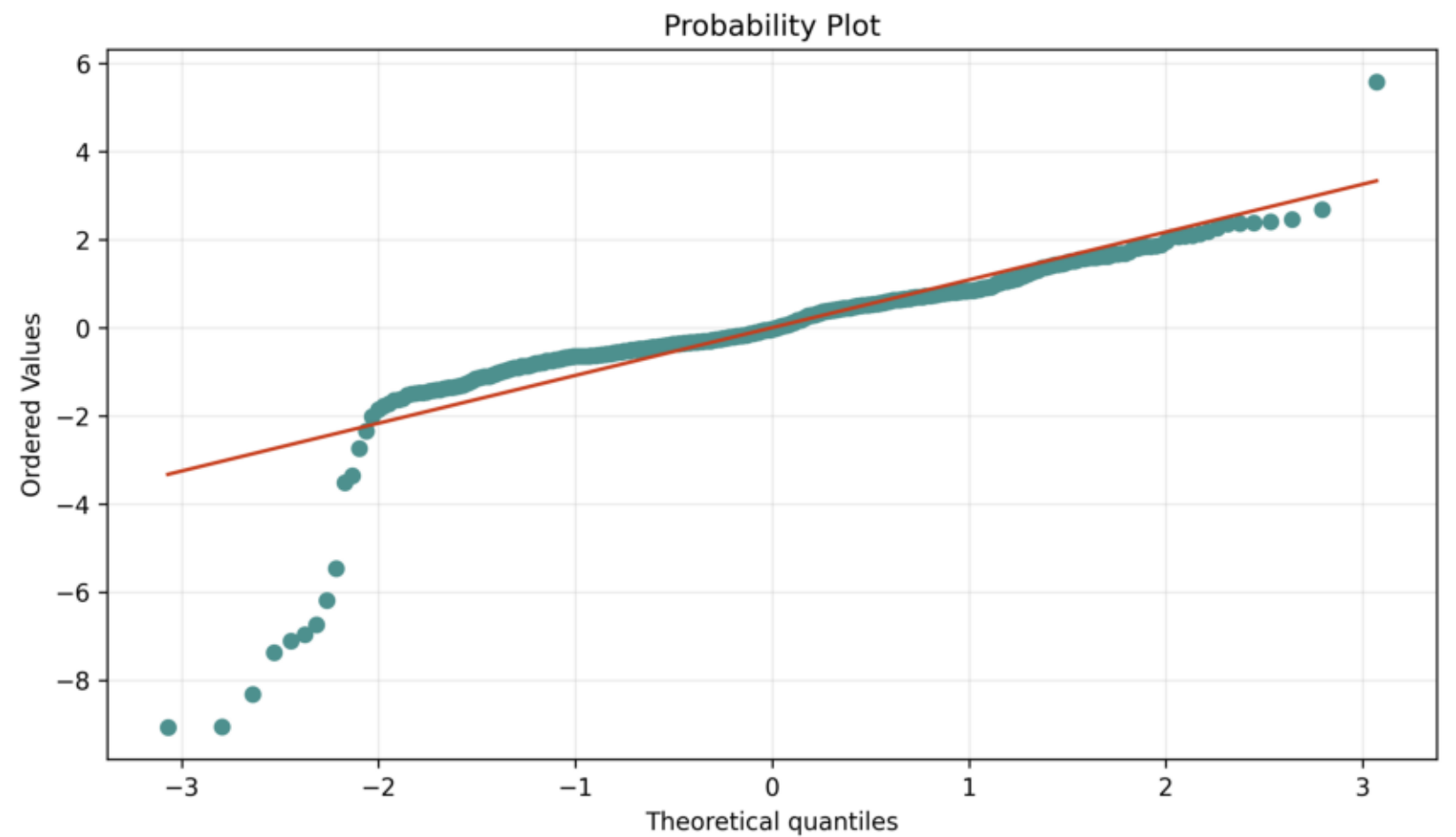
Residual distribution

Residuals centered around zero support a well-calibrated regression model.



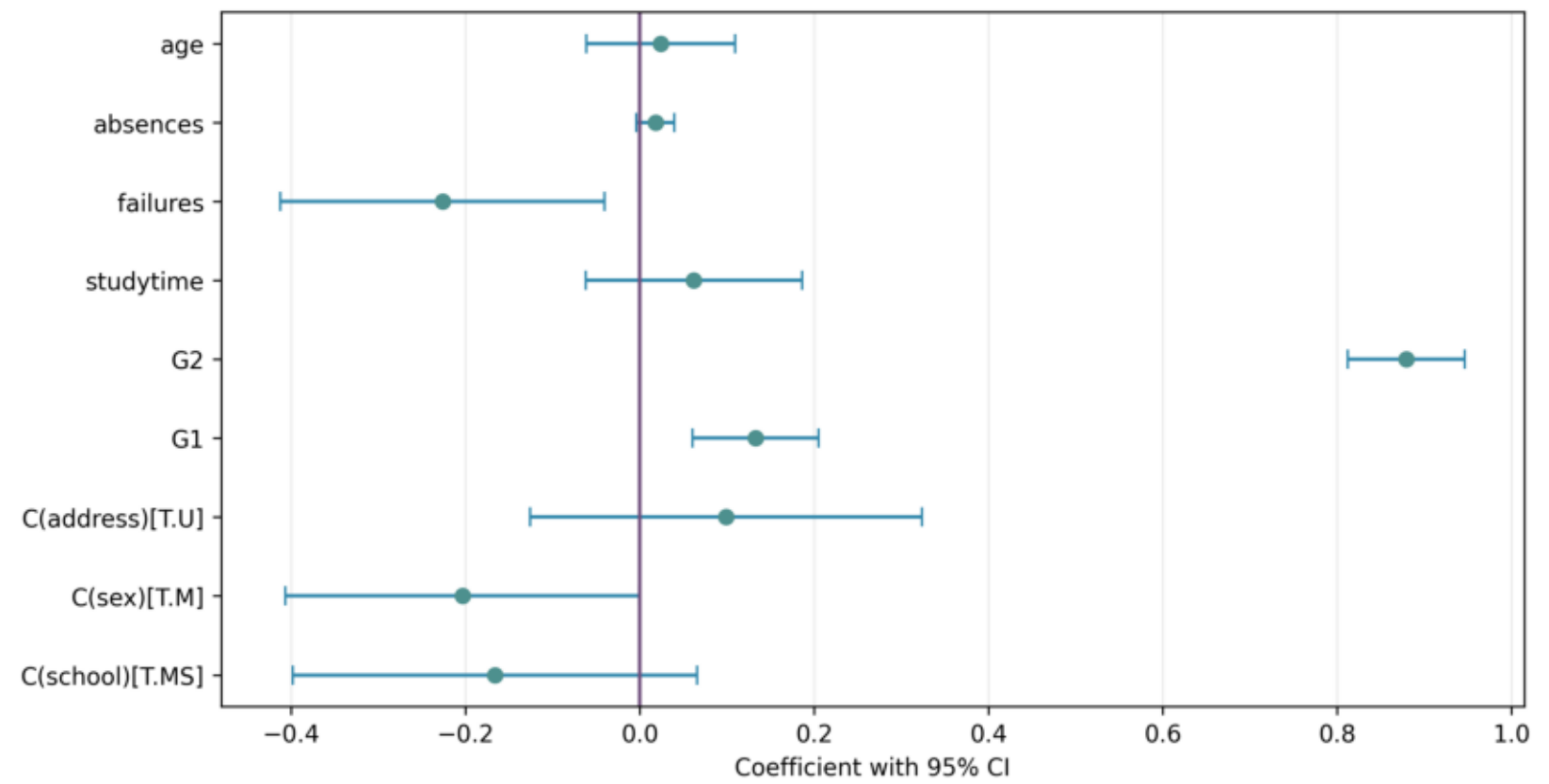
Normal Q-Q plot of residuals

Residual points near the line support the normality assumption.



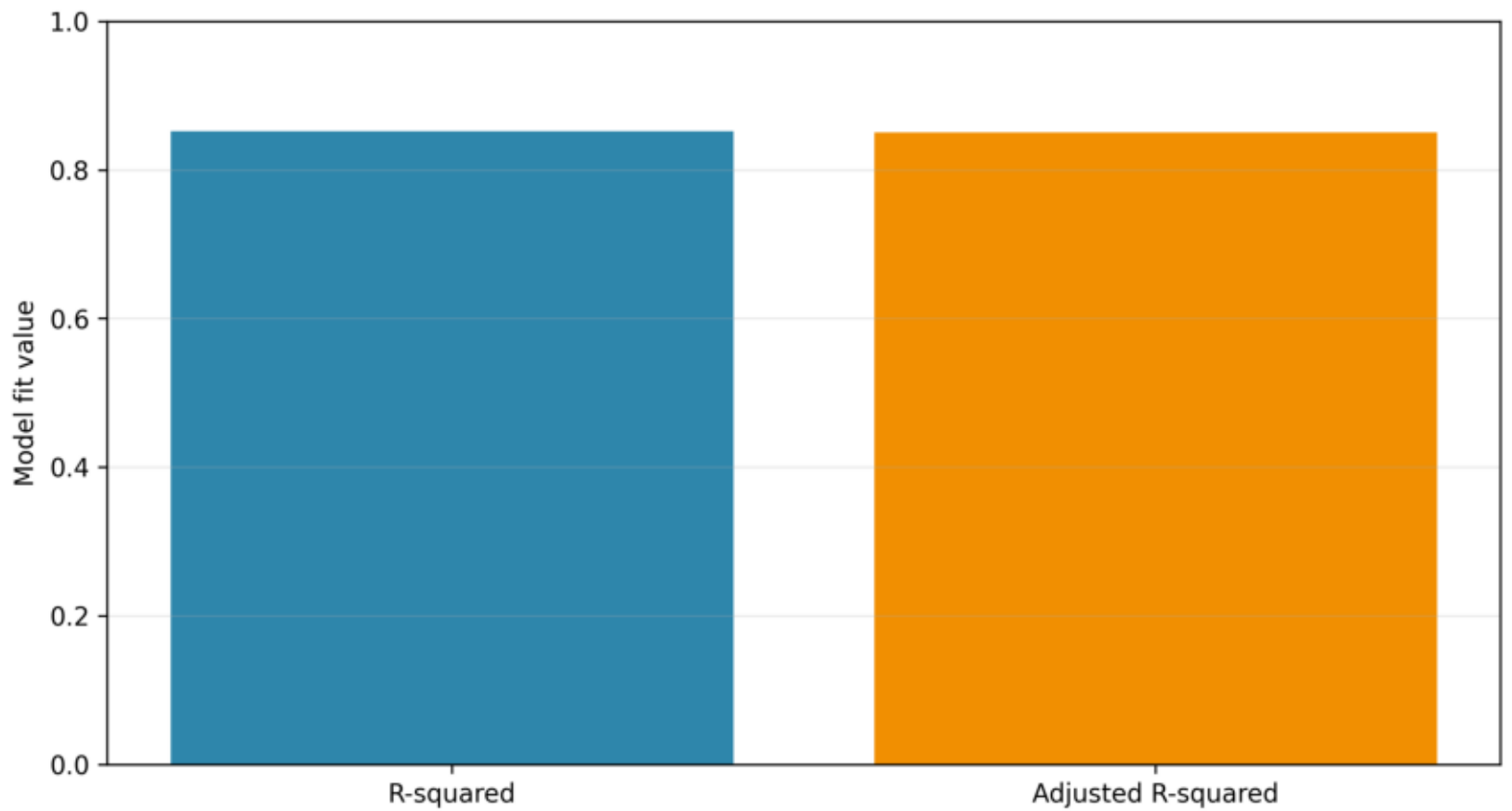
Regression coefficient plot

Confidence intervals not crossing zero indicate stronger predictor evidence.



Model fit summary

Adjusted R-squared penalizes unnecessary predictors and is central to reporting multiple regression.



Observed and predicted outcome by G1

This chart compares fitted regression values with observed outcomes across a key predictor.

