

Probit Regression

Test name: Probit Regression

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

Rows in original dataset: 649

Columns in original dataset: 33

Rows used in Probit Regression model: 649

Binary outcome created: probit_pass = 1 when G3 >= 10, otherwise 0

Observed event rate: 0.8459

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

Categorical predictors: school, sex, address

Formula: $\text{probit_pass} \sim \text{G1} + \text{G2} + \text{studytime} + \text{failures} + \text{absences} + \text{age} + \text{C(school)} + \text{C(sex)} + \text{C(address)}$

Family/link: binomial(link = probit).

Log likelihood: -89.2387

AIC: 198.4774

McFadden pseudo R squared: 0.6800

Classification accuracy at 0.50 cutoff: 0.9384

Interpretation: coefficients are on the latent standard normal z-score scale.

Positive coefficients increase the linear probit score and therefore increase predicted probability.

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

Reporting guide: state the binary outcome cutoff, probit link, coefficients, p-values, model fit, and classification results.

Charts generated: 8

Probit Regression coefficients

	term	probit_coefficient	standard_error	z_value	p_value	ci95_lower	ci95_upper	decision_alpha_0_05	interpretation
	Intercept	-12.814190	2.124178	-6.032541	1.614011e-09	-16.977502	-8.650878	Statistically significant	Positive values increase the latent probit score and therefore increase the predicted probability.
	C(school)[T.MS]	-0.588822	0.237304	-2.481293	1.309067e-02	-1.053930	-0.123714	Statistically significant	Positive values increase the latent probit score and therefore increase the predicted probability.
	C(sex)[T.M]	-0.312284	0.226364	-1.379567	1.677199e-01	-0.755950	0.131381	Not statistically significant	Positive values increase the latent probit score and therefore increase the predicted probability.
	C(address)[T.U]	-0.112358	0.225203	-0.498920	6.178360e-01	-0.553748	0.329032	Not statistically significant	Positive values increase the latent probit score and therefore increase the predicted probability.
	G1	0.331065	0.076525	4.326227	1.516850e-05	0.181079	0.481052	Statistically significant	Positive values increase the latent probit score and therefore increase the predicted probability.
	G2	0.800302	0.113398	7.057477	1.695528e-12	0.578046	1.022557	Statistically significant	Positive values increase the latent probit score and therefore increase the predicted probability.
	studytime	0.089479	0.147093	0.608319	5.429760e-01	-0.198817	0.377775	Not statistically significant	Positive values increase the latent probit score and therefore increase the predicted probability.
	failures	-0.185865	0.138095	-1.345923	1.783273e-01	-0.456525	0.084796	Not statistically significant	Positive values increase the latent probit score and therefore increase the predicted probability.
	absences	-0.024443	0.023140	-1.056308	2.908277e-01	-0.069797	0.020911	Not statistically significant	Positive values increase the latent probit score and therefore increase the predicted probability.
	age	0.209143	0.094820	2.205684	2.740615e-02	0.023299	0.394986	Statistically significant	Positive values increase the latent probit score and therefore increase the predicted probability.

Probit Regression model fit statistics

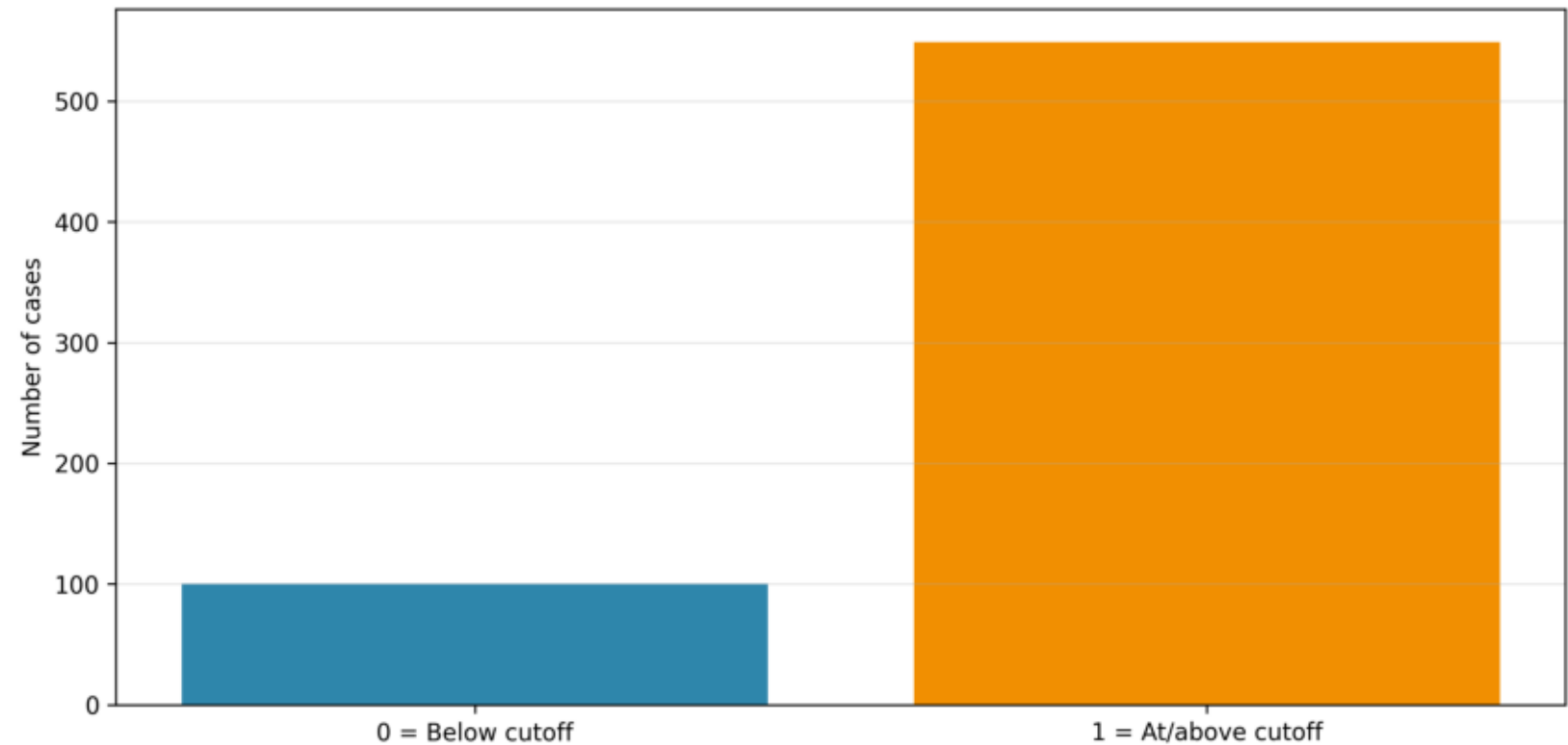
model	family	link	n	df_model	df_residual	log_likelihood	null_log_likelihood	aic	bic_deviance	mcFadden_pseudo_r2	classification_accuracy_cutoff_0_50	mean_predicted_probability	observed_event_rate	log_loss
Probit Regression	Binomial	Probit	649	9.0	639.0	-89.238714	-278.89277	198.477428	-3959.324078	0.680025	0.938367	0.84507	0.845917	0.137502

Probit Regression confusion matrix

actual_class	predicted_0	predicted_1
Actual 0	77	23
Actual 1	17	532

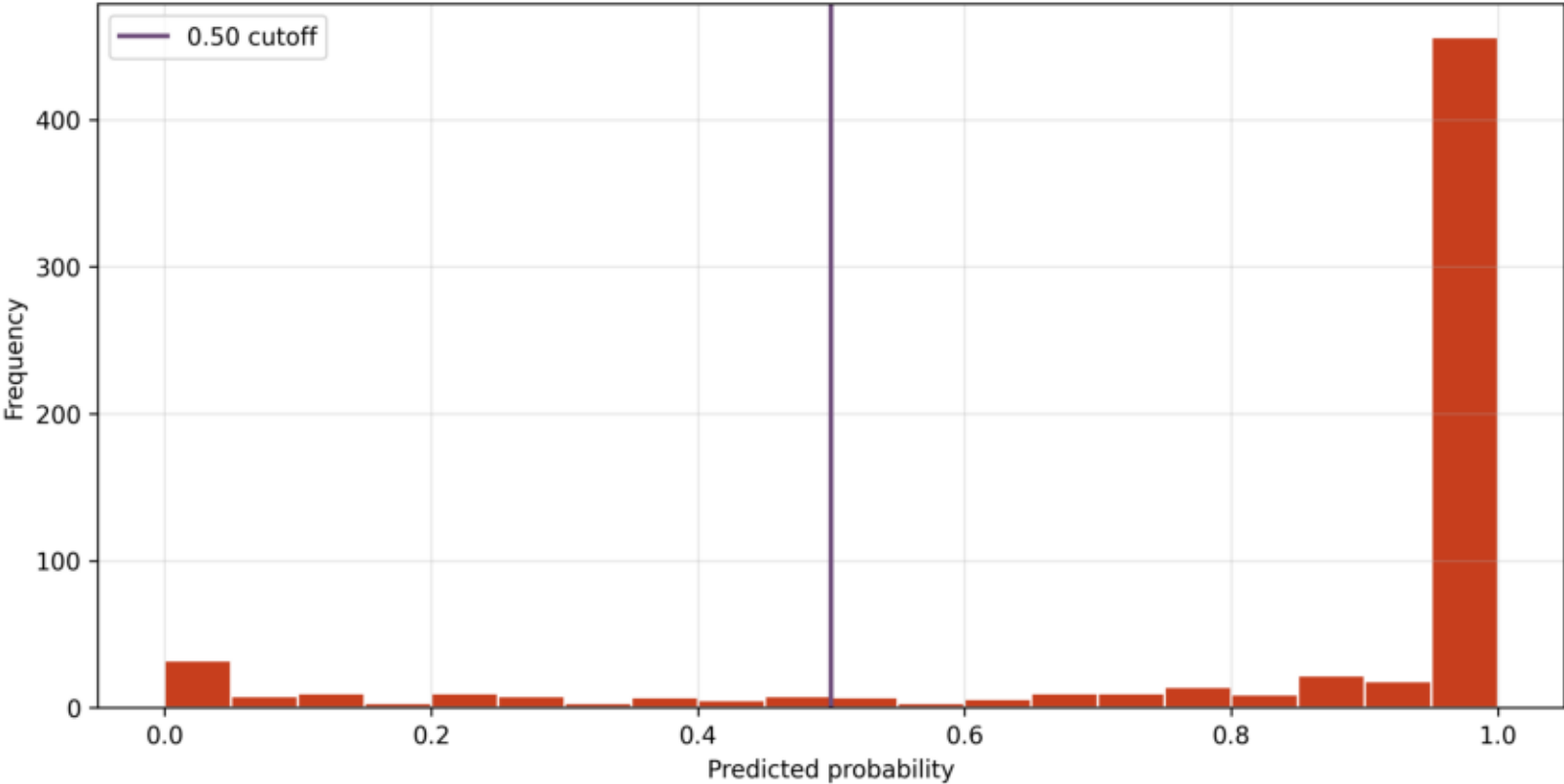
Binary outcome distribution

Outcome created from G3 using cutoff 10.



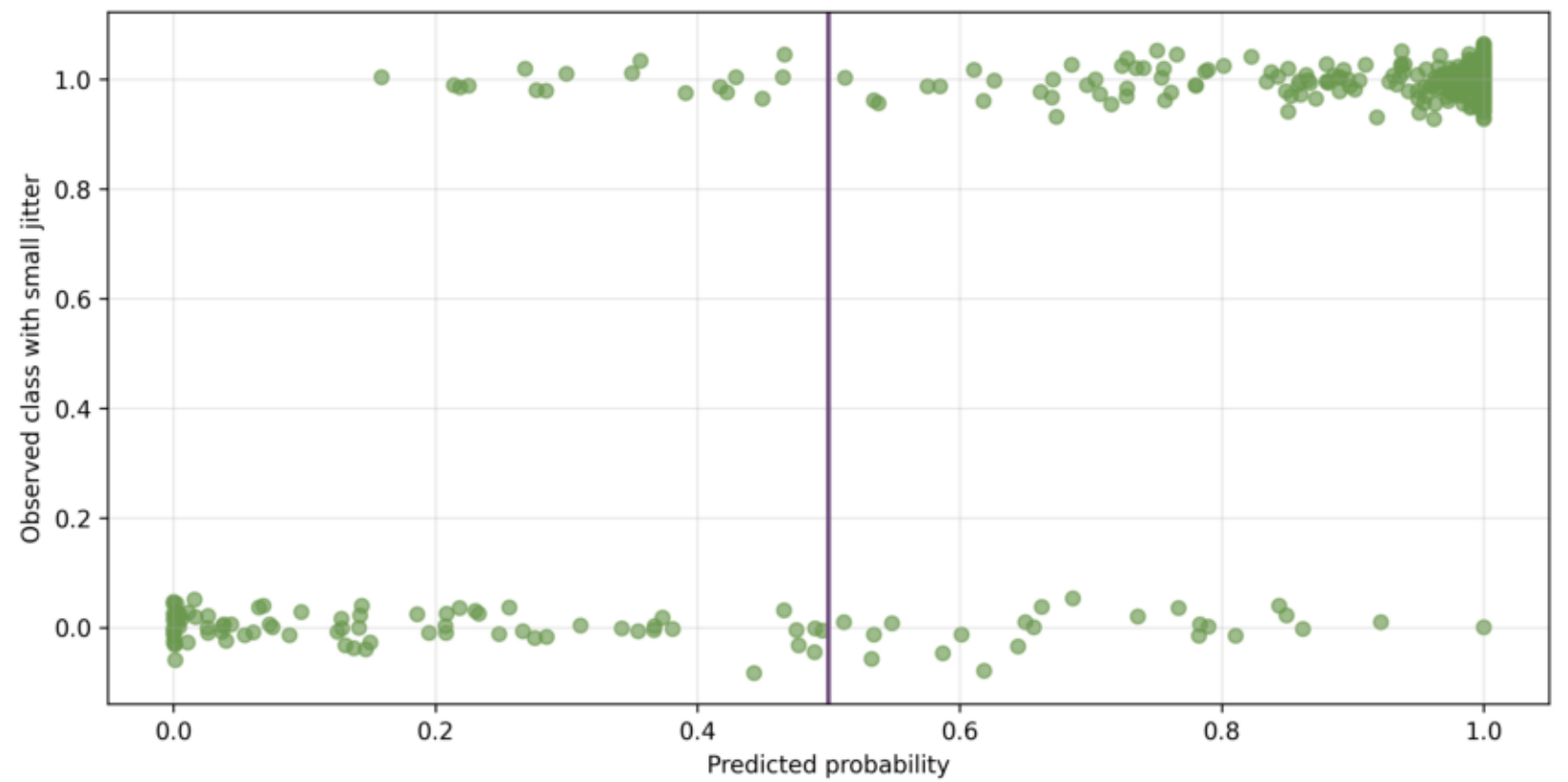
Predicted probability distribution

Probit predicted probabilities after applying the standard normal CDF to the linear predictor.



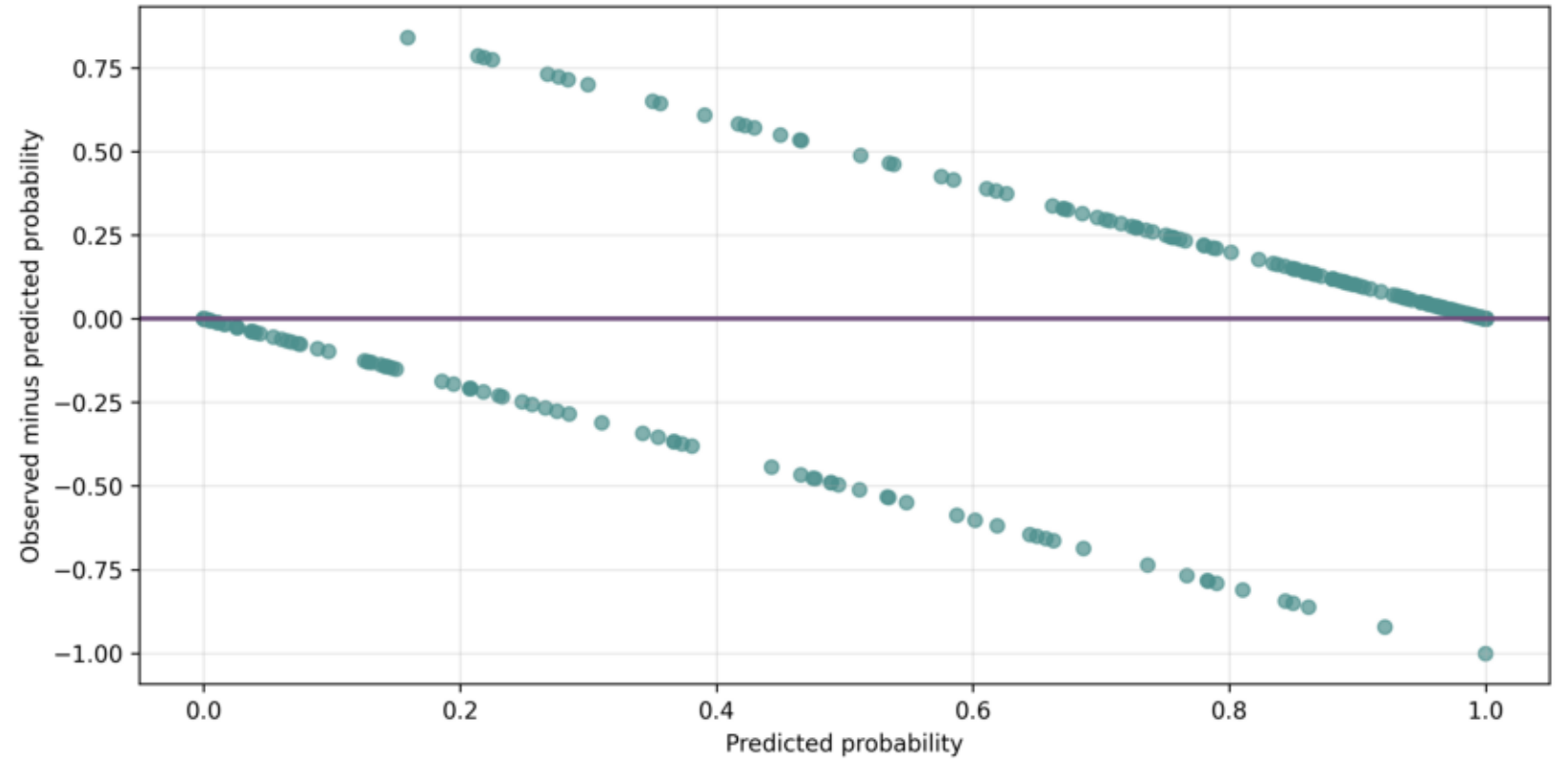
Observed class vs predicted probability

Cases near 0 have lower estimated probability; cases near 1 have higher estimated probability.



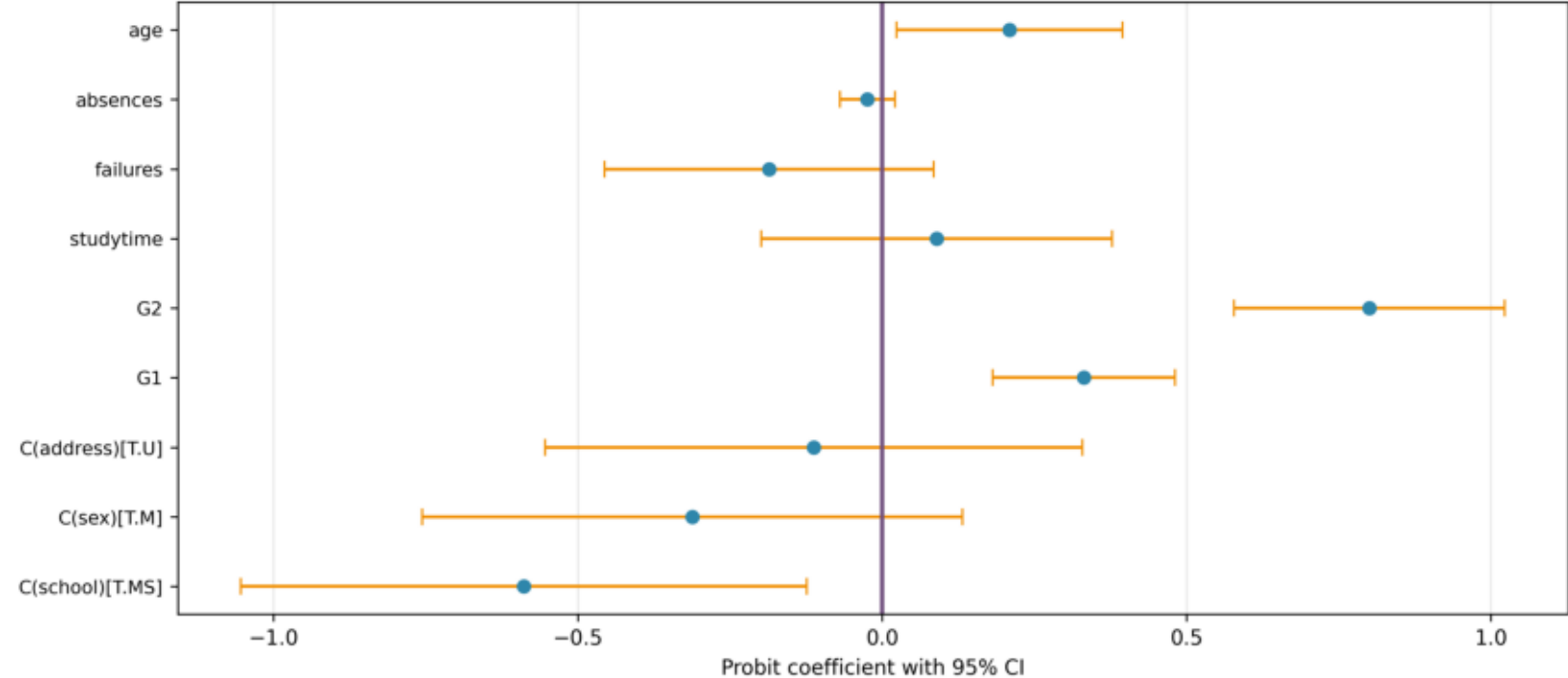
Probability residuals vs fitted probabilities

Residuals help detect systematic underprediction or overprediction.



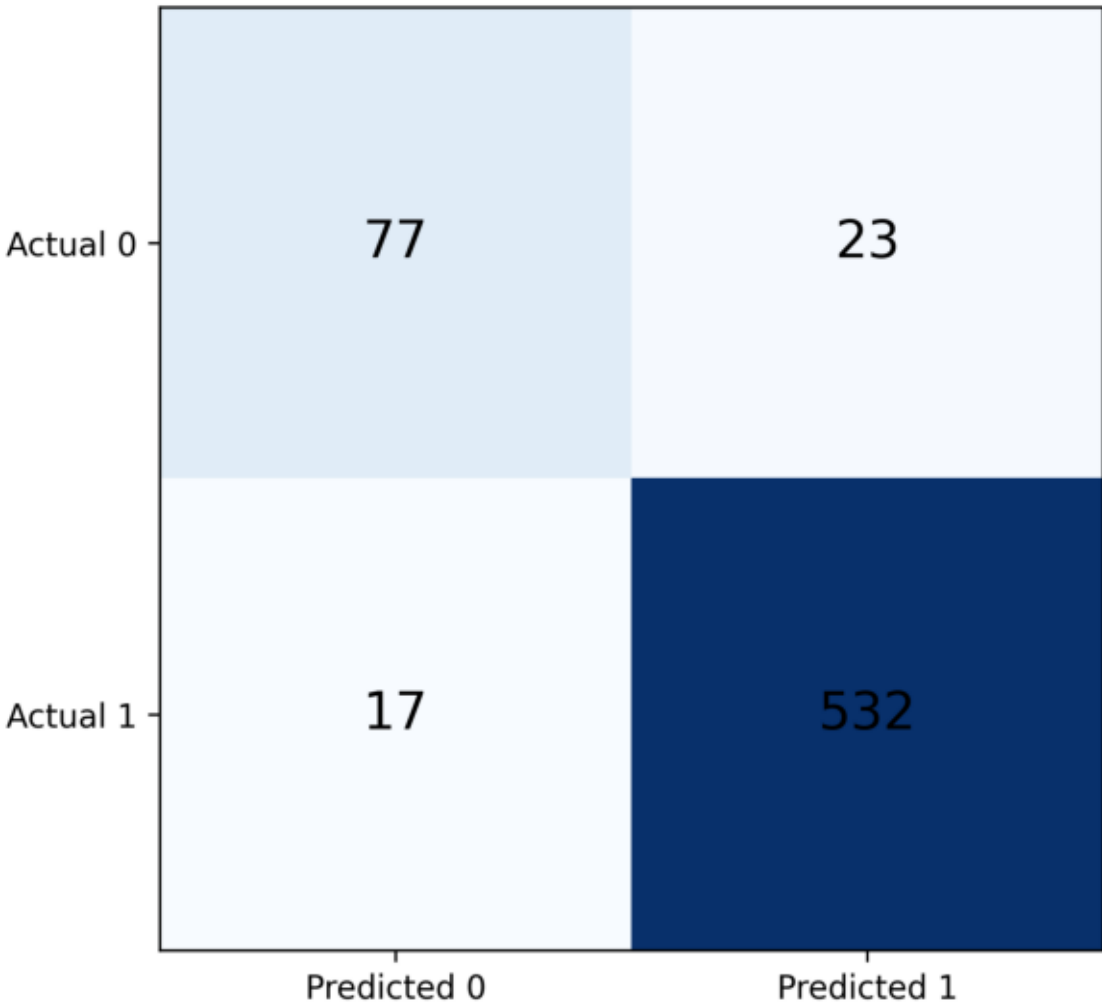
Probit coefficient plot

Positive coefficients increase the latent normal score and predicted probability.



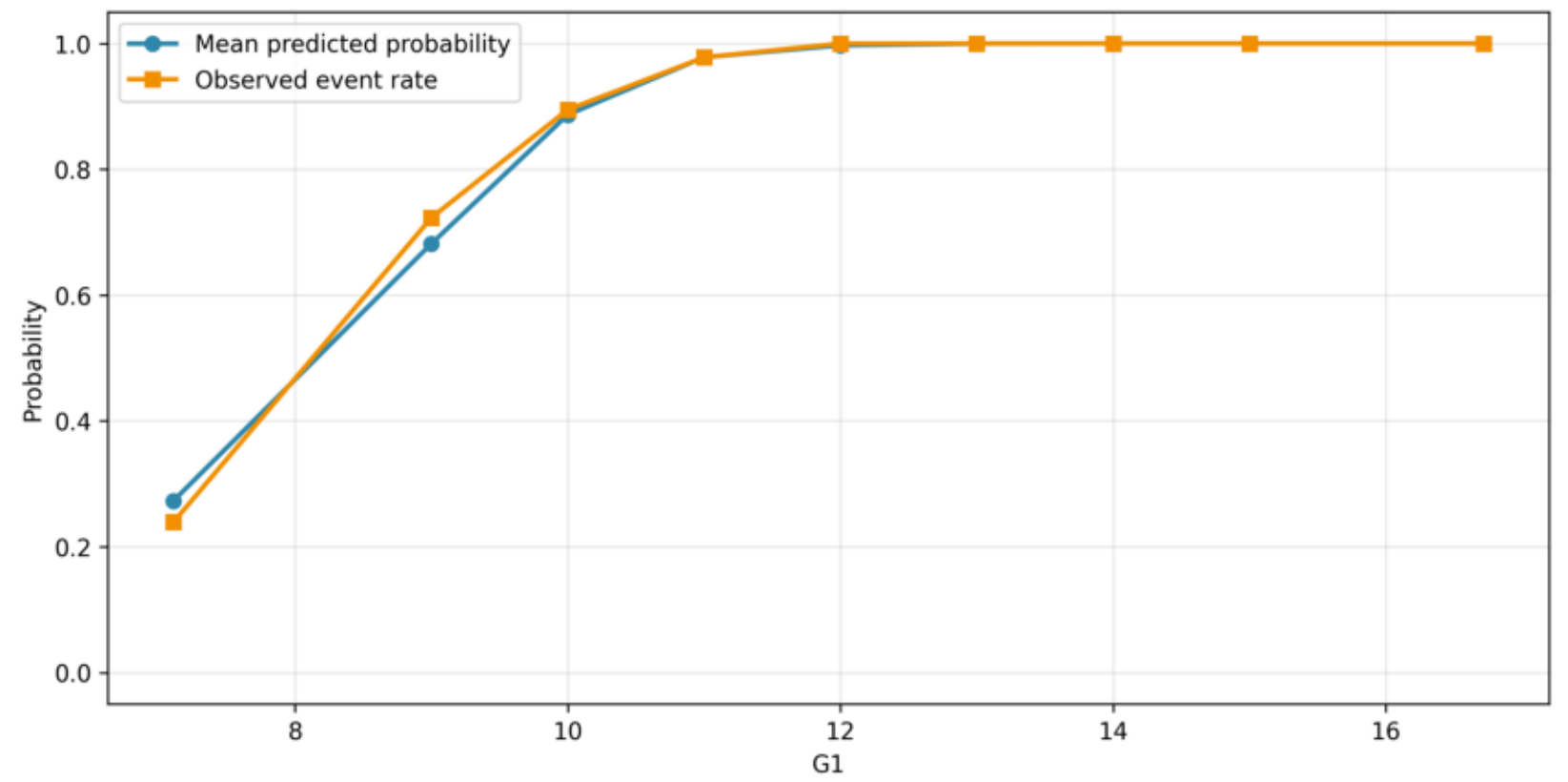
Confusion matrix at 0.50 cutoff

Classification counts from the fitted probit model.



Predicted probability by G1

Binned means compare observed event rate with fitted probit probabilities.



Linear predictor distribution

The probit model converts this z-score scale to probability using the standard normal CDF.

