Smoking and Mortality Risk (NHANES 2013–2014)

December 3, 2025

1 Data

This section details the source datasets, how they are obtained, and the basic shape and missingness patterns. DEMO and SMQ tables are joined on SEQN per NHANES cycle; mortality is linked via public-use files.

Data

file	wave	years	n	k
DEMO_H.csv	H	2013–2014	10 175.00	47.00
SMQ_H.csv	H	2013–2014	7168.00	32.00
NHANES_2013_2014_MORT_2019_PUBLIC.csv	NA	Mortality 2019 linkage	10 175.00	8.00

source: Cleaned matrices (label = dead) from clean-data/
NHANES cycles: H (2013–2014)
Observations (n): 6098
Variables (k): 8
Total missing cells: 0 (0.00%)

2 Cleaned Data

We construct X and y from the raw join: pick a single label (dead preferred) and drop rows only if the label is missing. Features are imputed (numeric: median + missing flags; categorical: Unknown level); age and PIR are standardized; we add age_sq and log_income. Categorical factors (smoker, sex) are used directly.
Label variable: dead
Features used: age, sex, smoker, smoker_code, age_sq, log_income
Initial n: 6111; **Dropped on label only:** 13; **Final n:** 6098
Columns with missing (pre-impute counts): income, dead
Modeled numeric features: age (standardized), age_sq, log_income (from PIR; raw PIR is standardized for preprocessing but not used directly in modeling) **Categoricals in X:** smoker and sex as factors. Race and education are excluded from X and kept only for descriptive/meta. **Smoker categories:** Never/Former/Current coded as 1/2/3 in smoker_code (counts shown in summary stats). **Head of Feature Matrix (X)**

age	sex	smoker	smoker_code	age_sq	log_income
1.17	male	Former	2.00	1.37	0.61
0.36	male	Current	3.00	0.13	1.02
1.33	male	Former	2.00	1.78	1.71
1.39	female	Never	1.00	1.92	1.79
0.47	male	Former	2.00	0.22	1.76
0.74	female	Never	1.00	0.54	1.79

Head of Label Vector (y)

dead
FALSE
TRUE
FALSE
TRUE
FALSE
FALSE

Head of Cleaned Full Frame

id	age	sex	income	smoker	smoker_code	dead	wave	age_missing	income_missing	smoke
73 557.00	1.17	male	−0.99	Former	2.00	FALSE	H	0.00	0.00	0.00
73 558.00	0.36	male	−0.40	Current	3.00	TRUE	H	0.00	0.00	0.00
73 559.00	1.33	male	1.32	Former	2.00	FALSE	H	0.00	0.00	0.00
73 561.00	1.39	female	1.62	Never	1.00	TRUE	H	0.00	0.00	0.00
73 562.00	0.47	male	1.49	Former	2.00	FALSE	H	0.00	0.00	0.00
73 564.00	0.74	female	1.62	Never	1.00	FALSE	H	0.00	0.00	0.00

Head of Meta (IDs and Missing Flags)

id	wave	age_missing	income_missing	smoker_code_missing
73 557.00	H	0.00	0.00	0.00
73 558.00	H	0.00	0.00	0.00
73 559.00	H	0.00	0.00	0.00
73 561.00	H	0.00	0.00	0.00
73 562.00	H	0.00	0.00	0.00
73 564.00	H	0.00	0.00	0.00

Imputation Counts by Variable

variable	n_imputed
age	0.00
income	480.00
sex	0.00
smoker	0.00
smoker_code	0.00

3 Feature/Label Dictionaries and Metrics

We document each feature and the label, then summarize numeric distributions and categorical compositions to contextualize model inputs. **Feature Dictionary (X)**

column	description
age	Age (years), standardized
sex	Sex factor: female/male
smoker	Smoking status factor: Never/Former/Current from SMQ020/040
smoker_code	Ordinal smoking code: Never=1, Former=2, Current=3; 0 if imputed
age_sq	Age squared, from standardized age
log_income	log1p(PIR) from imputed income

Label Dictionary (y)

label	description
dead	Mortality status from public-use linked mortality file (MORTSTAT==1)

X Metrics (Numeric)

	Unique	Missing Pct.	Mean	SD	Min	Median	Max
age	63	0	0.0	1.0	−1.6	0.0	1.8
smoker_code	3	0	1.6	0.8	1.0	1.0	3.0
age_sq	63	0	1.0	0.9	0.0	0.7	3.1
log_income	425	0	1.1	0.5	0.0	1.1	1.8

X Metrics (Categorical)

		N	%
sex	female	3187	52.3
	male	2911	47.7
	Unknown	0	0.0
smoker	Never	3523	57.8
	Former	1344	22.0
	Current	1231	20.2
	Unknown	0	0.0

y Metrics

y	count
FALSE	5632.00
TRUE	466.00

Class Imbalance and Class Weights

Let $\pi = P(y = 1)$ denote the prevalence of deaths. When y is imbalanced (π small), unweighted logistic loss can prioritize the majority class. A common remedy is to use inverse-prevalence weights: $w_1 = 1/\pi$ for positives and $w_0 = 1/(1 - \pi)$ for negatives, often normalized so the average weight equals one. Weighted log-likelihood becomes $L(\beta) = -\sum_i w_i \left[y_i \log p_i + (1 - y_i) \log(1 - p_i)\right]$, with $p_i = \text{logit}^{-1}(X_i\beta)$. **Prevalence and Suggested Class Weights**

quantity	value
n_total	6098.00
n_positive (y=1)	466.00
n_negative (y=0)	5632.00
prevalence (y=1)	0.08
w1 = 1/prevalence	13.09
w0 = 1/(1-prevalence)	1.08
w1_normalized	6.54
w0_normalized	0.54

4 Transformations

This section explains the transformation logic and motivations: label-only dropping preserves sample size; standardized numerics help comparability; categorical factors (smoker, sex) are used directly; smoker categories follow CDC coding of SMQ020/040.

Processing Notes

Joins and Scope: Person-level joins are performed on `SEQN` within each NHANES cycle (DEMO + SMQ), then rows are stacked across cycles.

Smoker Derivation (SMQ):

- If `SMQ020 == 2` (*has not smoked 100 cigarettes*): label as Never.
- If `SMQ020 == 1` and `SMQ040 == 3` (*currently do not smoke*): label as Former.
- If `SMQ020 == 1` and SMQ040 equals 1 or 2 (*every day or some days*): label as Current.
- Otherwise: smoker is set to NA (insufficient or inconsistent information).

Smoker Coding: Factor levels Never/Former/Current and an ordinal code `smoker_code` = 1/2/3 are created for modeling.

Other Harmonization:

- Sex normalized to `female/male` factor.
- Income set to NHANES PIR (`INDFMPIR`) and standardized jointly with age.
- Rows with missing among `id, age, sex, income, smoker, smoker_code,` and the label (`dead`) are dropped for the clean matrix.

Standardization: Numeric predictors (age, income) are centered and scaled to unit variance.

Label: The modeling label is `dead`. **Smoker Categories (Counts)**

smoker	count
Never	3523.00
Former	1344.00
Current	1231.00
Unknown	0.00

5 Models

We estimate two logistic models on the cleaned matrices: (1) a baseline weighted logit with HC1-robust standard errors, and (2) a weighted regularized logit selected by cross-validation.

Model 1: Logit (Weighted, HC1 robust)

Math: $y = \beta_0 + \beta_1 \cdot 1\{smoker = Former\} + \beta_2 \cdot 1\{smoker = Current\} + \beta_3 age + \beta_4 1\{sex = male\} + \beta_5 income + \varepsilon$.

Coefficients (robust SEs):

	title
(Intercept)	0.267
	(0.193)
smokerFormer	−0.540***
	(0.184)
smokerNever	−1.142***
	(0.188)
age	1.303***
	(0.095)
age_sq	0.451***
	(0.069)
sexmale	0.174
	(0.128)
log_income	−0.921***
	(0.138)
Num.Obs.	6098
RMSE	0.39
Std.Errors	Custom

* p <0.1, ** p <0.05, *** p <0.01

Classification metrics (in-sample)

Metric	Value
Accuracy	0.78
LogLoss	0.46
Brier	0.15
ROC_AUC	0.87
PR_AUC	0.38

Classification metrics (out-of-sample)

Metric	Value
Accuracy	0.79
LogLoss	0.46
Brier	0.15
ROC_AUC	0.88
PR_AUC	0.39

Model 2: Regularized Logit (CV, Weighted)

Math: $\hat{\beta} = \arg\min_{\beta} \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$, with λ chosen by cross-validation.

Nonzero coefficients:

lambda	term	estimate
lambda.min	(Intercept)	0.24
lambda.min	smokerFormer	−0.50
lambda.min	smokerNever	−1.11
lambda.min	age	1.29
lambda.min	age_sq	0.44
lambda.min	sexmale	0.16
lambda.min	log_income	−0.91
lambda.1se	(Intercept)	−0.13
lambda.1se	smokerNever	−0.54
lambda.1se	age	1.09
lambda.1se	age_sq	0.32
lambda.1se	log_income	−0.60

Classification metrics (in-sample)

Metric	Value
Accuracy	0.78
LogLoss	0.46
Brier	0.15
ROC_AUC	0.87
PR_AUC	0.38

Classification metrics (out-of-sample)

Metric	Value
Accuracy	0.79
LogLoss	0.46
Brier	0.15
ROC_AUC	0.88
PR_AUC	0.39

6 Diagnostics

We start with multicollinearity (VIF), then show per-model diagnostic plots.

Variance Inflation Factors (VIF).

```
               GVIF Df GVIF^(1/(2*Df))
smoker     1.179329  2        1.042098
age        1.109948  1        1.053541
age_sq     1.049561  1        1.024481
sex        1.035840  1        1.017762
log_income 1.075912  1        1.037262
```

Model 1: Logit

We estimate a weighted logistic regression; residual plots use deviance residuals. Predicted probabilities are also displayed as a 1–5 risk index.

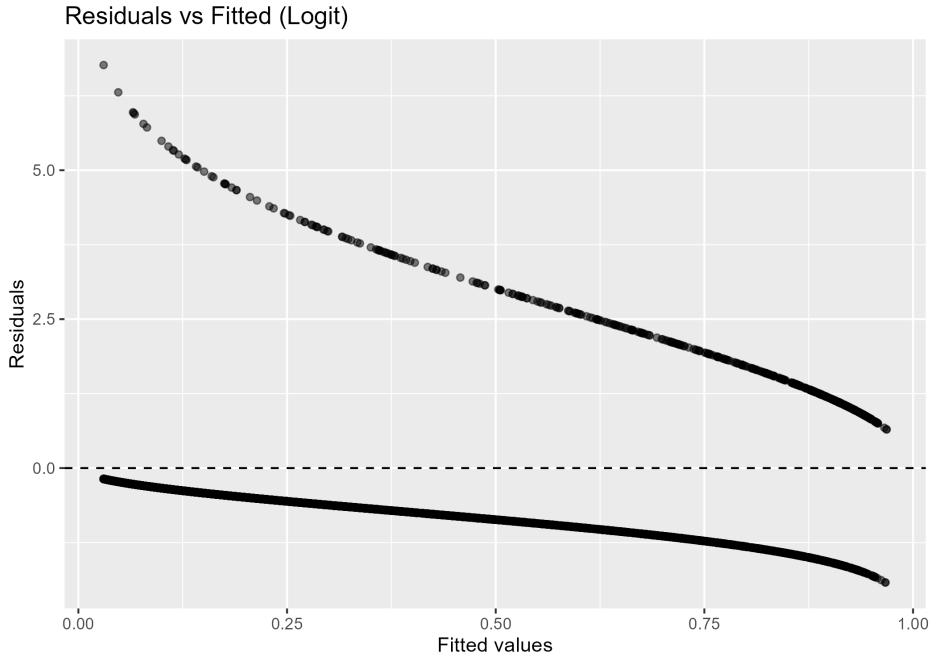

Figure 1: Logit: Residuals vs Fitted

Model 2: Regularized Logit (CV)

We fit a logistic regression with an L_1 penalty and 10-fold cross-validation, using the same class weights as Model 1. Residuals/QQ use predictions at λ_{min}.

7 Coefficient Comparison

Logit vs Regularized Logit (lambda.min and lambda.1se)

term	base_estimate	lasso_min	lasso_1se
(Intercept)	0.27	0.24	−0.13
age	1.30	1.29	1.09
age_sq	0.45	0.44	0.32
log_income	−0.92	−0.91	−0.60
sexmale	0.17	0.16	
smokerFormer	−0.54	−0.50	
smokerNever	−1.14	−1.11	−0.54

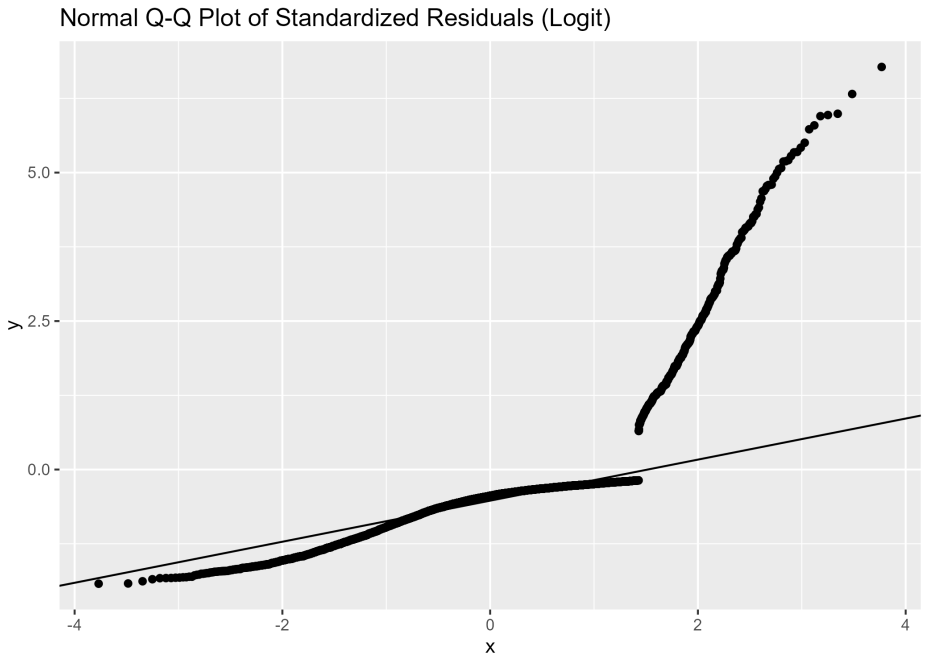

Figure 2: Logit: Normal Q-Q Plot

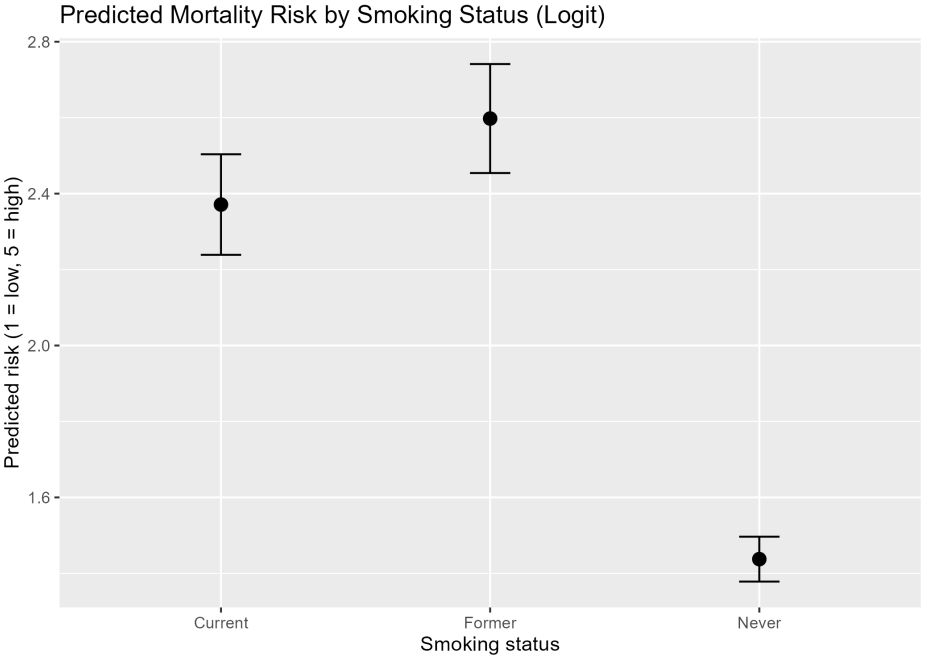

Figure 3: Logit: Predicted Risk by Smoking Status (scaled 1–5)

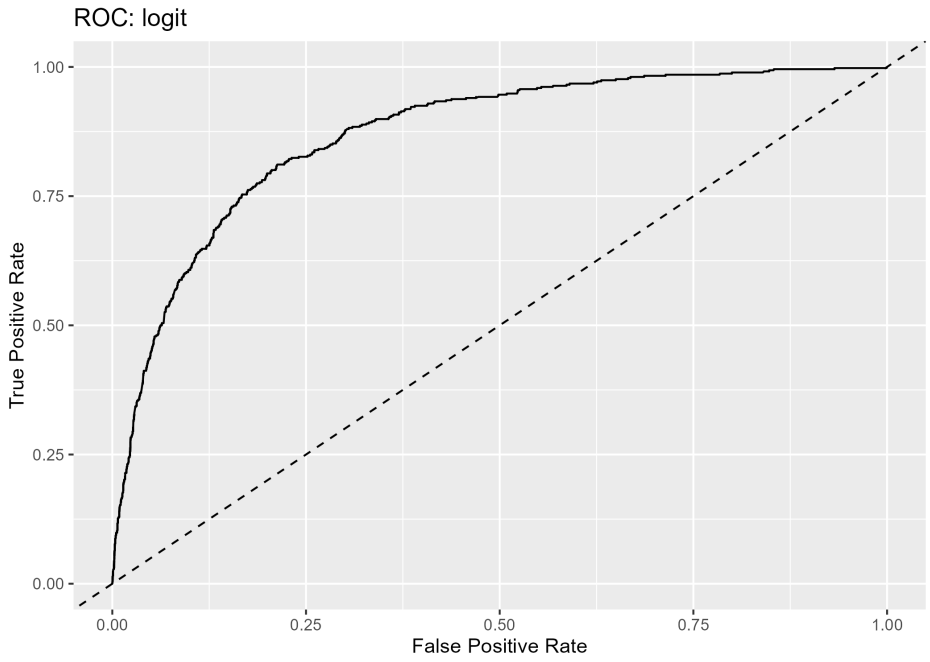

Figure 4: Logit: ROC Curve

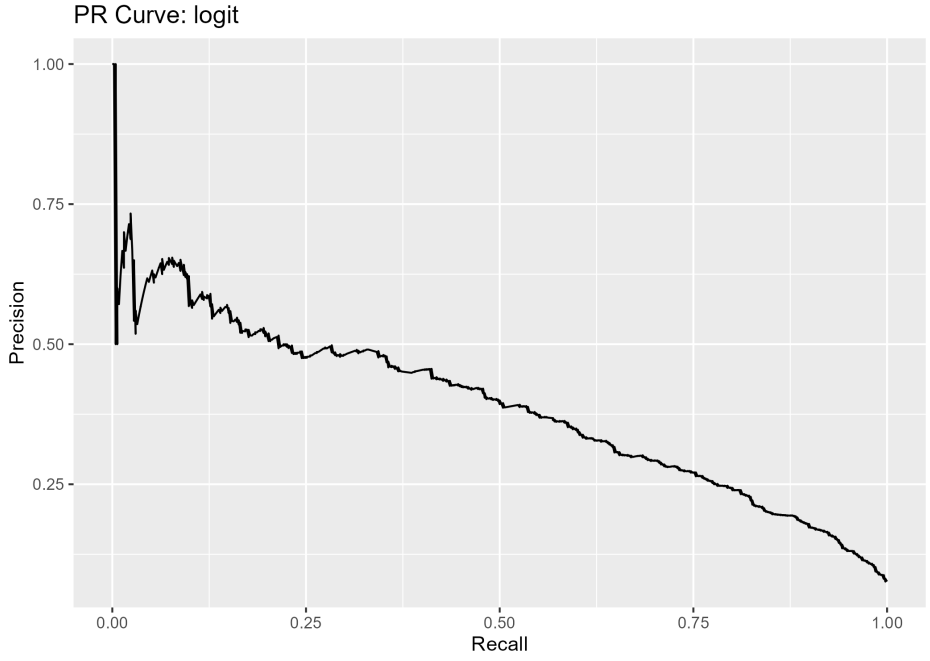

Figure 5: Logit: Precision–Recall Curve

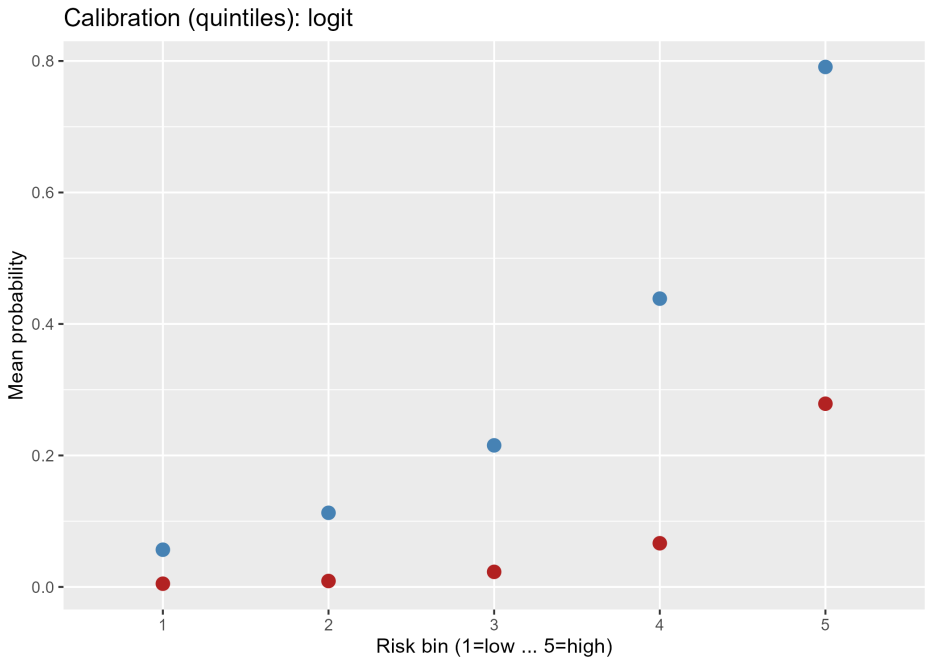

Figure 6: Logit: Calibration by Risk Quintile

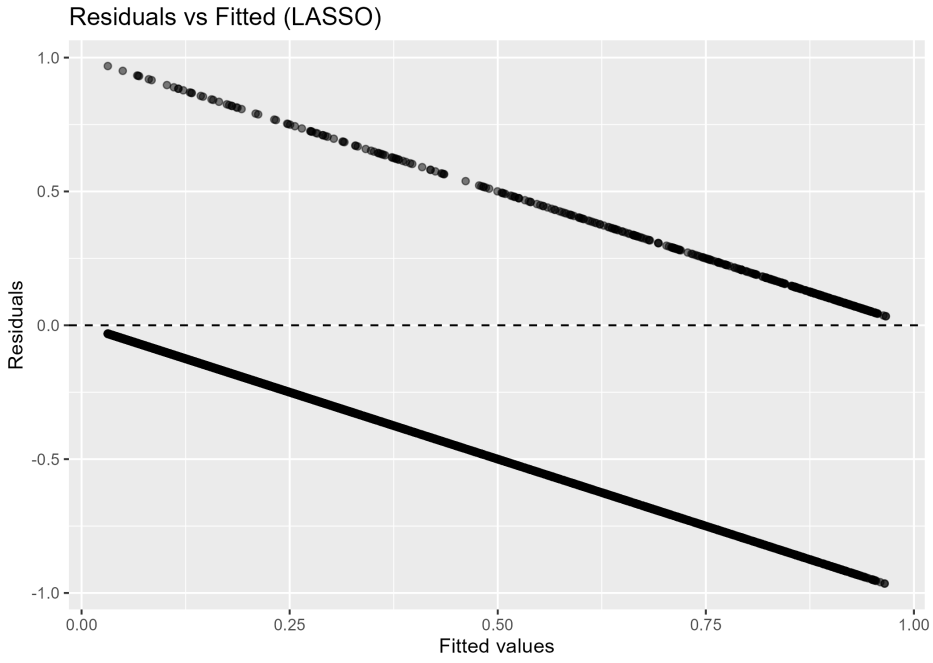

Figure 7: Regularized Logit: Residuals vs Fitted (lambda.min)

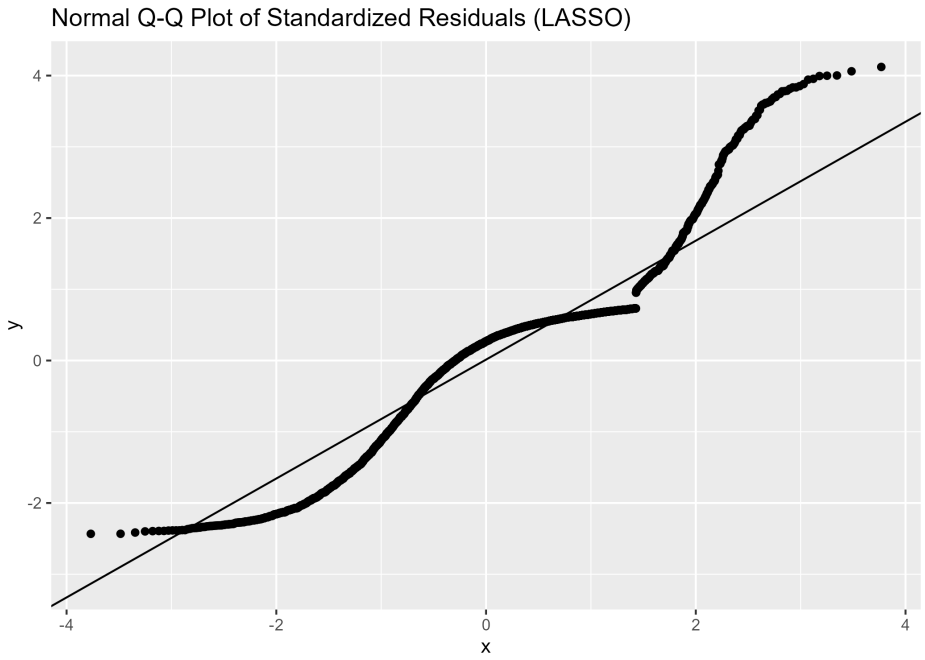

Figure 8: Regularized Logit: Normal Q-Q Plot (lambda.min)

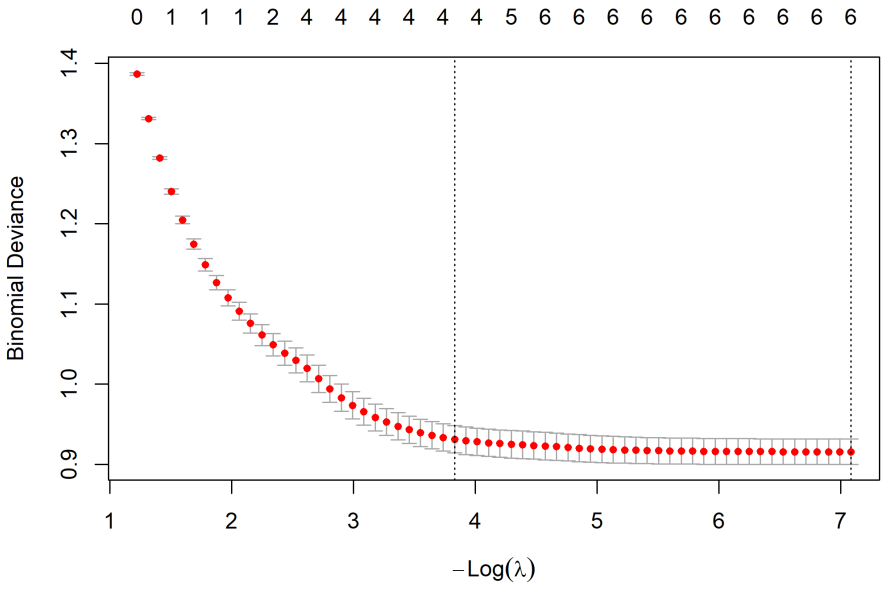

Figure 9: Regularized Logit: CV Error vs Lambda (glmnet)

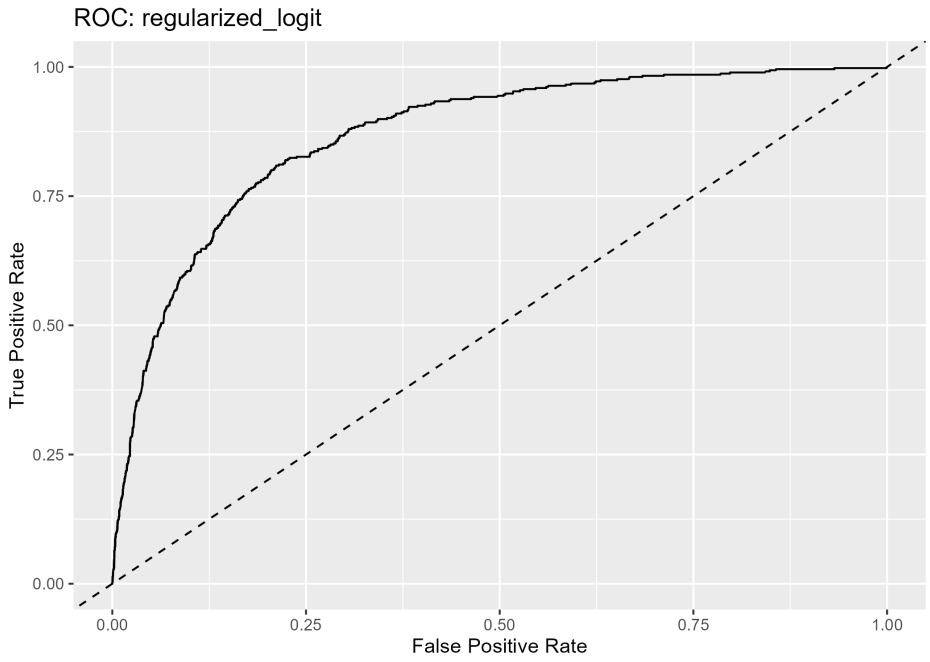

Figure 10: Regularized Logit: ROC Curve

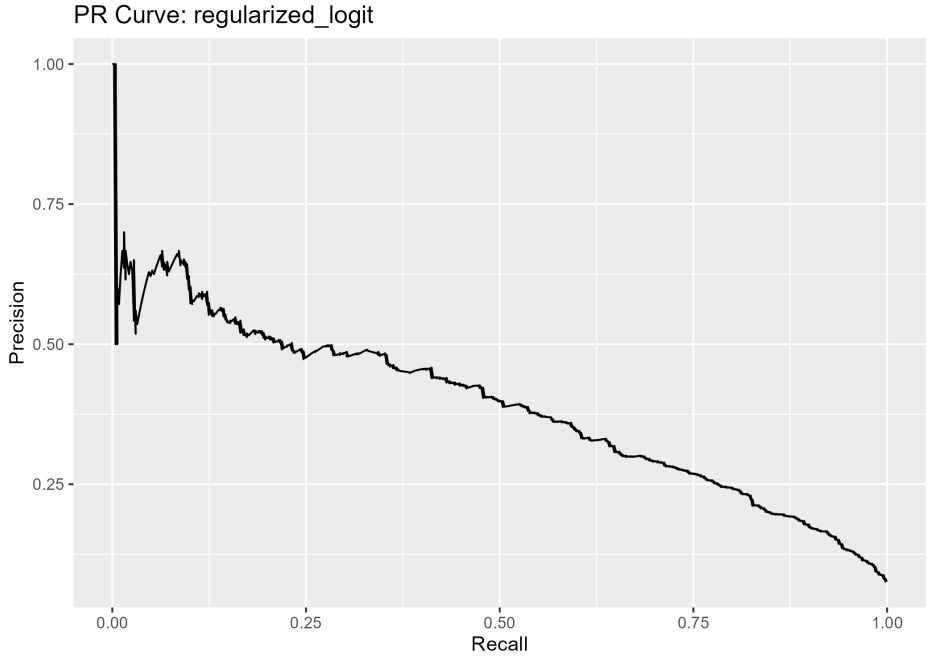

Figure 11: Regularized Logit: Precision–Recall Curve

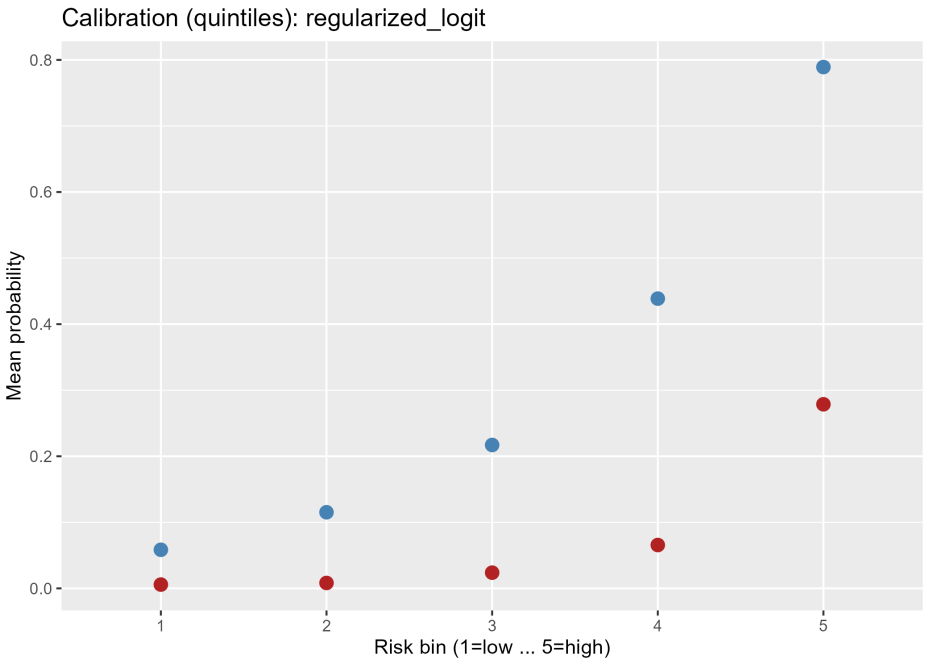

Figure 12: Regularized Logit: Calibration by Risk Quintile