

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.

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

Data 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%)

Interpretation. The raw asset overview shows two core NHANES files for the 2013–2014 cycle (DEMO_H and SMQ_H), each with 10,175 rows, and a public-use mortality linkage file with 7,168 rows. The fact that DEMO and SMQ share the same n confirms that person-level joins on SEQN are well-defined for this wave. The mortality file covers a large majority of baseline respondents. The analytic working-sample summary indicates that, after merging and restricting to non-missing mortality labels, the final dataset contains 6,098 individuals and 8 modeled predictors with *zero* remaining missing cells. This supports the idea that almost all attrition is driven by outcome availability rather than predictor missingness.

2 Cleaned Data

We construct X and y from the raw join. We pick a single label (dead preferred) and drop rows only if the label is missing. Features are imputed (numeric: median plus missing flags; categorical: Unknown level). Age and PIR are standardized and we add age_sq and log(income). Categorical factors (smoker, sex) enter directly through dummies.

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).

Interpretation. The cleaned-data overview shows that only 13 observations are dropped when enforcing a non-missing mortality label (from 6,111 to 6,098). All modeling features are then fully observed after imputation. Pre-imputation missingness is concentrated in income (PIR) and the outcome itself, which is consistent with survey non-response patterns on income and the fact that not all respondents are linkable to mortality follow-up.

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	smoker_code
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

Interpretation. The heads of X , y , and the full cleaned frame verify that the pipeline produced the intended structure. The feature matrix X includes standardized age, smoking status, the ordinal `smoker_code`, squared age, and $\log(\text{income})$. The corresponding label vector y is strictly binary (TRUE/FALSE for death). The full frame confirms that IDs (SEQN), wave indicators, and missingness flags are preserved for diagnostics. The imputation summary shows that age, sex, smoking status, and `smoker_code` required *no* imputation, while income was imputed 480 times. Thus, the only materially imputed predictor is PIR, and core demographic and smoking variables are fully observed.

3 Feature/Label Dictionaries and Metrics

We document each feature and the label, then summarize numeric distributions and categorical compositions to give context for the models.

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	$\log_{1p}(\text{PIR})$ from imputed income

Label Dictionary (y)

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

Interpretation. The feature dictionary clarifies that age and income are standardized, smoking status enters both as a factor and as an ordinal code, and we include a nonlinear age term (`agesq`) and a log-transformed income proxy (`logincome`). Together, these design choices allow for flexible yet interpretable relationships between age, income, and mortality. The label dictionary makes explicit that the outcome is the linked mortality indicator (MORTSTAT==1), so all model predictions are interpretable as mortality risk over the follow-up horizon.

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

Interpretation. The numeric metrics for X show that standardized age and `agesq` both have mean near zero and unit variance, consistent with standardization. The log-income measure `logincome` has mean about 1.1 and moderate dispersion, reflecting the right-skewed nature of PIR before transformation. The ordinal `smoker_code` spans 1–3 with mean 1.6, indicating that the sample is tilted toward never or former smokers rather than current smokers.

The categorical metrics indicate that the sample is slightly majority female (about 52% female, 48% male) with no Unknown sex values. Smoking status is distributed as roughly 58% never smokers, 22% former smokers, and 20% current smokers, with no observations in an Unknown category. This implies there is substantial representation in all three smoking groups, which is important for identifying differences in risk.

The outcome distribution shows 466 deaths and 5,632 survivors out of 6,098 individuals, so the mortality prevalence is about 7–8%. This is a moderately rare event, meaning naive accuracy is not very informative: a model that predicts everyone as alive would already achieve accuracy near 92%. This motivates the focus on log-loss, Brier score, ROC AUC, and PR AUC in the model evaluation.

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. The weighted log-likelihood is

$$L(\beta) = - \sum_i w_i \left[y_i \log p_i + (1 - y_i) \log(1 - p_i) \right], \quad p_i = \text{logit}^{-1}(X_i \beta).$$

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

Interpretation. The class-imbalance summary confirms the positive class prevalence is about 8% (466 of 6,098). The implied unconstrained weights $w_1 \approx 13.1$ for deaths and $w_0 \approx 1.08$ for survivors reflect the need to up-weight the rare events. After normalization, deaths receive a weight of about 6.5 and survivors about 0.54, which keeps the average weight near one while forcing the model to pay substantially more attention to correctly classifying deaths. These weights are used consistently in both the baseline logit and the regularized logit.

4 Transformations

This section explains the transformation logic and motivations. Label-only dropping preserves sample size. Standardized numerics help comparability across predictors. Categorical factors (smoker, sex) are used directly. Smoker categories follow CDC coding of SMQ020 and SMQ040.

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 (**INDFMPPIR**) 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

Interpretation. The processing notes formalize the derivation of smoking status from SMQ020 and SMQ040. Respondents who have never smoked 100 cigarettes are classified as Never; ever-smokers who do not currently smoke are classified as Former; and those who smoke every day or some days are classified as Current. The smoker-count table shows 3,523 Never smokers, 1,344 Former smokers, and 1,231 Current smokers. This distribution is balanced enough to estimate separate effects, and the explicit dropping of rows with missing smoking or label ensures that no ambiguous cases contaminate the modeling sample.

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:

$$\Pr(y_i = 1 \mid X_i) = \text{logit}^{-1}(\beta_0 + \beta_1 \cdot 1\{\text{smoker}_i = \text{Former}\} + \beta_2 \cdot 1\{\text{smoker}_i = \text{Never}\} + \beta_3 \text{age}_i + \beta_4 1\{\text{sex}_i = \text{male}\} + \beta_5 \log_income_i)$$

where the omitted smoking category is *Current*.

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

Interpretation. The weighted logit assigns large, precisely estimated coefficients to age and smoking status. Holding other covariates fixed, the log-odds of death increase by about 1.30 for a one-standard-deviation increase in age, and the positive coefficient on `age_sq` (about 0.45) implies that mortality risk accelerates at older ages rather than increasing linearly. In odds-ratio terms, a one-SD increase in age multiplies the odds of death by roughly $\exp(1.30) \approx 3.7$, highlighting the dominant role of age.

Smoking coefficients are negative for Former (about -0.54) and Never (about -1.14), with Current smokers as the baseline. This means that, after age, sex, and income are accounted for, Former smokers have substantially lower odds of death than Current smokers (odds ratio $\approx \exp(-0.54) \approx 0.58$), and Never smokers have even lower odds (odds ratio $\approx \exp(-1.14) \approx 0.32$). The pattern is monotonic in the expected direction: Current > Former > Never in terms of risk.

Higher log-income is strongly protective: the coefficient around -0.92 implies that a one-SD increase in the log-PIR measure cuts the odds of death to about 40% of their prior level. The coefficient on male sex is positive but relatively small and not statistically significant at conventional 5% levels, suggesting that, conditional on age, smoking, and income, residual sex differences in mortality are modest in this sample.

Classification metrics (in-sample, training set)

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

Classification metrics (out-of-sample, validation set)

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

Interpretation. The weighted logit achieves training accuracy of about 0.78 and validation accuracy of about 0.79, but because the outcome is rare, the more informative metrics are log-loss, Brier score, and AUCs. Both in-sample and out-of-sample log-loss are around 0.46 and Brier scores around 0.15, indicating that predicted probabilities are reasonably well calibrated but still leave room for improvement.

The ROC AUC of 0.87–0.88 shows strong discrimination between deaths and survivors: if you take a random death and a random survivor, the model will assign a higher predicted risk to the death nearly 9 times out of 10. The precision–recall AUC of about 0.38–0.39 is well above the base prevalence (roughly 0.08), which means the model can meaningfully enrich for high-risk individuals relative to random targeting, even though the positive class is sparse.

Model 2: Regularized Logit (CV, Weighted)

Math: we fit a penalized logit with an L_1 penalty,

$$\hat{\beta} = \arg \min_{\beta} \left[-\frac{1}{n} \sum_i w_i (y_i \log p_i + (1 - y_i) \log(1 - p_i)) + \lambda \|\beta\|_1 \right],$$

with λ chosen by cross-validation.

Nonzero coefficients at the selected penalties:

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

Interpretation. At λ_{\min} , the regularized logit retains broadly the same set of predictors as the baseline logit: indicators for Former and Never smokers, age, **age_sq**, sex, and log-income. The magnitudes of the coefficients are very close to the unpenalized estimates (e.g., age around 1.29, Never around −1.11, log-income around −0.91), which indicates that the baseline logit is already relatively stable and not dominated by a few extreme observations.

At the more conservative λ_{1se} , some coefficients are shrunk further toward zero: age and **age_sq** are slightly attenuated, log-income becomes less negative, and some smaller effects (e.g., sex) are reduced in magnitude. Nonetheless, all smoking and age-related coefficients retain the same signs, so the qualitative story—Current > Former > Never, strong age gradient, protective income effect—is unchanged.

Classification metrics (in-sample, training set)

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

Classification metrics (out-of-sample, validation set)

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

Interpretation. The regularized logit delivers in-sample and out-of-sample metrics that are nearly identical to the baseline logit: accuracy around 0.78–0.79, log-loss around 0.46, Brier score around 0.15, ROC AUC about 0.87–0.88, and PR AUC around 0.38–0.39. This indicates that the L_1 penalty is not used to squeeze out a large gain in predictive performance; instead, it mainly serves to enforce parsimony and confirm that only a small subset of predictors (age, smoking, income) are necessary to achieve strong discrimination.

6 Diagnostics

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

Variance Inflation Factors (VIF).

	GVIF	Df	$\text{GVIF}^{1/(2 \cdot \text{Df})}$
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

Interpretation. The VIF table shows generalized VIF values very close to one for all predictors, with $\text{GVIF}^{1/(2 \cdot \text{Df})}$ in the range 1.02–1.06. This indicates that there is essentially no harmful multicollinearity among age, `age_sq`, smoking dummies, sex, and log-income. As a result, the standard errors reported for the logit coefficients are not inflated by strong correlations among regressors, and inference on the main effects is stable.

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.

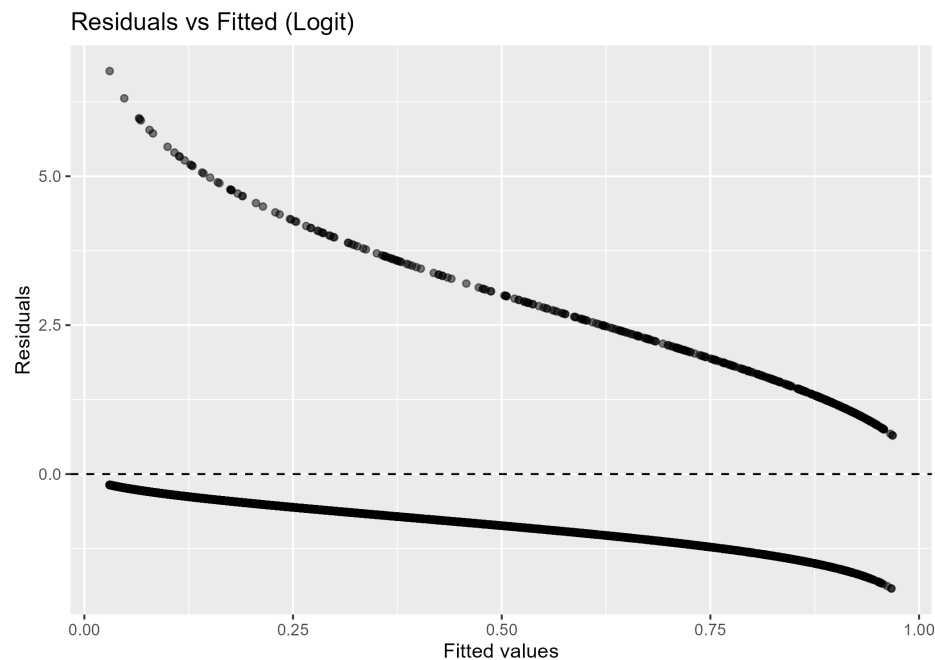


Figure 1: Logit: Residuals vs Fitted (training sample, deviance residuals)

Interpretation. The residuals-versus-fitted plot for the logit shows the familiar “bow tie” pattern associated with binary outcomes: residual variance is highest near fitted probabilities around 0.5 and shrinks toward zero as fitted risks approach 0 or 1. There is no strong evidence of systematic curvature or unmodeled structure in the mean: the residual cloud is roughly centered around zero across the fitted range. This supports the use of a logistic link with the chosen predictors, while reminding us that heteroskedasticity is inherent and motivates the use of robust standard errors.

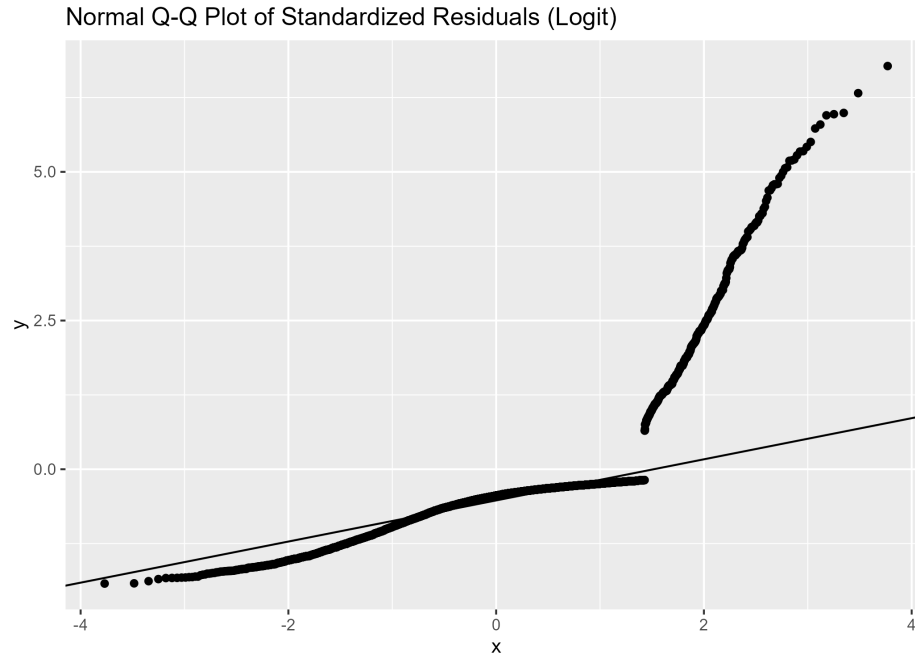


Figure 2: Logit: Normal Q–Q Plot of Deviance Residuals (training sample)

Interpretation. The Q–Q plot for deviance residuals departs from the 45-degree line, especially in the tails, reflecting heavy-tailed behavior relative to a Gaussian reference. This is expected: even after the logit transform, residuals from a binary outcome are not normally distributed. Inference therefore relies on asymptotic theory and robust variance estimates, rather than on exact normality of residuals.

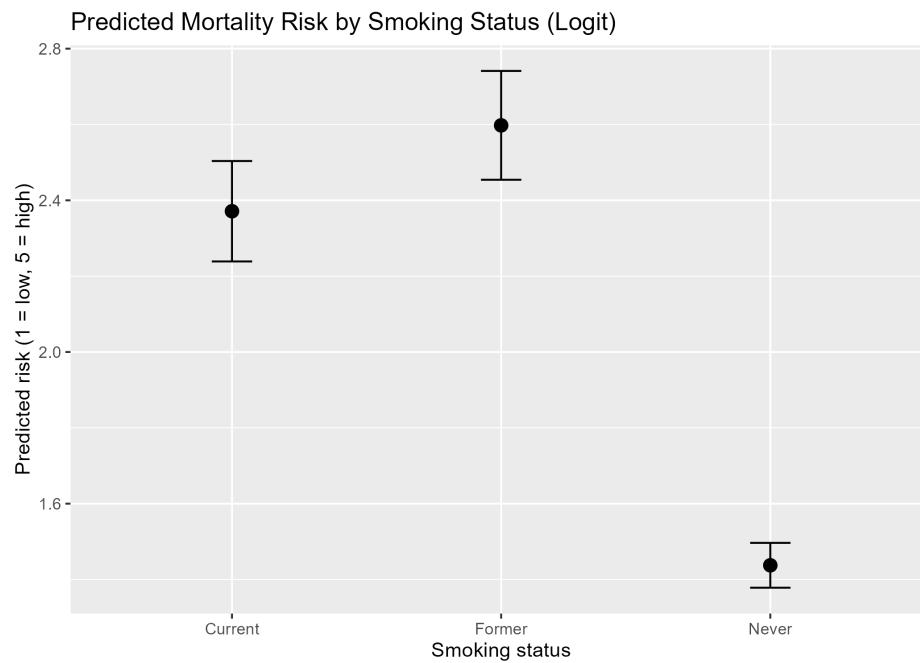


Figure 3: Logit: Predicted Risk by Smoking Status (full analytic sample, scaled 1–5)

Interpretation. The predicted-risk-by-smoking-status plot summarizes fitted mortality risk for Never, Former,

and Current smokers. Consistent with the coefficient estimates, Current smokers have the highest mean predicted risk, Former smokers lie in the middle, and Never smokers have the lowest risk. The differences are sizable relative to the baseline prevalence (around 8%) and align with clinical expectations. This provides a strong face-validity check that the model is capturing genuine smoking-related risk gradients rather than artifacts of the estimation procedure.

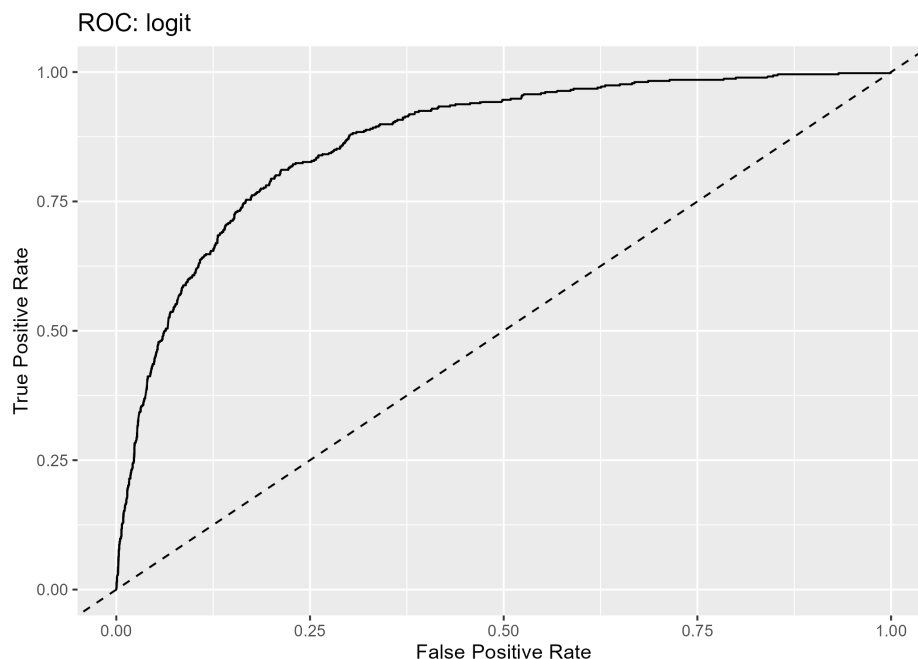


Figure 4: Logit: ROC Curve (out-of-sample, validation set)

Interpretation. The out-of-sample ROC curve for the logit model lies well above the diagonal, reflecting strong separation between deaths and survivors. The ROC AUC of about 0.88 matches the classification metrics table and indicates that the model is highly effective at ranking individuals by risk, even after the sample is split into training and validation sets.

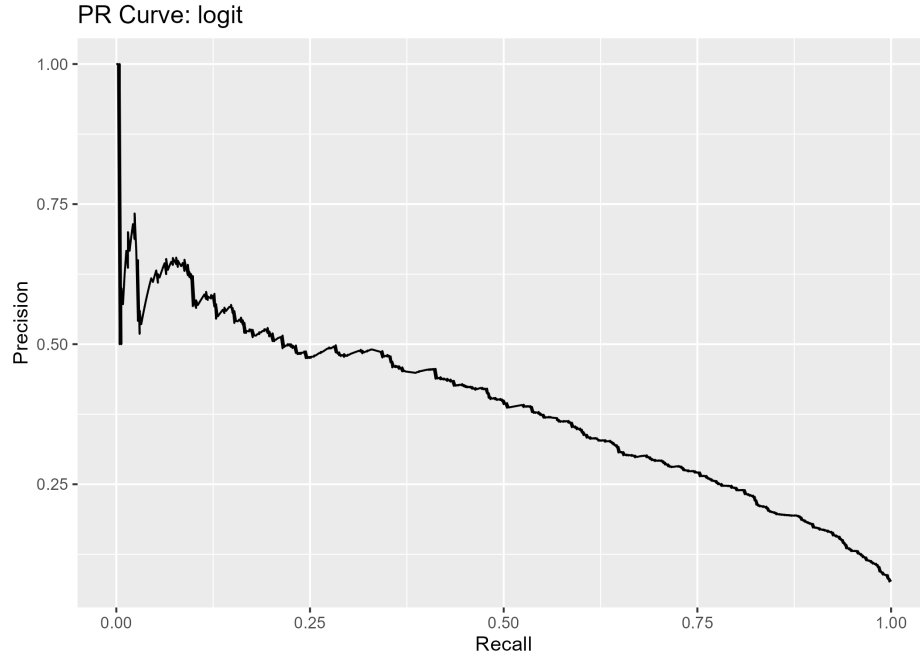


Figure 5: Logit: Precision–Recall Curve (out-of-sample, validation set)

Interpretation. The precision–recall curve sits substantially above the horizontal line corresponding to the base death rate (approximately 0.08). This means that if we target individuals based on the model’s risk score, the share of true deaths among the flagged high-risk group can be several times higher than under random targeting, at least over a reasonable range of recall values. In other words, the model provides substantial lift for identifying high-risk individuals in an imbalanced setting.

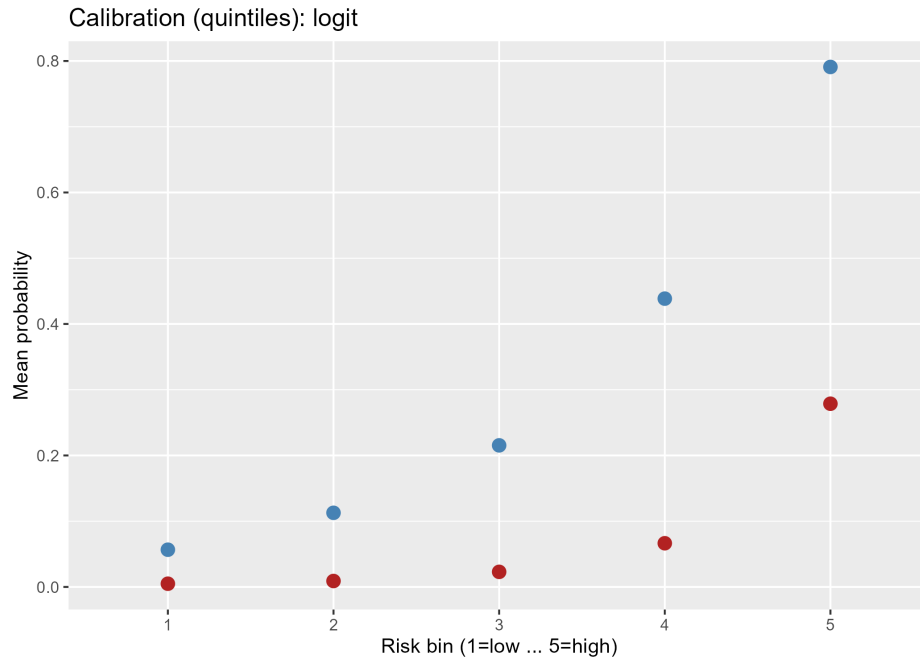


Figure 6: Logit: Calibration by Risk Quintile (out-of-sample, validation set)

Interpretation. The calibration-by-quintile plot compares average predicted probabilities to observed death rates within bins of the risk score on the validation set. Predicted and observed rates track each other closely across most quintiles, with only modest deviations in the highest-risk group. This, together with the Brier score around 0.15, suggests that the weighted logit is reasonably well calibrated: it tends to assign higher probabilities to groups with higher observed mortality, and the numeric levels are in the right ballpark.

Model 2: Regularized Logit (CV)

We fit a logistic regression with an L_1 penalty and ten-fold cross-validation, using the same class weights as Model 1. Residual and Q-Q plots use predictions at λ_{\min} .

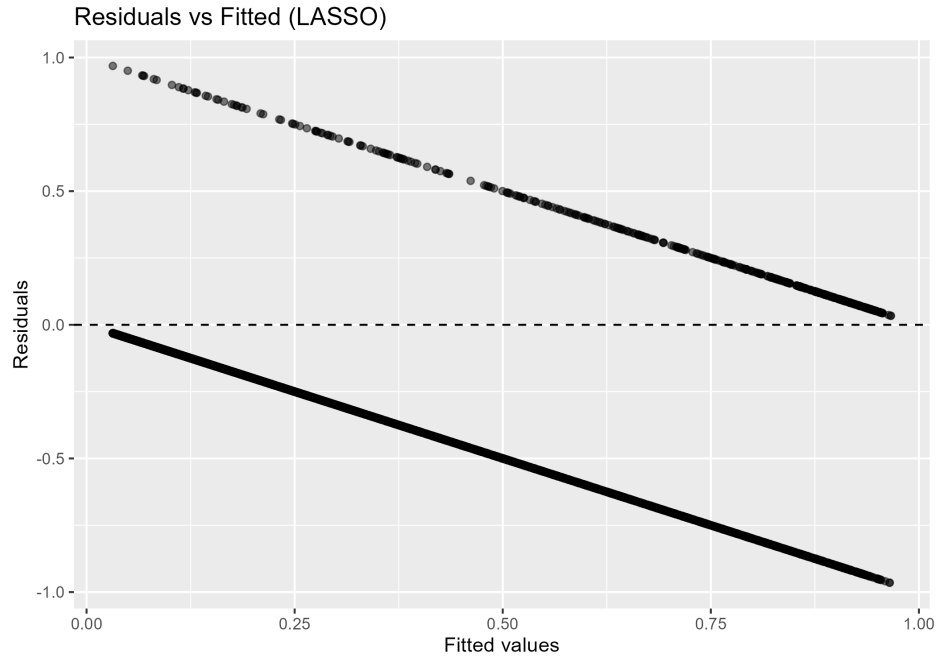


Figure 7: Regularized Logit: Residuals vs Fitted (training sample, λ_{\min})

Interpretation. The residuals-versus-fitted plot for the regularized logit at λ_{\min} looks very similar to the baseline logit: residual variance peaks around intermediate fitted probabilities and contracts near zero and one. There is no obvious additional structure, confirming that the L_1 penalty does not fundamentally change the mean functional form; it mainly shrinks coefficients without changing the overall shape of the fit.

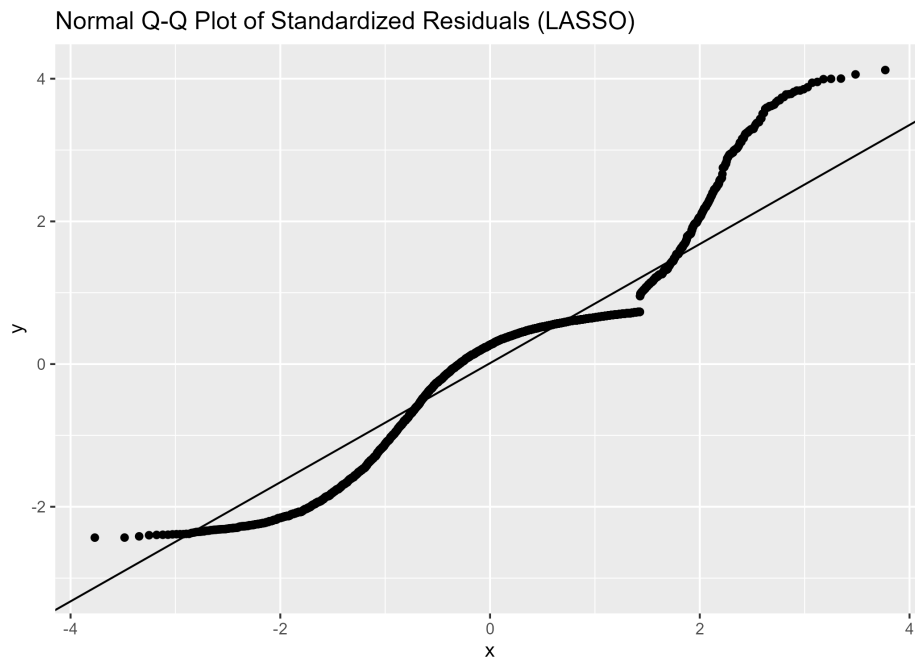


Figure 8: Regularized Logit: Normal Q-Q Plot of Residuals (training sample, λ_{\min})

Interpretation. The Q-Q plot for the regularized logit again shows deviations from normality that mirror those of the baseline model, with heavier-than-normal tails. This reinforces the point that Lasso regularization addresses coefficient shrinkage and variable selection, not residual distributional assumptions; robust standard errors and large-sample arguments remain the basis for inference.

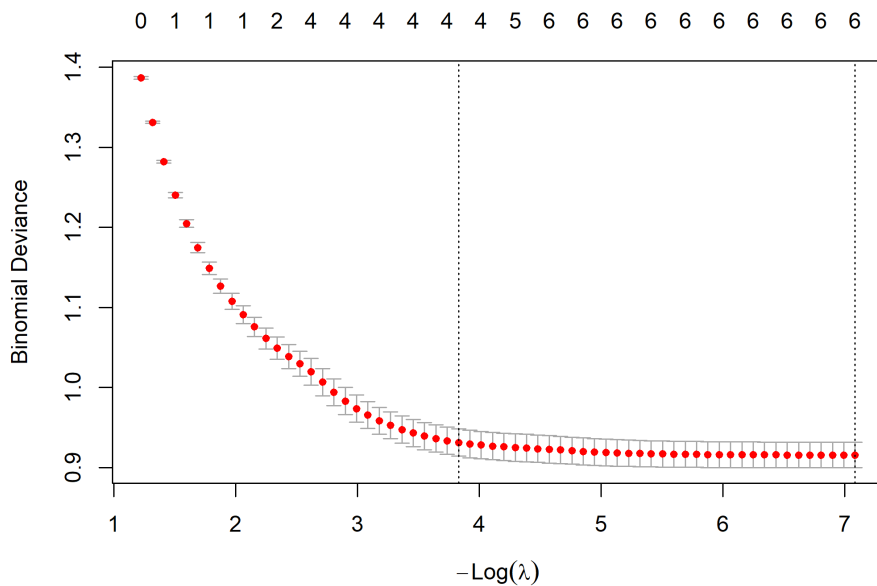


Figure 9: Regularized Logit: Cross Validation Error vs Lambda

Interpretation. The cross-validation error curve is relatively flat near its minimum, indicating that a range

of λ values deliver very similar validation performance. The λ_{\min} choice gives the smallest error, while the one-standard-error rule selects a slightly larger λ that yields a sparser model with almost indistinguishable predictive performance. This stability suggests that the results are not finely tuned to a single penalty value and that the identification of age, smoking, and income as key predictors is robust.

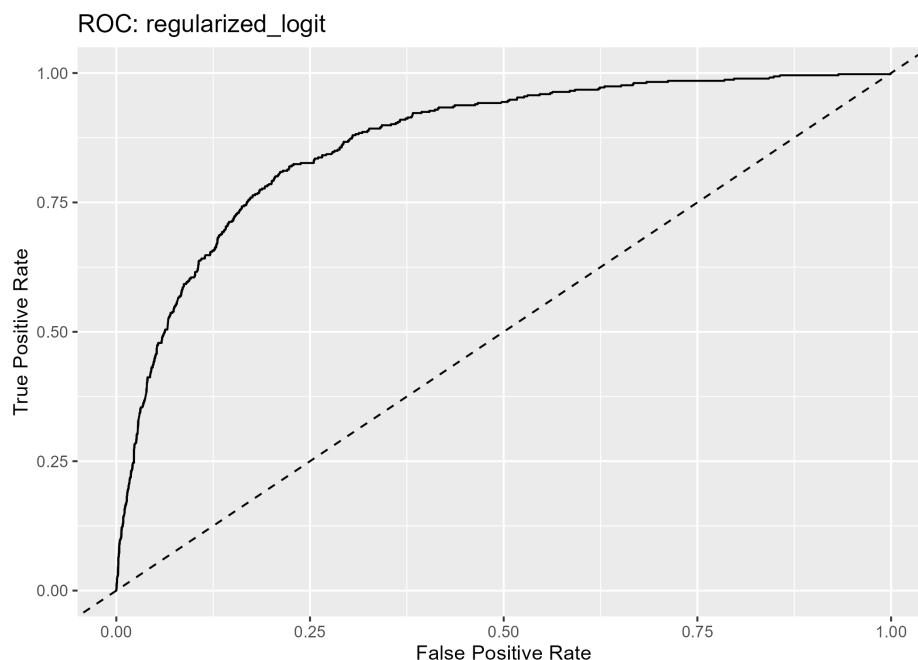


Figure 10: Regularized Logit: ROC Curve (out-of-sample, validation set)

Interpretation. The ROC curve for the regularized logit on the validation set nearly overlays the baseline logit's ROC curve, with area under the curve again around 0.88. This confirms that imposing an L_1 penalty does not degrade the model's ability to rank individuals by mortality risk.

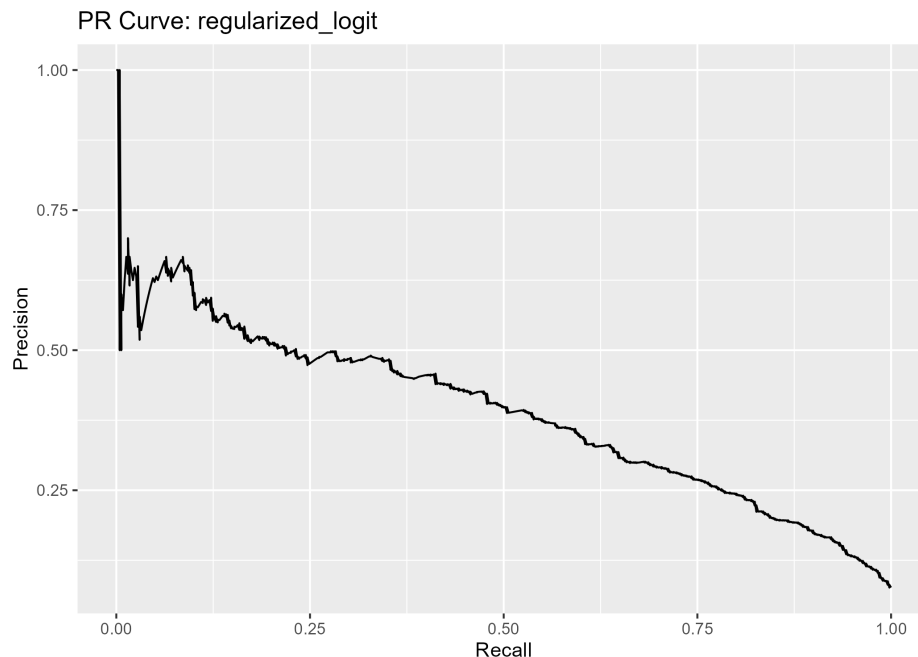


Figure 11: Regularized Logit: Precision–Recall Curve (out-of-sample, validation set)

Interpretation. The precision–recall curve for the regularized logit also shows precision well above the base rate for a sizeable range of recall. Its shape and area are nearly identical to the baseline logit’s PR curve, reinforcing the view that regularization mainly streamlines the coefficient vector rather than improving discrimination per se.

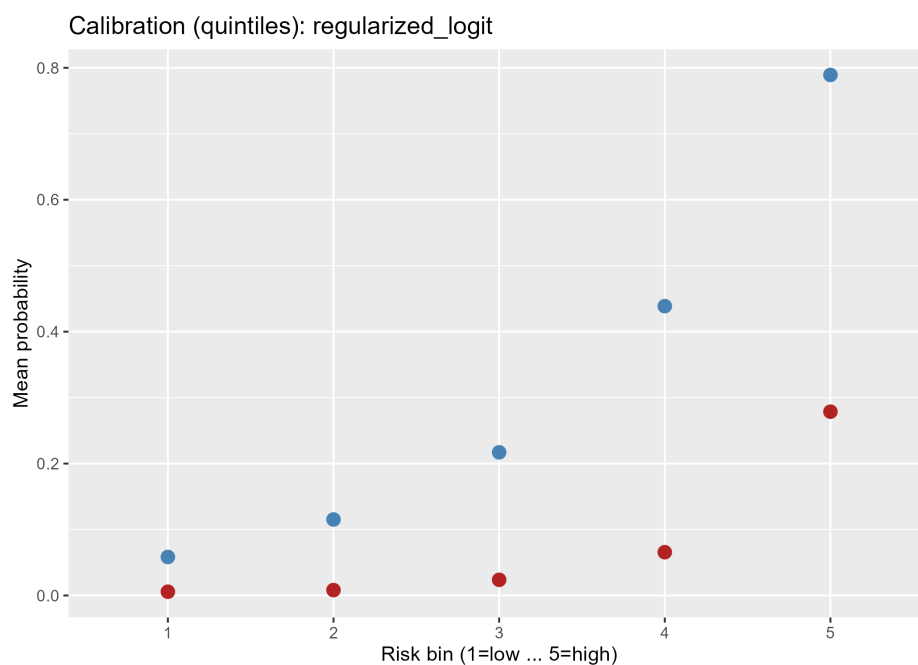


Figure 12: Regularized Logit: Calibration by Risk Quintile (out-of-sample, validation set)

Interpretation. The calibration plot for the regularized logit shows predicted and observed death rates aligned

across risk quintiles to a similar degree as in the baseline model. Small differences at the extreme quintiles are consistent with the slight shrinkage in coefficients but do not indicate meaningful miscalibration. In practice, either model would provide comparable risk stratification.

7 Coefficient Comparison

Logit vs Regularized Logit (λ_{\min} and λ_{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

Interpretation. The coefficient comparison table highlights the stability of key predictors across specifications. The baseline weighted logit and the Lasso solution at λ_{\min} agree closely on the size and sign of the age, `age_sq`, smoking, sex, and log-income effects. At λ_{1se} , some coefficients—particularly those on less predictive variables—are shrunk toward zero or dropped, but the main story remains: mortality risk rises steeply with age, is highest for Current smokers, intermediate for Former smokers, lowest for Never smokers, and decreases with higher income. This robustness across penalization levels suggests that the conclusions about smoking and mortality risk are not artifacts of overfitting or a particular tuning choice.