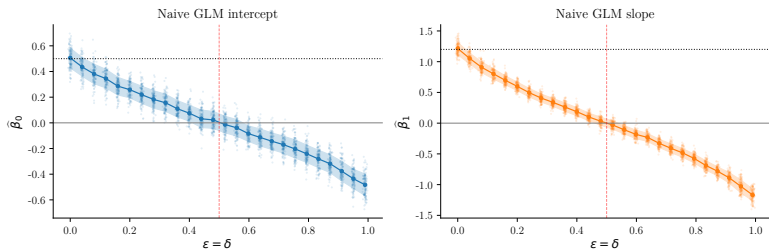


Abstract

Logistic regression is often used when we want to explain a binary outcome, such as benign/malignant or default/no default. The usual model assumes that the recorded label is correct. In many data sets this is too optimistic: a diagnostic test, manual annotation, or administrative record can flip a true label with known error rates ε and δ . If we fit ordinary logistic regression to these noisy labels, the model is fitting the wrong probability and the coefficients are pulled toward zero as the noise increases. The corrected likelihood instead models the label we actually observe, $q_i(\beta) = \delta + (1 - \varepsilon - \delta)p_i(\beta)$, and can recover the clean-label target when the error rates are known and the data still contain information. The trade-off is larger uncertainty: as $\varepsilon + \delta$ approaches 1, the labels contain less signal and the corrected fit becomes unstable. We study this bias–variance trade-off with closed-form calculations, Monte Carlo simulations, the Wisconsin breast-cancer data, and a covariate-dependent noise extension, using bounded L-BFGS-B and interior-fit diagnostics to show when the correction is trustworthy.

Naive fit: logistic regression

Ordinary logistic regression treats $\hat{y}_i \mid x_i$ as $\text{Bern}\{p_i(\beta)\}$, where $p_i(\beta) = H(x_i^\top \beta)$. Under label errors the observed probability is instead $q_i(\beta) = \delta + c p_i(\beta)$, $c = 1 - \varepsilon - \delta$. In the symmetric case, increasing noise pulls the observed probabilities toward 1/2, and the naive coefficient estimates shrink toward zero as the pole $\varepsilon + \delta = 1$ is approached. At the pole the labels carry no signal about β ; past it, the observed labels are more often flipped than correct, so the naive fit follows the sign-reversed direction.



Monte Carlo simulation with $n = 1000$ and $B = 200$ label-flip replicates per noise level. Dots: individual naive GLM fits. Solid: MC mean ± 1 sd. Dotted: true coefficients. Red dashed: pole.

Corrected log likelihood

The correction keeps the extra noise step in the model. With $p_i(\beta) = H(x_i^\top \beta)$ and constant flip rates,

$$q_i(\beta) = P(\hat{y}_i = 1 \mid x_i) = \delta + c p_i(\beta), \quad c = 1 - \varepsilon - \delta.$$

Since $\hat{y}_i \mid x_i \sim \text{Bern}\{q_i(\beta)\}$, the corrected log likelihood is

$$\ell_n^*(\beta) = \sum_{i=1}^n [\hat{y}_i \log q_i(\beta) + (1 - \hat{y}_i) \log(1 - q_i(\beta))].$$

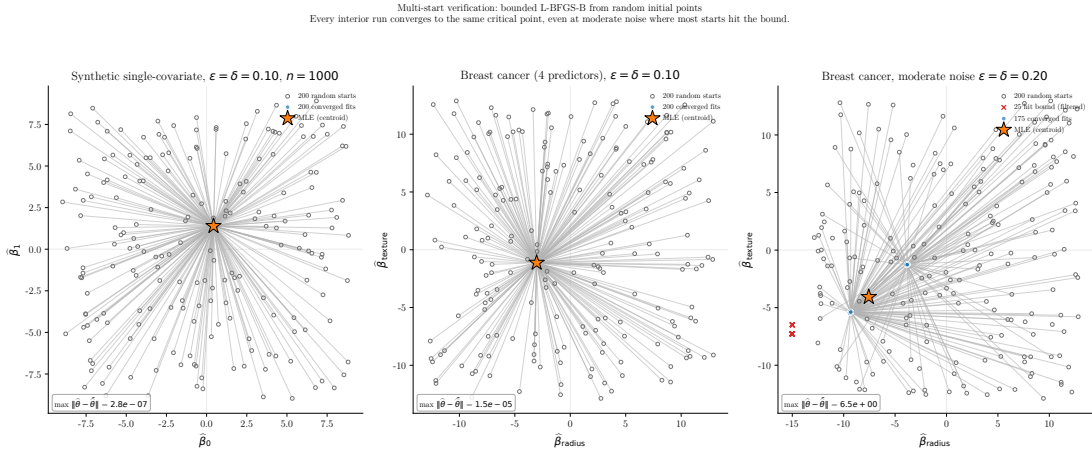
The score is

$$\nabla \ell_n^*(\beta) = \sum_i c \frac{p_i(1 - p_i)}{q_i(1 - q_i)} \{\hat{y}_i - q_i(\beta)\} x_i.$$

We fit β by maximizing ℓ_n^* . In code this means minimizing $-\ell_n^*$ with bounded L-BFGS-B, using the score above as the gradient.

Numerical optimizer: L-BFGS-B

The corrected objective has no closed-form maximizer and is not globally concave once label error is present. We optimize on a compact box $B_M = [-M, M]^p$. An interior solution is accepted; a boundary hit warns that the likelihood is flat, near the pole, or affected by quasi-separation. We track $\hat{p}_{\text{conv}}$, the fraction of Monte-Carlo fits whose coordinates remain inside the box.



Optimizer diagnostic, not label Monte Carlo: $B = 200$ random starts in the bounded box. Low-noise starts converge to one point; at $\varepsilon = \delta = 0.20$ on breast-cancer data, 175 interior runs split 119/56 between two local maxima.

No covariate: bias and variance

The no-covariate model makes the bias–variance trade-off explicit. If $y_i \sim \text{Bern}(p)$ and $\hat{y}_i \sim \text{Bern}(p^*)$, $p^* = \delta + cp$, then

$$\hat{p} = \frac{\bar{y} - \delta}{1 - \varepsilon - \delta}, \quad \mathbb{E} \hat{p} = p, \quad \text{Var} \hat{p} = \frac{1}{n} \frac{p^*(1 - p^*)}{c^2}.$$

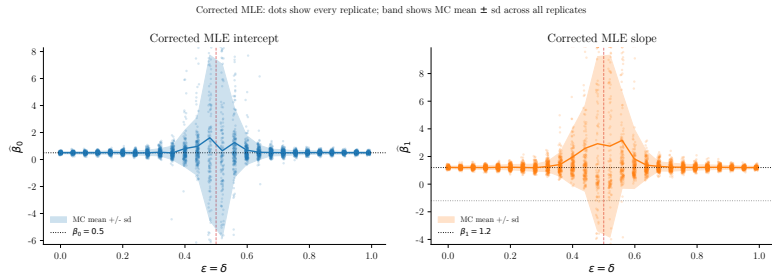
The correction is unbiased, but pays variance factor $p^*(1 - p^*)/[c^2 p(1 - p)]$ relative to clean data, reducing to $1/c^2$ at $p = 1/2$. In Monte Carlo checks ($p = 0.3$, $n = 1000$, $B = 4000$), the mean stays at 0.300 and Wald coverage stays near 0.95 away from the pole.

References

J. A. Hausman, J. Abrevaya, and F. M. Scott-Morton (1998). *Misclassification of the dependent variable in a discrete-response setting*, Journal of Econometrics.
R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu (1995). *A limited memory algorithm for bound constrained optimization*, SIAM J. Sci. Comput.
W. H. Wolberg, O. L. Mangasarian, and N. Street (1993). *Breast Cancer Wisconsin (Diagnostic)*, UCI Machine Learning Repository.
H. Hofmann (1994). *Statlog (German Credit Data)*, UCI Machine Learning Repository.
N. L. Hjort (2025). *Logistic Regression with Errors in Outcome Classification*, University of Oslo.

Corrected fit: adjusted log regression3

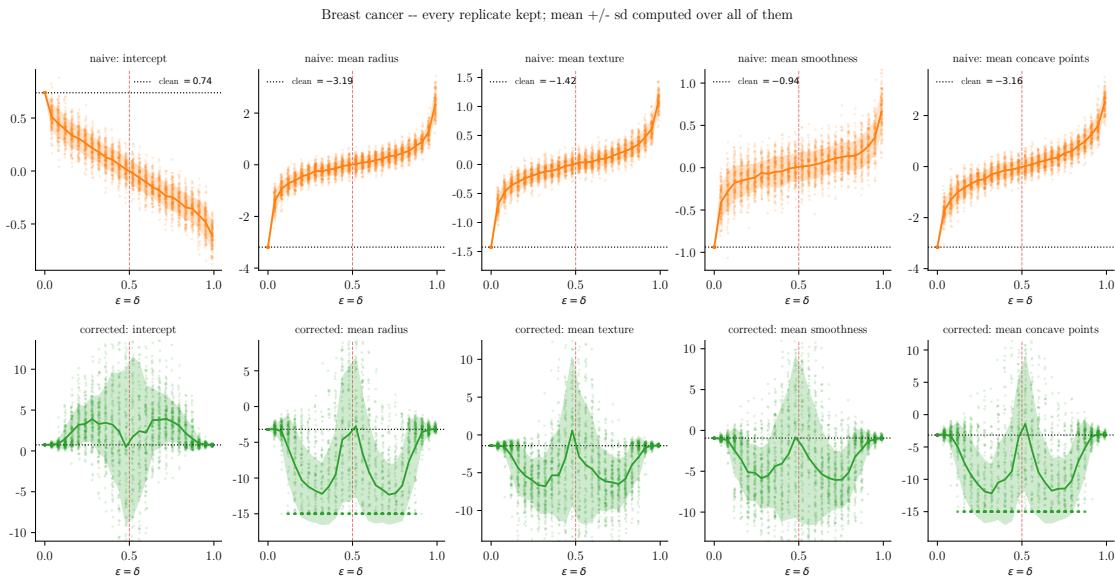
Reintroduce one covariate: $x_i \sim N(0, 1)$, $y_i \mid x_i \sim \text{Bern}\{H(\beta_0 + \beta_1 x_i)\}$ with $(\beta_0, \beta_1) = (0.5, 1.2)$, and flip labels symmetrically. We maximize ℓ_n^* with $|\beta_j| \leq 10$, warm-starting past the pole at $-\hat{\beta}_{\text{naive}}$. The corrected MLE is centered near the true coefficients on both sides of the pole; the visible price is the expanding Monte-Carlo spread.



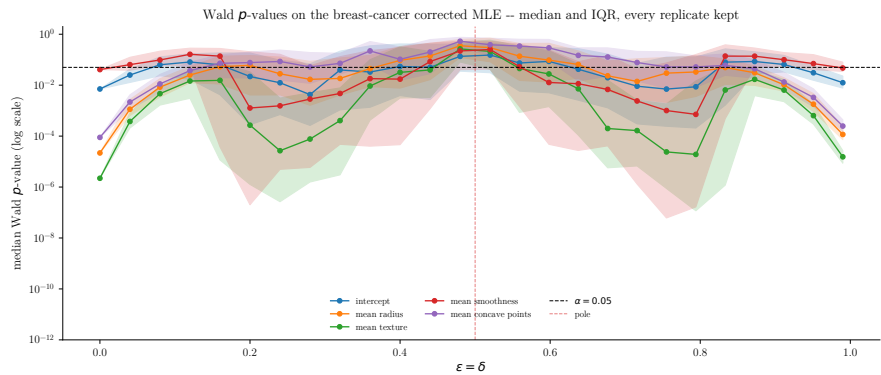
Same Monte Carlo draws as the naive fit, $B = 200$ per noise level. Dots: corrected MLEs. Solid and shaded band: MC mean ± 1 sd. The mean stays near (β_0, β_1) while variance grows near the pole.

Real data: Breast cancer

As a semi-synthetic stress test we use the Wisconsin breast-cancer data: 569 labelled fine-needle aspirates, a stratified 398/171 split, and four standardized predictors (mean radius, texture, smoothness, concave points). The clean-label fit is treated as the target, and we artificially flip labels at known rates. At $\varepsilon = \delta = 0.10$, the naive refit strongly attenuates radius, texture, and concave points; the corrected refit moves back toward the clean baseline, but with standard errors about 2–4 times wider. On this near-separating design, moderate noise also creates boundary hits and multiple local maxima. Around $\varepsilon \approx 0.25$, Wald tests lose power even while interior point estimates still cluster near the clean target.



Monte Carlo label flips on the fixed breast-cancer training set. Top: naive GLM paths. Bottom: corrected MLE paths. Solid: MC mean; shaded: MC mean ± 1 sd; dashed horizontal: clean-label baseline.



Wald tests over Monte Carlo label flips. Curves show median p -values over interior corrected fits; bands are IQRs. Strong predictors cross $\alpha = 0.05$ around $\varepsilon \approx 0.25$.

Covariate-dependent noise

Each observation may have its own label-error rates, $\varepsilon_i = \varepsilon(x_i)$ and $\delta_i = \delta(x_i)$, as when diagnostic accuracy or credit miscoding depends on measured features. The observed probability is still

$$q_i(\beta) = \delta_i + a_i p_i(\beta), \quad a_i = 1 - \varepsilon_i - \delta_i,$$

so only observations with $a_i \neq 0$ identify β . The important difference is practical: using one average error rate corrects the marginal amount of label noise, but not the way noise is tied to the predictors. If duration drives miscoding, the remaining bias concentrates on the duration coefficient and its standard error is too optimistic.

In German Credit we flip labels with a duration-dependent false-negative rate, $\varepsilon_i = 0.02 + 0.20H\{(\text{duration}_i - \bar{d})/s_d\}$, and keep $\delta_i = 0.05$. The duration coefficient moves 0.187 (naive) $\rightarrow$ 0.219 (constant $\bar{\varepsilon}$) $\rightarrow$ 0.301 (per- i), against clean target 0.316. A small audit can therefore help only if it reveals which features drive the error rates and gives a well-conditioned clean sub-design.

