

Logistic Regression

Aayush Bajaj

August 31, 2026

logistic regression is the method that seems only ever to be used for *classification* yet insists on calling itself regression. the resolution: it **is** regression — of the log-odds of a bernoulli success probability onto a linear predictor. this page develops it the honest way, as a generalised linear model: bernoulli response, canonical logit link, likelihood fitted by fisher scoring, inference through the deviance. the machine-learning reading (cross-entropy loss, linear decision boundaries) falls out at the end as a corollary.

*this page follows my math5806
(applied regression analysis)
week 4 notes, which follow
dobson and barnett*

Contents

1 motivation	1
2 the glm view	2
2.1 bernoulli in canonical form	2
2.2 the canonical link is the logit	2
3 the likelihood	3
4 fitting: fisher scoring is iterated weighted least squares	4
4.1 irls from scratch	5
5 inference	5
5.1 deviance	5
5.2 wald vs likelihood ratio	7
5.3 model comparison	7
6 interpretation: odds ratios	7
7 diagnostics	8
7.1 residuals	8
7.2 separation	8
7.3 overdispersion and link checks	8
8 the machine-learning view	8
9 extensions: more than two classes	9
10 notebook	9
11 see also	9

1 motivation

why not just regress the 0/1 labels with least squares? take the **Default** data from the **ISLR** package (the running example in james et al.'s *an introduction to statistical*

learning): predict credit-card default from account balance. the linear fit $\hat{y} = \mathbf{x}^\top \beta$ fails on three counts:

- the response is **binary**, not continuous — a straight line happily predicts probabilities below 0 (small balances) and above 1 (large ones).
- the response is **bounded** in $[0, 1]$; a line is not.
- the variance is **not constant**: $\text{Var}(Y_i) = \pi_i(1 - \pi_i)$ depends on the mean, so the homoscedastic gauss–markov story collapses.

the glm fix is not to bend the data to the line but to bend the line to the data: keep the linear predictor $\eta_i = \mathbf{x}_i^\top \beta$, and connect it to the mean through a **link function** $g(\mu_i) = \eta_i$ chosen to respect the geometry of the response.

2 the glm view

2.1 bernoulli in canonical form

define $Y_i = 1$ for “success” (probability π_i) and $Y_i = 0$ for “failure” (probability $1 - \pi_i$). a glm requires the response distribution to sit in the exponential family, written in the canonical form

*this is dobson and barnett’s
parametrisation of the
exponential family, which
math5806 uses throughout*

$$f(y_i | \theta_i) = \exp[y_i b(\theta_i) + c(\theta_i) + d(y_i)]. \quad (1)$$

the bernoulli pmf factors straight into this shape:

$$f(y_i | \pi_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i} = \exp \left[y_i \log \left(\frac{\pi_i}{1 - \pi_i} \right) + \log(1 - \pi_i) \right],$$

so $b(\pi) = \log \frac{\pi}{1-\pi}$, $c(\pi) = \log(1 - \pi)$, $d(y) = 0$. the general exponential-family identities $\mathbb{E}(Y) = -c'(\theta)/b'(\theta)$ and $\text{Var}(Y) = [b''(\theta)c'(\theta) - c''(\theta)b'(\theta)]/[b'(\theta)]^3$ then hand back $\mathbb{E}(Y_i) = \pi_i$ and $\text{Var}(Y_i) = \pi_i(1 - \pi_i)$ after a line of algebra.¹

2.2 the canonical link is the logit

the coefficient multiplying y_i — the **natural parameter** $b(\pi_i)$ — is exactly the log-odds. the canonical link is the choice $g = b$, i.e.

$$g(\pi_i) = \text{logit}(\pi_i) = \log \left(\frac{\pi_i}{1 - \pi_i} \right) = \mathbf{x}_i^\top \beta, \quad (2)$$

so the linear predictor lives on the log-odds scale: unbounded in both directions, exactly where a linear function is comfortable. inverting is a two-line rearrangement:

$$\frac{\pi_i}{1 - \pi_i} = \exp[\mathbf{x}_i^\top \beta] \implies \pi_i = \frac{\exp[\mathbf{x}_i^\top \beta]}{1 + \exp[\mathbf{x}_i^\top \beta]},$$

¹with $\theta = \pi$: $b'(\pi) = \frac{1}{\pi(1-\pi)}$ and $c'(\pi) = -\frac{1}{1-\pi}$, so $-c'/b' = \frac{\pi(1-\pi)}{1-\pi} = \pi$. the variance identity likewise reduces to $\pi(1 - \pi)$ after differentiating once more.

the **sigmoid** (inverse-logit). it maps the whole real line into $(0, 1)$, is symmetric about $\pi = \frac{1}{2}$, and steepest there.

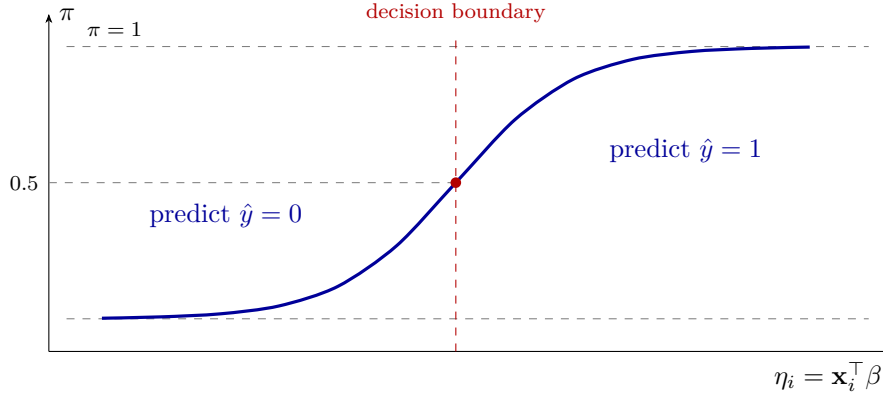


Figure 1: the sigmoid: probability against the linear predictor. the dashed guides at 0 and 1 are the bounds a straight-line fit ignores; the vertical line is the 0.5 decision boundary.

the whole glm can be drawn as a pipeline — the only moving part relative to ordinary least squares is the middle box:

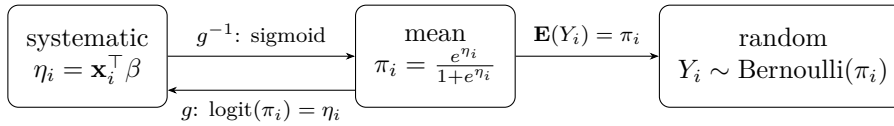


Figure 2: anatomy of the model: a linear systematic component, a link bridging scales, and a bernoulli random component.

3 the likelihood

for independent Y_i , the joint likelihood is

$$f(y_1, \dots, y_N \mid \boldsymbol{\pi}) = \prod_{i=1}^N \pi_i^{y_i} (1 - \pi_i)^{1-y_i} = \exp \left[\sum_{i=1}^N y_i \log \left(\frac{\pi_i}{1 - \pi_i} \right) + \sum_{i=1}^N \log(1 - \pi_i) \right], \quad (3)$$

and substituting the sigmoid gives the log-likelihood as a function of β :

$$\ell(\beta; \mathbf{y}) = \sum_{i=1}^N \left[y_i \mathbf{x}_i^\top \beta - \log(1 + \exp[\mathbf{x}_i^\top \beta]) \right]. \quad (4)$$

differentiating via the chain rule $\frac{d\ell_i}{d\beta_j} = \frac{d\ell_i}{d\theta_i} \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\beta_j}$ yields the **score function** of a general glm,

$$U_j = \sum_{i=1}^N \frac{(y_i - \mu_i) x_{ij}}{\text{Var}(Y_i)} \left(\frac{d\mu_i}{d\eta_i} \right), \quad (5)$$

and here the canonical link earns its name: for the logit, $\frac{d\mu_i}{d\eta_i} = \pi_i(1 - \pi_i) = \text{Var}(Y_i)$, so the ratio cancels and the score equations collapse to

$$\mathbf{U} = \mathbf{X}^\top (\mathbf{y} - \boldsymbol{\pi}(\boldsymbol{\beta})) = \mathbf{0}. \quad (6)$$

these look like the normal equations of least squares, but $\boldsymbol{\pi}$ is nonlinear in $\boldsymbol{\beta}$: **there is no closed form**. (contrast ordinary least squares, where the analogous system is linear and solves in one shot.) the same estimation machinery carries over unchanged when $Y_i \sim \text{Bin}(n_i, \pi_i)$ is binomial rather than bernoulli — only the number of trials enters the bookkeeping.

4 fitting: fisher scoring is iterated weighted least squares

newton-flavoured ascent replaces the intractable equations with a sequence of linear ones. the **method of scoring** updates

$$\hat{\boldsymbol{\beta}}^{(m)} = \hat{\boldsymbol{\beta}}^{(m-1)} + [\mathcal{I}^{(m-1)}]^{-1} \mathbf{u}^{(m-1)}, \quad (7)$$

where $\mathcal{I} = \mathbb{E}[\mathbf{U}\mathbf{U}^\top]$ is the fisher information. because $\mathbb{E}[(Y_i - \mu_i)(Y_l - \mu_l)] = 0$ for $i \neq l$, the information reduces to

$$\mathcal{I}_{jk} = \sum_{i=1}^N \frac{x_{ij}x_{ik}}{\text{Var}(Y_i)} \left(\frac{d\mu_i}{d\eta_i} \right)^2 \quad \Longleftrightarrow \quad \mathcal{I} = \mathbf{X}^\top \mathbf{W} \mathbf{X}, \quad w_{ii} = \frac{1}{\text{Var}(Y_i)} \left(\frac{d\mu_i}{d\eta_i} \right)^2. \quad (8)$$

multiply the scoring update through by $\mathcal{I}^{(m-1)}$ and the right-hand side rearranges into $\mathbf{X}^\top \mathbf{W} \mathbf{z}$ with the **working response**

$$z_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}^{(m-1)} + (y_i - \mu_i) \left(\frac{d\eta_i}{d\mu_i} \right), \quad (9)$$

so each iteration solves a weighted least-squares problem,

$$\mathbf{X}^\top \mathbf{W}^{(m-1)} \mathbf{X} \hat{\boldsymbol{\beta}}^{(m)} = \mathbf{X}^\top \mathbf{W}^{(m-1)} \mathbf{z}^{(m-1)}, \quad (10)$$

with $\mathbf{W}$ and $\mathbf{z}$ recomputed at each step — **iteratively reweighted least squares** (irls). the working response is a first-order linearisation of $g(y_i)$ about the current mean; the weights downweight observations whose response is noisy on the linear-predictor scale (Hastie, Trevor and Tibshirani, Robert and Friedman, Jerome, 2009).

specialising to the logit link (use $\frac{d\mu_i}{d\eta_i} = \pi_i(1 - \pi_i)$ and $\text{Var}(Y_i) = \pi_i(1 - \pi_i)$):

- **weights** $w_{ii} = \pi_i(1 - \pi_i)$ — maximal at $\pi_i = \frac{1}{2}$, vanishing as fits saturate towards 0 or 1;
- **working response** $z_i = \eta_i + \frac{y_i - \pi_i}{\pi_i(1 - \pi_i)}$;
- **canonical bonus**: with the canonical link the natural parameter *is* the linear predictor, so the hessian of ℓ is non-random and equals $-\mathbf{X}^\top \mathbf{W} \mathbf{X}$ exactly — newton–raphson and fisher scoring are the *same* algorithm.

this is why R's summary.glm reports 'number of fisher scoring iterations'; for non-canonical links the two schemes genuinely differ

4.1 irls from scratch

the course's running dataset is `trade.union` from the `SemiPar` R package: 534 workers from the 1985 us current population survey, binary union membership regressed on wage and age. twenty lines of R reproduce `glm()` to machine precision:

```
1 library(SemiPar)
2 data("trade.union")
3
4 y <- trade.union$union.member
5 X <- cbind(1, trade.union$wage, trade.union$age) # design matrix, N x 3
6
7 irls <- function(X, y, tol = 1e-10, maxit = 25) {
8   beta <- rep(0, ncol(X)) # start at pi = 0.5
9   for (m in 1:maxit) {
10    eta <- drop(X %*% beta) # linear predictor
11    pi <- 1 / (1 + exp(-eta)) # inverse logit
12    w <- pi * (1 - pi) # working weights (canonical
13    <- link)
14    z <- eta + (y - pi) / w # working response
15    beta.new <- solve(t(X) %*% (w * X), t(X) %*% (w * z))
16    if (max(abs(beta.new - beta)) < tol) { beta <- beta.new; break }
17    beta <- beta.new
18   }
19   se <- sqrt(diag(solve(t(X) %*% (w * X)))) # inverse Fisher information
20   list(beta = drop(beta), se = se, iter = m)
21 }
22
23 fit <- irls(X, y)
24 glm.fit <- glm(union.member ~ wage + age, data = trade.union, family =
25   <- binomial)
26
27 cat("irls beta:", format(fit$beta, digits = 7), " iters:", fit$iter, "\n")
28 cat("glm() beta:", format(coef(glm.fit), digits = 7), "\n")
29 cat("irls se: ", format(fit$se, digits = 7), "\n")
30 cat("glm() se: ", format(sqrt(diag(vcov(glm.fit))), digits = 7), "\n")
31
32 pi.hat <- 1 / (1 + exp(-drop(X %*% fit$beta)))
33 dev <- -2 * sum(y * log(pi.hat) + (1 - y) * log(1 - pi.hat))
34 cat("irls deviance:", format(dev, digits = 7), " glm deviance:",
35   <- format(deviance(glm.fit), digits = 7), "\n")
```

```
irls beta: -2.97609519  0.06516919  0.02186120  iters: 6
glm() beta: -2.97609519  0.06516919  0.02186120
irls se:  0.426686048 0.020117231 0.009722013
glm() se:  0.42668511 0.02011721 0.00972200
irls deviance: 485.5239  glm deviance: 485.5239
```

six iterations from a cold start; the standard errors are the square roots of the diagonal of $\mathcal{I}^{-1} = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1}$, the asymptotic covariance of the mle.

5 inference

5.1 deviance

the **deviance** compares the fitted model against the saturated one (a parameter per observation). for the binomial model $Y_i \sim \text{Bin}(n_i, \pi_i)$,

$$D = 2 \sum_{i=1}^N \left[y_i \log \left(\frac{y_i}{n_i \hat{\pi}_i} \right) + (n_i - y_i) \log \left(\frac{n_i - y_i}{n_i - n_i \hat{\pi}_i} \right) \right], \quad (11)$$

asymptotically equivalent to the pearson chi-squared statistic $P^2 = \sum_i (y_i - n_i \hat{\pi}_i)^2 / [n_i \hat{\pi}_i (1 - \hat{\pi}_i)]$, and under the null that the model is correct, $D \sim \chi^2(N - p)$ approximately.

the caveat that matters: this χ^2 result leans on the counts n_i being reasonably large per covariate pattern. with **ungrouped binary data** ($n_i = 1$, the usual case with continuous covariates) the approximation fails outright — the saturated model's parameter count grows with N , and D becomes a function of the fitted values alone, useless as an absolute goodness-of-fit measure. two escapes:

the lecture's version: both D and the pearson statistic are 'influenced by small frequencies, typical of continuous covariates'; there is some evidence the pearson statistic is the better behaved of the two

- **grouped fits:** pool observations sharing (or binned into) covariate patterns and fit the binomial model to (y_i, n_i) pairs — the **hosmer–lemeshow** idea. on the trade-union data, grouping by distinct wage leaves the estimates and standard errors untouched but changes the deviance bookkeeping entirely (residual deviance 239.92 on 236 df, versus 490.50 on 532 ungrouped) — same model, different goodness-of-fit scale.
- **differences of deviances:** for nested models the offending saturated terms cancel, and $D_0 - D_1 \sim \chi^2(p_1 - p_0)$ is trustworthy even for binary data. this is the workhorse.

worked on the trade-union fit — test $H_0 : \beta_{\text{wage}} = \beta_{\text{age}} = 0$ against the intercept-only model:

```
1 suppressMessages(library(SemiPar)); data("trade.union")
2 fit <- glm(union.member ~ wage + age, data = trade.union, family = binomial)
3
4 # prediction at wage 6.5, age 56 (course example)
5 eta <- sum(coef(fit) * c(1, 6.5, 56))
6 cat("pi.new =", exp(eta) / (1 + exp(eta)), "\n")
7
8 # deviance (likelihood-ratio) test of beta1 = beta2 = 0
9 d0 <- fit$null.deviance; d1 <- deviance(fit)
10 cat("d0 - d1 =", d0 - d1, " vs  chi^2_{2,0.95} =", qchisq(0.95, df = 2), "\n")
11
12 # odds ratios and 95% wald intervals
13 or <- exp(cbind(estimate = coef(fit), confint.default(fit)))
14 print(round(or, 4))
```

```
pi.new = 0.209446
d0 - d1 = 17.56029 vs  chi^2_{2,0.95} = 5.991465
      estimate 2.5 % 97.5 %
(Intercept)  0.0510 0.0221 0.1177
wage         1.0673 1.0261 1.1103
age          1.0221 1.0028 1.0418
```

null deviance 503.08 drops to 485.52 when wage and age enter; the drop of 17.56 dwarfs the χ^2_2 critical value 5.99, so the covariates earn their keep. the predicted membership probability for a 56-year-old on a wage of 6.5 is 0.209 — below the 0.5 threshold, classify as non-member (though nothing is sacred about 0.5; a conservative screen might trigger at 0.1).

5.2 wald vs likelihood ratio

- **wald:** $z = \hat{\beta}_j / \text{se}(\hat{\beta}_j)$ is asymptotically standard normal — these are the z-values R prints in `summary()`. cheap (one fit), but computed from the curvature *at the mle*, and it degrades badly when the likelihood is asymmetric — pathologically so near separation, where the wald statistic can shrink to non-significance precisely because the effect is enormous (the hauck–donner effect).
- **likelihood ratio:** refit the null model and compare $C = 2[\ell(\hat{\pi}; \mathbf{y}) - \ell(\tilde{\pi}; \mathbf{y})] \sim \chi^2(p-1)$ against the minimal (intercept-only) model, or use deviance differences as above for any nested pair. dearer, but the asymptotics are more reliable and invariant to reparametrisation (Wasserman, Larry, 2010).
- **confidence intervals on odds ratios:** form the wald interval on the log-odds scale, $\hat{\beta}_j \pm z_{1-\alpha/2} \text{se}(\hat{\beta}_j)$, then exponentiate — never build a symmetric interval around $e^{\hat{\beta}_j}$ directly. the table above: a dollar of hourly wage multiplies the odds of membership by 1.067, 95% ci (1.026, 1.110).

5.3 model comparison

when models are not nested, likelihood-based scores with complexity penalties take over: $\text{AIC} = -2\ell(\hat{\pi}; \mathbf{y}) + 2p$ and $\text{BIC} = -2\ell(\hat{\pi}; \mathbf{y}) + p \log N$. on the trade-union fits, aic prefers wage + age (491.52) over wage alone (494.50) and the null (505.08), while the harsher bic prefers wage alone (503.06 vs 504.37) — a tidy reminder the two criteria answer different questions. the **pseudo- R^2** , $[\ell(\tilde{\pi}) - \ell(\hat{\pi})]/\ell(\tilde{\pi})$, reads as the proportional log-likelihood improvement over the minimal model (a humble 0.035 here) — but unlike the linear-regression R^2 it has no variance-explained interpretation and inflates with every added predictor.

6 interpretation: odds ratios

because the model is linear in log-odds, a one-unit increase in x_j adds β_j to the log-odds — equivalently, **multiplies the odds by e^{β_j}** , holding the other covariates fixed. from the fit above: $e^{0.0652} = 1.067$ per wage dollar, $e^{0.0219} = 1.022$ per year of age.

three traps the course flags:

- β_j is **not** the change in probability. the marginal effect $\partial\pi/\partial x_j = \beta_j \pi(1 - \pi)$ depends on where you sit on the sigmoid: largest at $\pi = \frac{1}{2}$, negligible in the saturated tails. only the *sign* transfers unconditionally: positive β_j means increasing x_j increases π , and that is all.
- an **odds ratio is not a relative risk**. if disease probability is 0.8 in one group and 0.2 in another, the odds ratio is $\frac{0.8/0.2}{0.2/0.8} = 16$ while the prevalence ratio (relative risk) is $0.8/0.2 = 4$. quoting an or as “16 times more likely” is wrong whenever the outcome is common.
- estimating prevalence ratios directly means a **log link** on a bernoulli response — and that model routinely fails to converge, because the log maps probabilities to the non-positive half-line while the linear predictor roams all of $\mathbb{R}$ (the trade-union data throw R’s “no valid set of coefficients” error on the spot). the logit’s unbounded range is not a nicety; it is what makes unconstrained optimisation of β legitimate.

7 diagnostics

7.1 residuals

raw residuals are useless when responses are 0/1; glms standardise by the model's own variance. for $Y_i \sim \text{Bin}(n_i, \pi_i)$:

- **pearson residuals** $P_i = \frac{y_i - n_i \hat{\pi}_i}{\sqrt{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}}$, with $\sum_i P_i^2 = P^2$; standardise by leverage as $P_i / \sqrt{1 - h_{ii}}$ with h_{ii} from the hat matrix of the final irls solve.
- **deviance residuals** $d_i = \text{sign}(y_i - n_i \hat{\pi}_i) \sqrt{2 \left[y_i \log \frac{y_i}{n_i \hat{\pi}_i} + (n_i - y_i) \log \frac{n_i - y_i}{n_i - n_i \hat{\pi}_i} \right]}$, with $\sum_i d_i^2 = D$.

plot them against each continuous covariate, against omitted candidates, in measurement order, and against the fitted probabilities. expect a rude surprise on binary data: with $y_i \in \{0, 1\}$ the residual at a given $\hat{\pi}_i$ can take only two values, so every plot degenerates into **two smooth curves** (positives above, negatives below) and eyeballing is hopeless. superimpose a scatterplot smoother — on the trade-union fit a smoothing spline through the pearson residuals reveals the mean is underestimated at middling predictions and overestimated at the extremes, invisible to the naked eye. residual plots are simply less informative for glms than for linear models; lean on the formal statistics too.

7.2 separation

if some hyperplane perfectly separates the two classes, the likelihood increases monotonically as $\|\beta\| \rightarrow \infty$ along the separating direction: **the mle does not exist**. symptoms: coefficients and standard errors ballooning across iterations, fitted probabilities pinned at 0 and 1, R muttering **fitted probabilities numerically 0 or 1 occurred**. it is most common in small samples with many covariates — precisely where you would like the model most. remedies: penalised likelihood (firth's bias-reduction, or the ridge/lasso route of regularised regression), a bayesian prior, or admitting the data cannot identify the parameter.

7.3 overdispersion and link checks

- with grouped data the variance can exceed the binomial $n_i \pi_i (1 - \pi_i)$ — flagged by D far above $N - p$; causes include omitted covariates and dependence between trials. the quasi-likelihood fix scales the variance by an extra parameter ϕ (**family = quasibinomial** in R). for ungrouped binary data, overdispersion is not even identifiable — another reason to prefer grouped fits when the design permits.
- is the logit even the right link? embed it in a parametric family such as $g(\pi, \alpha) = \log \left[\frac{(1-\pi)^{-\alpha} - 1}{\alpha} \right]$, which recovers the logit at $\alpha = 1$ and the log-log link $\log[-\log(1-\pi)]$ as $\alpha \rightarrow 0$; estimating α from the data tests the logit against its neighbours.

8 the machine-learning view

the ml literature rederives all of the above with different vocabulary; the dictionary is short:

- **cross-entropy is the negative bernoulli log-likelihood**. the “log loss”

$-\sum_i [y_i \log \hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i)]$ minimised by every deep-learning framework is exactly $-\ell(\beta; \mathbf{y})$, and its minimiser is the mle. see loss functions for the family portrait — this identity is also why the deviance is twice the gap between achieved and perfect log loss.

- **the decision boundary is linear.** classify $\hat{y} = 1$ when $\hat{\pi} > \frac{1}{2}$, i.e. when $\mathbf{x}^\top \beta > 0$: a hyperplane, just like the perceptron and the linear svm. the models differ not in the boundary's shape but in how they choose it — maximum likelihood on probabilities versus mistake-driven updates versus maximum margin — and logistic regression alone hands back calibrated probabilities with standard errors attached.
- **regularisation is penalised likelihood.** adding $\lambda \|\beta\|_2^2$ or $\lambda \|\beta\|_1$ to the negative log-likelihood gives ridge and lasso logistic regression (regularised regression); besides taming variance it cures separation, since the penalised objective always has a finite minimiser.

9 extensions: more than two classes

for a nominal response with $K > 2$ categories (the lecture's emergency-room triage: stroke vs drug overdose vs epileptic seizure — the labels 1, 2, 3 carry no order, which is exactly why regressing on them linearly is nonsense), the model becomes **multinomial logistic regression**: pick a baseline category and fit $K - 1$ simultaneous logits, $\log(\pi_{ik}/\pi_{iK}) = \mathbf{x}_i^\top \beta_k$ — softmax, to the ml crowd. when the categories *are* ordered (disease severity, likert scales), the **proportional-odds model** is thriftier: model the cumulative logits $\text{logit } \Pr(Y_i \leq k) = \alpha_k - \mathbf{x}_i^\top \beta$ with one shared slope vector and $K - 1$ thresholds, so a covariate shifts the whole ordinal scale coherently. both inherit the entire glm toolkit above — exponential family, irls, deviance — with more bookkeeping. dobson & barnett's *an introduction to generalized linear models* (the math5806 text) develops both in chapter 8.

10 notebook

an executed ml-flavoured companion (numpy gradient descent on the log loss, decision surfaces):

11 see also

- generalised linear models — the exponential-family engine this page instantiates
- estimation and inference — score, information, and why wald and likelihood-ratio tests can disagree
- ordinary least squares — the glm you get with an identity link and gaussian noise
- regularised regression — penalised likelihood, and the cure for separation
- loss functions — cross-entropy and friends
- support vector machines — same linear boundary, margin-based objective
- the perceptron — same boundary again, mistake-driven fit

references

Hastie, Trevor and Tibshirani, Robert and Friedman, Jerome (2009). *The Elements of Statistical Learning*, Springer.

Wasserman, Larry (2010). *All of Statistics: A Concise Course in Statistical Inference*, Springer.