Logistic regression

What if the response Y is categorical?

Still interested in discriminative models, i.e. models for $P\{Y|X\}$

Assume

Y	Probability
1	$P\{Y=1\} = \pi$
0	$P\{Y=0\} = 1-\pi$

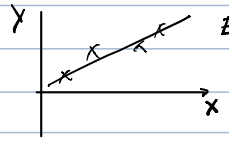
$Y \sim \text{Bernoulli}(\pi)$, $E\{Y\} = \pi$, $\text{Var}\{Y\} = \pi(1-\pi)$

Unlike Normal distribution, mean and variance depend on the same parameter

Question: how does $P\{Y=1\}$ depend on predictor X?

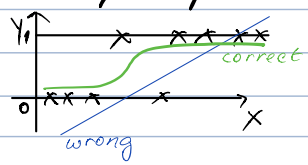
Specify hypothesis $h_\theta(x): X \to Y = \{0,1\}$

Continuous response:



$E\{Y|X\} = \underbrace{\theta_0 + \theta_1 X}_{h_\theta(x)}$

Binary response:



$\pi = E\{Y|X\} = \underbrace{g(\theta_0 + \theta_1 X)}_{h_\theta(x)}$ — g: mean response function or activation function

$g^{-1}(E\{Y|X\}) = \theta_0 + \theta_1 X$ — g^{-1}: link function

Good choice for $g(\cdot)$: a CDF of a probability distribution

$g(t)$ = CDF of $N(0,1)$ $\Rightarrow$ probit regression (no analytical expression)

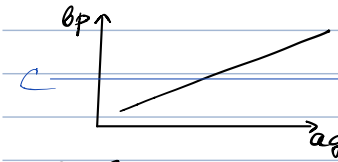
$g(t) = \frac{e^t}{1+e^t}$ = CDF of logistic distribution $\Rightarrow$ logistic regression

$$E\{Y|X\} = g(\underbrace{\theta_0 + \theta_1 X}_{\text{linear function}})$$

non-linear function $\Rightarrow$ not a linear model $\Rightarrow$ a generalized linear model $\Rightarrow$ no analytical solution for param. estimate (numeric solution only)

Motivation for probit regression: Latent variable

Example: a linear regression for blood pressure as function of age:



$bp = \theta_0 + \theta_1 age + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$

Only observe hypertension: $Y = \begin{cases} 1, & bp > c \\ 0, & bp < c \end{cases}$

$$P\{Y=1\} = P\{bp > c\} = P\{a + b\,age + \varepsilon > c\} = P\{\varepsilon < (a-c) + b\,age\}$$

$$= P\left\{\underbrace{\frac{\varepsilon}{\sigma}}_{\sim N(0,1)} < \underbrace{\frac{a-c}{\sigma}}_{\theta_0} + \underbrace{\frac{b}{\sigma}}_{\theta_1} age\right\} = \underbrace{\Phi}_{\text{CDF of } N(0,1)}(\theta_0 + \theta_1 age)$$

Logistic regression:

$$\pi = E\{Y|X\} = \frac{e^{\theta_0+\theta_1 X}}{1+e^{\theta_0+\theta_1 X}} = \frac{1}{1+e^{-(\theta_0+\theta_1 X)}}$$

- monotonic increasing if $\theta_1 > 0$
- monotonic decreasing if $\theta_1 < 0$

$$\log\frac{E\{Y|X\}}{1-E\{Y|X\}} = \log\frac{\pi}{1-\pi} = \theta_0 + \theta_1 X$$

$\log \frac{E\{Y|X=x+1\}/(1-E\{Y|X=x+1\})}{E\{Y|X=x\}/(1-E\{Y|X=x\})} = \log(\text{odds ratio})$

Interpretation of θ_1:

$$\log\underbrace{\frac{E\{Y|X=x+1\}}{1-E\{Y|X=x+1\}}}_{\text{Odds}(X=x+1)} - \log\underbrace{\frac{E\{Y|X=x\}}{1-E\{Y|X=x\}}}_{\text{Odds}(X=x)} = \theta_1$$

Parameter estimation: no error term; least squares not appropriate ⇒ Maximum Likelihood

$Y_i \overset{ind}{\sim} \text{Bernoulli}(\pi_i)$, where $\pi_i \overset{\text{model}}{=} \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$

Likelihood $L = \prod_{i=1}^{N} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$ (indep. · Bernoulli distribution)

$$\log L = \sum_{i=1}^{N} Y_i \log(\pi_i) + (1-Y_i)\log(1-\pi_i) \overset{\text{plug in } \pi_i}{=} \sum_{i=1}^{N}\left[Y_i \log \frac{1}{1+e^{-(\beta_0+\beta_1 X_i)}} + (1-Y_i)\log\left(1-\frac{1}{1+e^{(\beta_0+\beta_1 X_i)}}\right)\right]$$

$\frac{1}{1+e^{\beta_0+\beta_1 X}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$; convex

To maximize $\log L$, set $\nabla_\beta \log L = 0$. It helps to use chain rule:

Define $\pi(z) = \frac{1}{1+e^{-z}}$. $\pi'(z) = \frac{d}{dz}\frac{1}{1+e^{-z}} = \frac{1}{(1+e^{-z})^2}e^{-z} = \frac{1}{1+e^{-z}}\cdot\left(1-\frac{1}{1+e^{-z}}\right) = \pi(z)(1-\pi(z))$

In this notation, $\log L = \sum_{i=1}^{N} Y_i \log(\pi(\beta_0+\beta_1 X_i) + (1-Y_i)\log[1-\pi(\beta_0+\beta_1 X_i)]$

$$\frac{d\log L}{d\beta_1} = \frac{d\log L}{d\pi}\cdot\frac{d\pi}{dz}\cdot\frac{dz}{d\beta_1} = \sum_{i=1}^{N}\left[Y_i \frac{1}{\pi(\beta_0+\beta_1 X_i)} - (1-Y_i)\frac{1}{1-\pi(\beta_0+\beta_1 X_i)}\right]\cdot \pi(\beta_0+\beta_1 X_i)\left(1-\pi(\beta_0+\beta_1 X_i)\right)\cdot X_i$$

$$= \sum_{i=1}^{N}(Y_i - \pi(\beta_0+\beta_1 X_i))X_i$$

In multivariate case, batch gradient descent:

Initialize $\beta_{p\times 1}$

Repeat until convergence {

  For $j = 1, \dots, p$ {

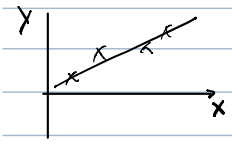
    $\beta_j \leftarrow \beta_j + \alpha \sum_{i=1}^{N}(Y_i - \pi(X_i^T\beta))X_{ij}$

  }

}

Practical comment: grouped data

Continuous response:

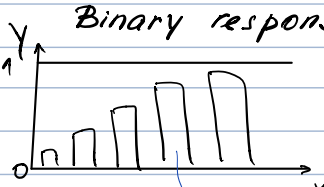


$Y = \beta_0+\beta_1 X+\varepsilon,\ \varepsilon \sim N(0,\sigma^2)$

If $\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \Rightarrow \text{Var}\{\bar{Y}_i\} = \frac{\sigma^2}{n_i}$

Weighted regression: same fit

Binary response:



proportion of $Y=1$ out of n_i obs with same X

Data: $(X_i, \bar{Y}_i, n_i)$

$Y_i \sim \text{Binomial}(n_i, \pi_i)$

$E\{\bar{Y}_i\} = \pi_i$

$\text{Var}\{\bar{Y}_i\} = \frac{\pi_i(1-\pi_i)}{n_i}$

Log-likelihood:

$$\log \prod_{i=1}^{N}\binom{n_i}{Y_i}\pi_i^{Y_i}(1-\pi_i)^{(n_i-Y_i)}$$

$$= \sum_{i=1}^{N}\log\binom{n_i}{Y_i} + \sum_{i=1}^{N}[Y_i\log\pi_i + (n_i-Y_i)\log(1-\pi_i)] = \text{const} + \sum_{i=1}^{N}\sum_{j=1}^{n_i}[Y_{ij}\log(\pi_i) + (1-Y_{ij})\log(1-\pi_i)]$$

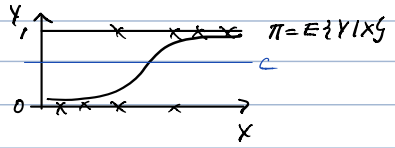
$\log\binom{n_i}{Y_i}$: does not depend on π_i, can ignore

$Y_i = \sum_{j=1}^{n_i} Y_{ij}$, where $Y_{ij} = \begin{cases}1, & \pi_i \\ 0, & 1-\pi_i\end{cases}$

$\sum_{j=1}^{n_i}[Y_{ij}\log(\pi_i) + (1-Y_{ij})\log(1-\pi_i)]$: Log-likelihood of individual data

⇒ individual and grouped data give same $\hat{\beta}$ if we know n_i

Logistic regression for classification



Classify $\hat{y} = \begin{cases} 1, & \text{if } \widehat{E\{Y|X\}} \geq c \\ 0, & \text{if } \widehat{E\{Y|X\}} < c \end{cases}$

- this is the reason for the name "discriminative model"

Summary of classification:

		Decision 0	Decision 1	Total
Truth	0	TN	FP	N
	1	FN	TP	P

counts of observations
confusion matrix

Prevalence: $\frac{P}{P+N}$

False positive rate $= \frac{FP}{N}$

Sensitivity = recall = true positive rate $= \frac{TP}{P}$

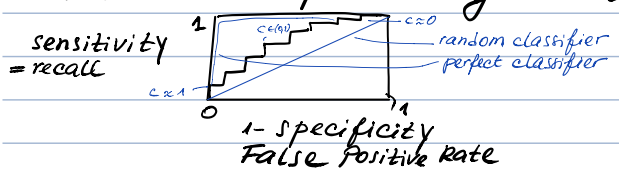
Specificity = selectivity = true negative rate $= \frac{TN}{N}$

Precision = positive predictive value $= \frac{TP}{TP+FP}$

False positive rate $= \frac{FP}{N}$

Accuracy $= \frac{TP+TN}{P+N}$

Receiver operating characteristic (ROC) curve



- Area under the ROC curve summarizes the predictive ability over all cutoffs
- More optimistic (i.e., higher) on the training set than on the validation set

Decision boundary

The classification cutoff separates points by a hyperplane:

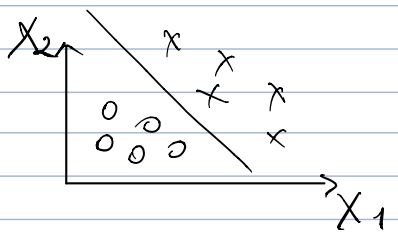
$$\hat{Y} = 1 \text{ if } \hat{\pi} = \frac{1}{1+e^{-(\delta_0+\delta_1 X)}} \geq \frac{1}{2} \iff 1 + e^{-(\delta_0+\delta_1 X)} \leq 2$$

$$e^{-(\delta_0+\delta_1 x)} \leq 1$$

$$-(\delta_0 + \delta_1 x) \leq 0$$

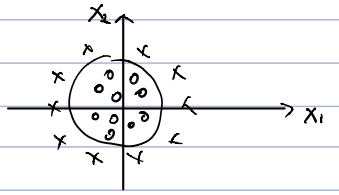
$$\delta_0 + \delta_1 x \geq 0$$

decision boundary is a hyperplane →



$$Y = 1 \text{ if } \hat{\pi} = \frac{1}{1+e^{-(3-X_1+X_2)}} \geq \frac{1}{2}$$

$$\Rightarrow 3 - X_1 + X_2 \geq 0 \Rightarrow X_2 \leq 3 - 1 \cdot X_1$$



Higher-order terms can express complex decision boundaries

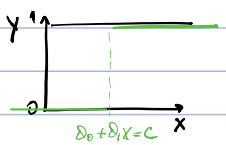
$$\frac{1}{1+e^{-(-1+0\cdot X_1+0\cdot X_2+1\cdot X_1^2+1\cdot X_2^2)}} \geq 0.5 \iff X_1^2 + X_2^2 \geq 1$$

Comments:

- Linear separability. MLE does not have a unique solution
  $\text{var}(\delta) \to \infty$
- Alternative ways to derive linear boundary:

$$f(X) = \begin{cases} 1, & \text{if } X_i^T \delta \geq c \\ 0, & \text{if } X_i^T \delta < c \end{cases}$$

{ $g(x)$: step function



- the perceptron
"firing of neurons"

No probabilistic interpretation; no MLE

- **Variable selection**: no residuals → no R^2, no MSE

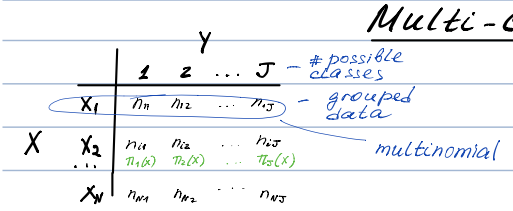
$AIC = -2\log L + 2p$
$BIC = -2\log l + \log N \cdot p$
} find subsets of predictors that minimize AIC or BIC

train / dev / test sets or cross-validation - as in linear regression

- **Regularization**

maximize $\log L - \lambda \|\beta\|^2$ → other penalties/algos as in regression

Multi-class classification



$\pi_j(x) = P\{Y = j \mid X = x\}$, $\sum_{j=1}^{J} \pi_j(x) = 1$

Multinomial logistic regression:

$$\log \frac{\pi_j(x)}{\pi_1(x)} = \alpha_j + \beta_j X, \quad j = 2, \dots, J$$

arbitrary category as baseline, to satisfy $\sum_{j=1}^{J} \pi_j = 1$

separate params for each category

} useful if interested in param. interpretation

Solving for π_j:

$$\pi_1(x) = \frac{1}{1 + \sum_{j=1}^{J} e^{\alpha_j + \beta_j X}}, \quad \pi_j = \frac{e^{\alpha_j + \beta_j X}}{1 + \sum_{j=1}^{J} e^{\alpha_j + \beta_j X}}, \; j = 2, \dots, J$$

MLE: $\log \text{Likelihood}_i = \log \left[\prod_{j=1}^{J} \pi_j(x)^{y_{ij}} \right] = \sum_{j=2}^{J} y_{ij} \log \pi_j(x) + \left(1 - \sum_{j=2}^{J} y_{ij}\right) \log \pi_1(x_i)$

for one data point, to simplify

Alternative parametrization: softmax regression

$$\pi_j(x) = \frac{e^{\alpha_j + \beta_j X}}{\sum_{j=1}^{J} e^{\alpha_j + \beta_j X}}$$

} more convenient if goal is prediction

- only $J-1$ free parameter sets
- set $\alpha_1 = \beta_1 = 0$ to have the previous case