Logistic Regression

Statistical Machine Learning

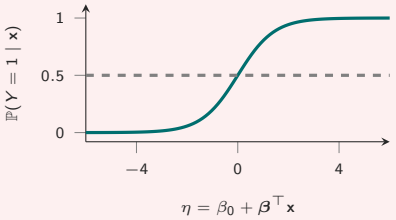
A. Sharma

Lecture 4

By the end of this session you will be able to:

- explain *why* linear regression fails for a yes/no response,
- state the logistic model and read β_j on the odds scale,
- describe how the coefficients are estimated, and why the problem is well behaved.

A knight meets a dragon. Does the knight survive?



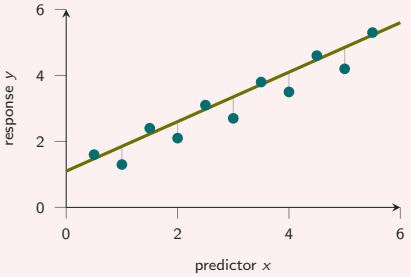
The logistic (sigmoid) curve

Recall: linear regression

The setup you already know. A *continuous* response $y \in \mathbb{R}$ from predictors $\mathbf{x} \in \mathbb{R}^d$:

$$y = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Here $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$.
Equivalently $\mathbb{E}[y \mid \mathbf{x}] = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}$, i.e. the mean is linear in the predictors.



Least squares minimises the squared vertical residuals. The response is continuous

Recall: linear regression

The setup you already know. A *continuous* response $y \in \mathbb{R}$ from predictors $\mathbf{x} \in \mathbb{R}^d$:

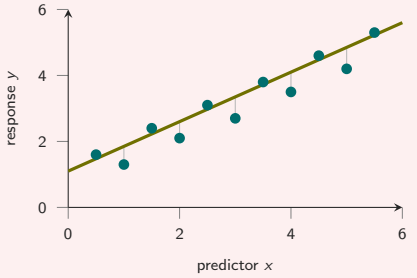
$$y = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Here $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$.
Equivalently $\mathbb{E}[y \mid \mathbf{x}] = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}$, i.e. the mean is linear in the predictors.

Fitting. Least squares,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_i (y_i - \beta_0 - \boldsymbol{\beta}^\top \mathbf{x}_i)^2 = (X^\top X)^{-1} X^\top \mathbf{y},$$

which is also the maximum-likelihood estimate under Gaussian noise.



Least squares minimises the squared vertical residuals. The response is continuous

Recall: linear regression

The setup you already know. A *continuous* response $y \in \mathbb{R}$ from predictors $\mathbf{x} \in \mathbb{R}^d$:

$$y = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

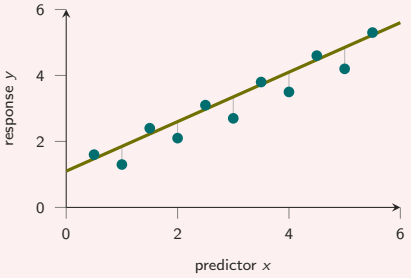
Here $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$.
Equivalently $\mathbb{E}[y \mid \mathbf{x}] = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}$, i.e. the mean is linear in the predictors.

Fitting. Least squares,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_i (y_i - \beta_0 - \boldsymbol{\beta}^\top \mathbf{x}_i)^2 = (X^\top X)^{-1} X^\top \mathbf{y},$$

which is also the maximum-likelihood estimate under Gaussian noise.

Reading it. β_j = change in expected y per unit of x_j, others fixed.



Least squares minimises the squared vertical residuals. The response is continuous

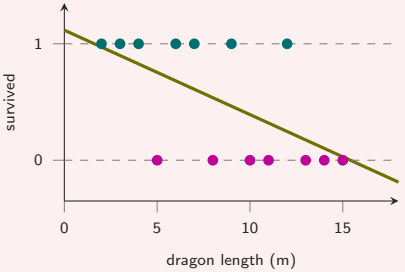
What if the response is not continuous but 0/1, yes/no, spam/no spam?

Why not just fit a straight line?

The problem. The response is a *label*, not a number:

$$Y \in \{0, 1\}, \qquad x \in \mathbb{R},$$

We want $p(x) = \mathbb{P}(Y = 1 \mid x)$.
Our data: 14 knights, each faced a dragon of known length and $Y = 1$ if the knight survived.



Ordinary least square is useless: a 0.5 m dragon gets $P > 1$, an 18 m dragon gets $P < 0$

Why not just fit a straight line?

The problem. The response is a *label*, not a number:

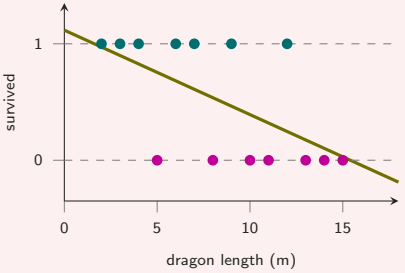
$$Y \in \{0, 1\}, \qquad x \in \mathbb{R},$$

We want $p(x) = \mathbb{P}(Y = 1 \mid x)$.
Our data: 14 knights, each faced a dragon of known length and $Y = 1$ if the knight survived.

Try least squares: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

- $\rightarrow$ fitted values leave $[0, 1]$, therefore not probabilities,
- $\rightarrow$ the errors cannot be normal, they take two values.

We need a model mapping a linear predictor into $(0, 1)$.



Ordinary least square is useless: a 0.5 m dragon gets $P > 1$, an 18 m dragon gets $P < 0$

The logistic model

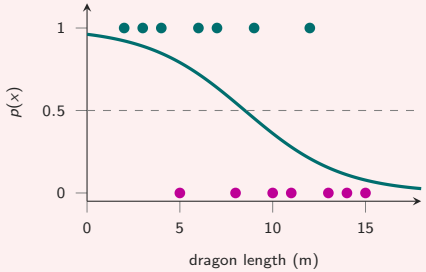
We will first develop ideas and model when predictor x is one dimensional, i.e. $x \in \mathbb{R}$.

The logistic model

The model. Keep the linear predictor, deform it:

$$p(x) = \sigma\big(\beta_0 + \beta_1 x\big), \qquad \sigma(z) = \frac{1}{1 + e^{-z}}.$$

Equivalently $Y \mid x \sim \text{Bernoulli}\big(p(x)\big)$.



Same data, logistic fit: bounded in $(0, 1)$, crossing 0.5 at 8.5 m

The logistic model

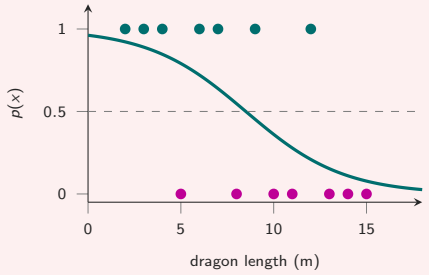
The model. Keep the linear predictor, deform it:

$$p(x) = \sigma\big(\beta_0 + \beta_1 x\big), \qquad \sigma(z) = \frac{1}{1 + e^{-z}}.$$

Equivalently $Y \mid x \sim \text{Bernoulli}\big(p(x)\big)$.

Bernoulli (RECALL)

Think of a single (biased or unbiased) coin flip:
$\mathbb{P}(Y = 1) = p$, $\mathbb{P}(Y = 0) = 1 - p$.
$\mathbb{E}[Y] = p$, and $\mathrm{Var}(Y) = p(1 - p)$.



Same data, logistic fit: bounded in $(0, 1)$, crossing 0.5 at 8.5 m

The logistic model

The model. Keep the linear predictor, deform it:

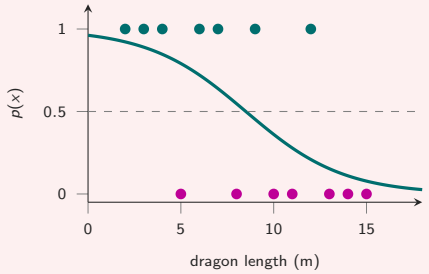
$$p(x) = \sigma\big(\beta_0 + \beta_1 x\big), \qquad \sigma(z) = \frac{1}{1 + e^{-z}}.$$

Equivalently $Y \mid x \sim \text{Bernoulli}\big(p(x)\big)$.

Useful properties of σ.

$$\sigma : \mathbb{R} \to (0,1), \quad \sigma(0) = \tfrac{1}{2}, \quad \sigma(-z) = 1 - \sigma(z),$$

$$\sigma'(z) = \sigma(z)\big(1 - \sigma(z)\big) > 0.$$



Same data, logistic fit: bounded in $(0, 1)$, crossing 0.5 at 8.5 m

The logistic model

The model. Keep the linear predictor, deform it:

$$p(x) = \sigma\big(\beta_0 + \beta_1 x\big), \qquad \sigma(z) = \frac{1}{1 + e^{-z}}.$$

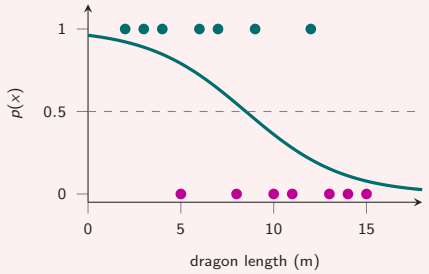
Equivalently $Y \mid x \sim \text{Bernoulli}\big(p(x)\big)$.

Useful properties of σ.

$$\sigma : \mathbb{R} \to (0,1), \quad \sigma(0) = \tfrac{1}{2}, \quad \sigma(-z) = 1 - \sigma(z),$$

$$\sigma'(z) = \sigma(z)\big(1 - \sigma(z)\big) > 0.$$

- → fitting our 14 knights gives $\hat\beta_0 = 3.23$, $\hat\beta_1 = -0.38$ (we will see later how to estimate these, DON'T WORRY!),
- → $\hat\beta_1 < 0$: survival falls with size, $|\hat\beta_1|$ sets the steepness (THINK!).



Same data, logistic fit: bounded in $(0, 1)$, crossing 0.5 at 8.5 m

The logistic model

So, the logistic regression model is

$$p(x) = \sigma\big(\beta_0 + \beta_1 x\big), \qquad \sigma(z) = \frac{1}{1 + e^{-z}}.$$

A straight line $\beta_0 + \beta_1 x$, squashed into a probability.

The logistic model

So, the logistic regression model is

$$p(x) = \sigma\big(\beta_0 + \beta_1 x\big), \qquad \sigma(z) = \frac{1}{1 + e^{-z}}.$$

A straight line $\beta_0 + \beta_1 x$, squashed into a probability.

Two questions remain:

What does β_1 *mean*?

How do we *find* β_0, β_1?

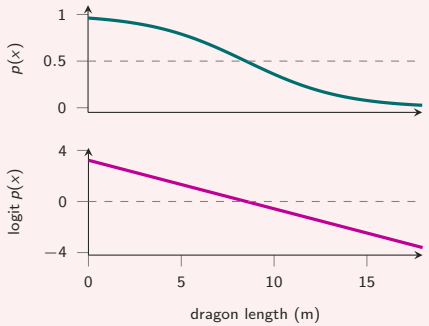
Reading the coefficients: odds and log-odds

Definition. The *odds* of survival,

$$\text{odds}(x) = \frac{p(x)}{1-p(x)} \in (0, \infty).$$

Using the *logit link*:

$$\text{logit}\big(p(x)\big) = \log \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x = 3.23 - 0.38\, x.$$



Curved on the probability scale, straight on the log-odds scale (slope $\hat{\beta}_1$)

Reading the coefficients: odds and log-odds

Definition. The *odds* of survival,

$$\text{odds}(x) = \frac{p(x)}{1-p(x)} \in (0, \infty).$$

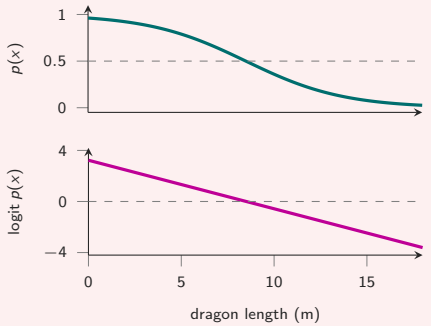
Using the *logit link*:

$$\text{logit}\big(p(x)\big) = \log \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x = 3.23 - 0.38\, x.$$

Interpretation. One extra metre of dragon, (keeping everything else fixed):

$$\frac{\text{odds}(x+1)}{\text{odds}(x)} = e^{\beta_1} \quad \text{(the } \textit{odds ratio}\text{)}.$$

Careful: e^{β_1} multiplies the *odds*, not the probability. The change in p is largest near $p = 0.5$.



Curved on the probability scale, straight on the log-odds scale (slope $\hat{\beta}_1$)

Reading the coefficients: odds and log-odds

Definition. The *odds* of survival,

$$\text{odds}(x) = \frac{p(x)}{1-p(x)} \in (0, \infty).$$

Using the *logit link*:

$$\text{logit}\big(p(x)\big) = \log \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x = 3.23 - 0.38\, x.$$

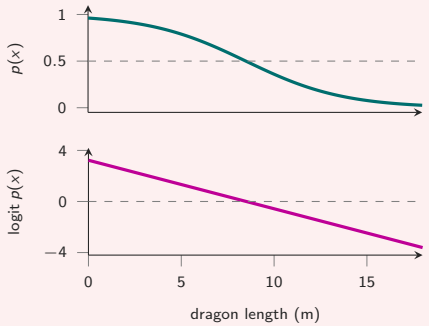
Interpretation. One extra metre of dragon, (keeping everything else fixed):

$$\frac{\text{odds}(x+1)}{\text{odds}(x)} = e^{\beta_1} \quad \text{(the } \textit{odds ratio}\text{)}.$$

Careful: e^{β_1} multiplies the *odds*, not the probability. The change in p is largest near $p = 0.5$.

Our fit

$\hat{\beta}_1 = -0.38 \Rightarrow e^{-0.38} \approx 0.68$: each extra metre of dragon cuts the knight's survival odds by about 32%.



Curved on the probability scale, straight on the log-odds scale (slope $\hat{\beta}_1$)

Fitting the model: maximum likelihood

Log-likelihood. With $p_i = \sigma(\beta_0 + \beta_1 x_i)$ and independent knights,

$$\ell(\beta) = \sum_{i=1}^{n} \Big[y_i \log p_i + (1 - y_i) \log(1 - p_i) \Big].$$

Minimising $-\ell$ is exactly the *cross-entropy* (log-loss) of machine learning.

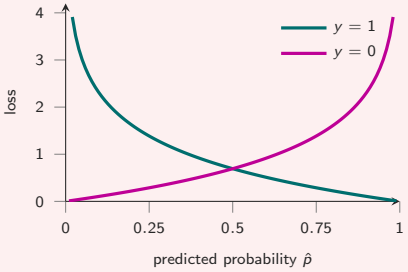
From likelihood to log-likelihood

Each knight is Bernoulli: $\mathbb{P}(Y_i{=}y_i) = p_i^{y_i}(1 - p_i)^{1-y_i}$.

Independence $\Rightarrow$ multiply: $L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}$.

Take log (product $\to$ sum), maximiser unchanged:

$$\ell = \log L = \sum_{i=1}^{n} \big[y_i \log p_i + (1 - y_i) \log(1 - p_i) \big].$$



Confident and wrong is punished without limit.

Fitting the model: maximum likelihood

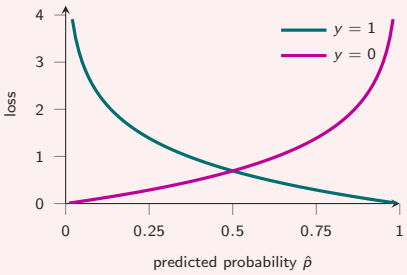
Log-likelihood. With $p_i = \sigma(\beta_0 + \beta_1 x_i)$ and independent knights,

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Big[y_i \log p_i + (1 - y_i) \log(1 - p_i) \Big].$$

Minimising $-\ell$ is exactly the *cross-entropy* (log-loss) of machine learning.

Score and Hessian. With X $(n \times 2)$ the design matrix,

$$\nabla \ell = X^\top (\mathbf{y} - \mathbf{p}), \qquad \nabla^2 \ell = -X^\top W X, \quad W = \operatorname{diag}\big(p_i(1-p_i)\big).$$



Confident and wrong is punished without limit.

From likelihood to log-likelihood

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i} \xrightarrow{\log}$$

$$\ell = \sum_{i=1}^{n} \big[y_i \log p_i + (1 - y_i) \log(1 - p_i) \big].$$

Fitting the model: maximum likelihood

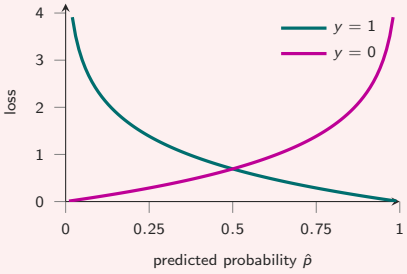
Log-likelihood. With $p_i = \sigma(\beta_0 + \beta_1 x_i)$ and independent knights,

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Big[y_i \log p_i + (1 - y_i) \log(1 - p_i) \Big].$$

Minimising $-\ell$ is exactly the *cross-entropy* (log-loss) of machine learning.

Score and Hessian. With X $(n \times 2)$ the design matrix,

$$\nabla \ell = X^\top (\mathbf{y} - \mathbf{p}), \qquad \nabla^2 \ell = -X^\top W X, \quad W = \operatorname{diag}\big(p_i(1-p_i)\big).$$



Confident and wrong is punished without limit.

- → $W \succeq 0$, so ℓ is *concave*: no spurious local maxima.
- → $\nabla \ell = 0$ has no closed form, therefore iterate (Newton = iteratively reweighted least squares).
- → Our 14 knights $\Rightarrow (\hat{\beta}_0, \hat{\beta}_1) = (3.23, -0.38)$.

From likelihood to log-likelihood

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i} \xrightarrow{\log}$$

$$\ell = \sum_{i=1}^{n} \big[y_i \log p_i + (1 - y_i) \log(1 - p_i) \big].$$

Fitting = walking downhill on the loss surface

The loss. Maximising ℓ is minimising the negative log-likelihood $\mathrm{NLL}(\boldsymbol{\beta}) = -\ell(\boldsymbol{\beta})$. Its Hessian $X^\top W X \succeq 0$, so the surface is a *single convex bowl* $\rightarrow$ one minimum & no traps.

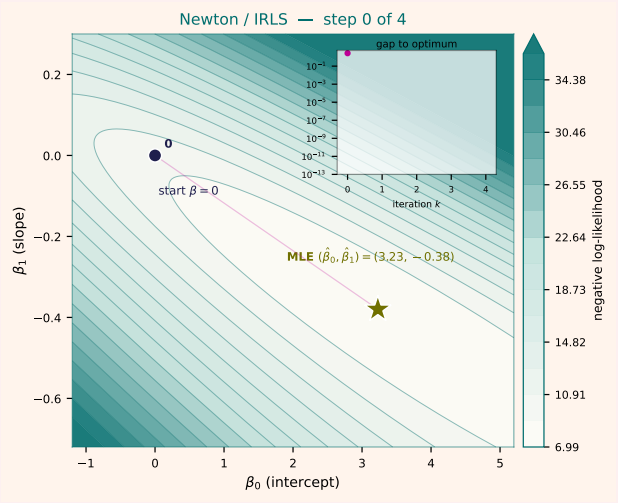
Newton / IRLS step.

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k + \left(X^\top W_k X\right)^{-1} X^\top (\mathbf{y} - \mathbf{p}_k).$$

We will revisit optimization part in next lecture.

Start at $\boldsymbol{\beta} = 0$ ("no effect"). The NLL falls

$$9.70 \;\rightarrow\; 7.10 \;\rightarrow\; 6.995 \;\rightarrow\; 6.9931.$$



Contours of the NLL over (β_0, β_1); the inset tracks the gap to the optimum

Fitting = walking downhill on the loss surface

The loss. Maximising ℓ is minimising the negative log-likelihood $\mathrm{NLL}(\boldsymbol{\beta}) = -\ell(\boldsymbol{\beta})$. Its Hessian $X^\top W X \succeq 0$, so the surface is a *single convex bowl* $\rightarrow$ one minimum & no traps.

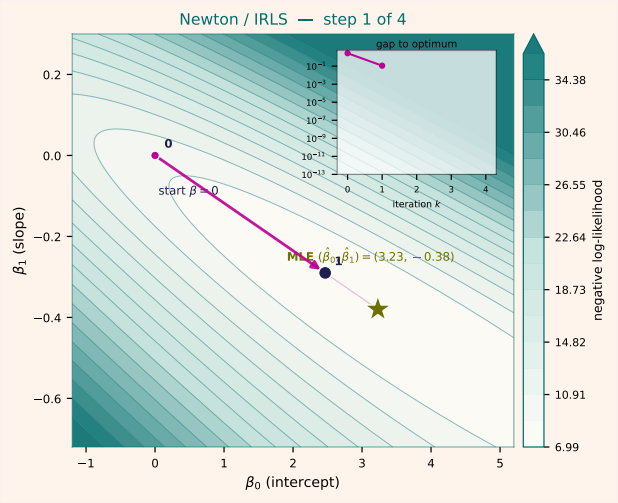
Newton / IRLS step.

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k + \left(X^\top W_k X\right)^{-1} X^\top (\mathbf{y} - \mathbf{p}_k).$$

We will revisit optimization part in next lecture.

Start at $\boldsymbol{\beta} = 0$ ("no effect"). The NLL falls

$$9.70 \;\rightarrow\; 7.10 \;\rightarrow\; 6.995 \;\rightarrow\; 6.9931.$$



Contours of the NLL over (β_0, β_1); the inset tracks the gap to the optimum

Fitting = walking downhill on the loss surface

The loss. Maximising ℓ is minimising the negative log-likelihood $\mathrm{NLL}(\boldsymbol{\beta}) = -\ell(\boldsymbol{\beta})$. Its Hessian $X^\top W X \succeq 0$, so the surface is a *single convex bowl* $\rightarrow$ one minimum & no traps.

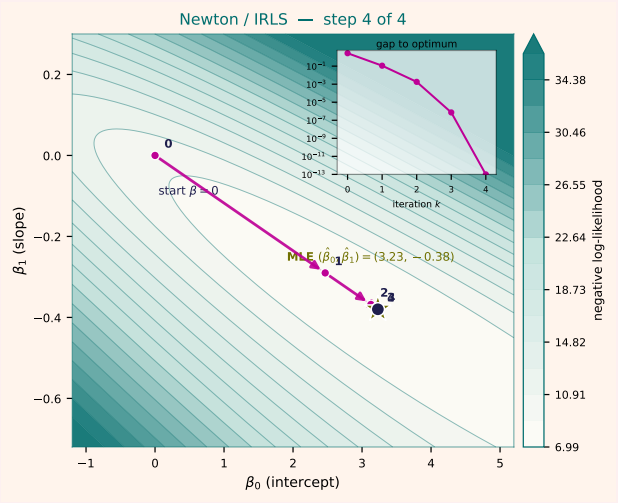
Newton / IRLS step.

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k + \left(X^\top W_k X\right)^{-1} X^\top (\mathbf{y} - \mathbf{p}_k).$$

We will revisit optimization part in next lecture.

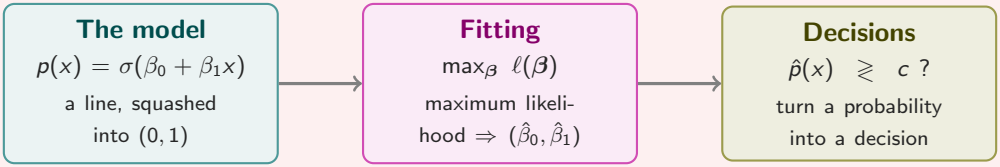
Start at $\boldsymbol{\beta} = 0$ ("no effect"). The NLL falls

$$9.70 \;\rightarrow\; 7.10 \;\rightarrow\; 6.995 \;\rightarrow\; 6.9931.$$



Contours of the NLL over (β_0, β_1); the inset tracks the gap to the optimum

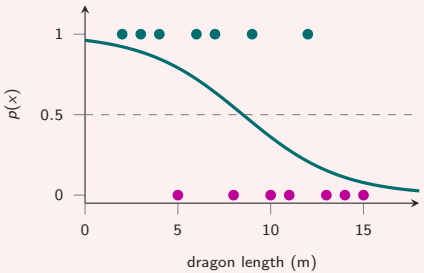
Where we are


The model
$p(x) = \sigma(\beta_0 + \beta_1 x)$
a line, squashed into $(0, 1)$
Fitting
$\max_{\beta} \ \ell(\beta)$
maximum likelihood $\Rightarrow (\hat{\beta}_0, \hat{\beta}_1)$
Decisions
$\hat{p}(x) \ \gtrless \ c$?
turn a probability into a decision
what is the model?
how do we find it?
how do we use it?

From probabilities to decisions

The object. The model returns a probability. We need a "rule" to turn it into a decision:

$$\hat{y} = 1 \iff \hat{p}(x) \geq c, \qquad (\text{default } c = \tfrac{1}{2}).$$



Our fitted model. Now we choose a decision rule.

From probabilities to decisions

The object. The model returns a probability. We need a "rule" to turn it into a decision:

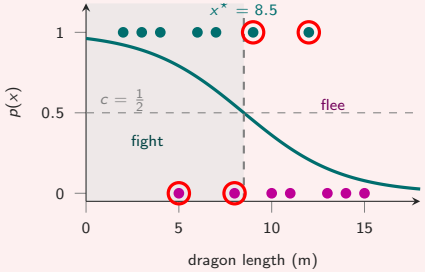
$$\hat{y} = 1 \iff \hat{p}(x) \geq c, \qquad (\text{default } c = \tfrac{1}{2}).$$

The default, $c = \frac{1}{2}$. Since σ is increasing,

$$\hat{p}(x) \geq \tfrac{1}{2} \iff \hat{\beta}_0 + \hat{\beta}_1 x \geq 0 \iff x \leq x^\star = -\frac{\hat{\beta}_0}{\hat{\beta}_1} = 8.5 \text{ m}.$$

Fight if the dragon is under 8.5 m, else flee.

But the knights at 5 and 8 m are sent to fight, and they fall. ☹



$c = \frac{1}{2}$: cutoff at 8.5 m; 4 knights on the wrong side.

From probabilities to decisions

The object. The model returns a probability. We need a "rule" to turn it into a decision:

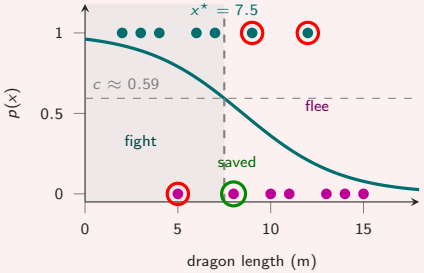
$$\hat{y} = 1 \iff \hat{p}(x) \geq c, \qquad (\text{default } c = \tfrac{1}{2}).$$

Raise the bar. Take $c \approx 0.59$:

$$\hat{p}(x) \geq 0.59 \iff x \leq x^\star = 7.5 \text{ m}.$$

Now the 8 m knight is told to flee and survives.

Errors $4 \to 3$.



$c \approx 0.59$: cutoff at 7.5 m; the 8 m knight is saved; 3 errors left.

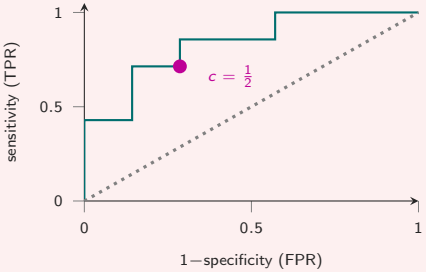
Then what is the right threshold?

At $c = \frac{1}{2}$ (cutoff 8.5 m), our 14 knights give:

	predict survive	predict perish
survived	5 (TP)	2 (FN)
perished	2 (FP)	5 (TN)

Accuracy $= 10/14 \approx 71\%$.

The two FP are the knights at 5 and 8 m, sent to fight, and killed. ☹



Each threshold is one point, $c = \frac{1}{2}$ sits here.

Then what is the right threshold?

At $c = \frac{1}{2}$ (cutoff 8.5 m), our 14 knights give:

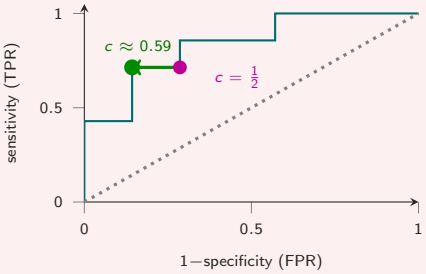
	predict survive	predict perish
survived	5 (TP)	2 (FN)
perished	2 (FP)	5 (TN)

Accuracy $= 10/14 \approx 71\%$.

The two FP are the knights at 5 and 8 m, sent to fight, and killed. ☹

For the next lecture. Every c is one point on the *ROC curve*. The $c = \frac{1}{2}$ point is *not* best: at $c \approx 0.59$ (cutoff 7.5 m) we catch the *same* survivors but raise *one fewer* false alarm.

$\Rightarrow$ the 8 m knight is saved ☺ accuracy $11/14 \approx 79\%$, errors $4 \to 3$.

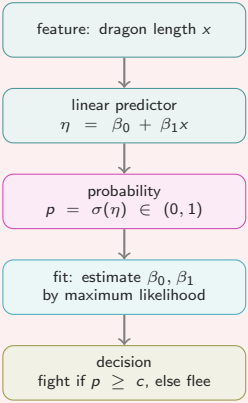


Raise c: move **left** to a better point, same sensitivity, one fewer false alarm. AUC$\approx$ 0.84.

Summary

This lecture.

- → Introduced the logistic regression model, $p(x) = \sigma(\beta_0 + \beta_1 x)$, a Bernoulli response with a logit link, and interpreted β_1 on the odds scale (e^{β_1} is an odds ratio).
- → Estimated the parameters by maximum likelihood; showed the log-likelihood is concave and solved it by Newton/IRLS method.
- → Turned predicted probabilities into decisions via a threshold.



The whole model on one line

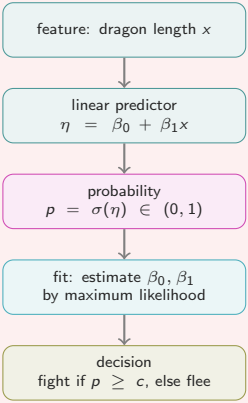
Summary

This lecture.

- → Introduced the logistic regression model, $p(x) = \sigma(\beta_0 + \beta_1 x)$, a Bernoulli response with a logit link, and interpreted β_1 on the odds scale (e^{β_1} is an odds ratio).
- → Estimated the parameters by maximum likelihood; showed the log-likelihood is concave and solved it by Newton/IRLS method.
- → Turned predicted probabilities into decisions via a threshold.

Next lectures.

- → Higher dimensional situation, discussion on IRLS.
- → Model comparison and selection using deviance.

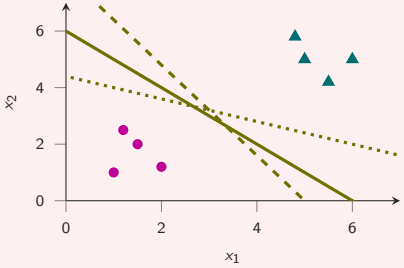


The whole model on one line

THINK

No marks, no quiz, no single right answer. Just sit with each one for a moment.

1. What if the two classes are *perfectly separated* by a line?
   Hint: the likelihood keeps rewarding a steeper fit, so $\|\hat{\boldsymbol{\beta}}\| \to \infty$ and the MLE escapes to infinity. Could a penalty on $\|\boldsymbol{\beta}\|$ (regularisation) help?
2. Our dragon had *one* feature. What changes when we have, $\mathbf{x} \in \mathbb{R}^d$?
3. In Newton/IRLS we wrote $\left(X^\top W X\right)^{-1}$. Do we ever actually *need* that inverse?
   Solve a linear system, or invert-then-multiply....hmmmm....which is cheaper, and which is safer?
4. Keep the sigmoid, but replace $\beta_0 + \beta_1 x$ with a general $\phi(\mathbf{x}; \boldsymbol{\beta})$. What do you gain and what might you lose?
   Curved boundaries come for free, but is the log-likelihood still concave?



Question 1 in a picture: separable data, infinitely many perfect boundaries, and the likelihood keeps preferring steeper ones.

THINK

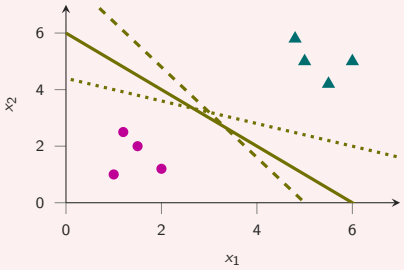
No marks, no quiz, no single right answer. Just sit with each one for a moment.

1. What if the two classes are *perfectly separated* by a line?
Hint: the likelihood keeps rewarding a steeper fit, so $\|\hat{\boldsymbol{\beta}}\| \to \infty$ and the MLE escapes to infinity. Could a penalty on $\|\boldsymbol{\beta}\|$ (regularisation) help?

2. Our dragon had *one* feature. What changes when we have, $\mathbf{x} \in \mathbb{R}^d$?

3. In Newton/IRLS we wrote $\left(X^\top W X\right)^{-1}$. Do we ever actually *need* that inverse?
Solve a linear system, or invert-then-multiply....hmmmm....which is cheaper, and which is safer?

4. Keep the sigmoid, but replace $\beta_0 + \beta_1 x$ with a general $\phi(\mathbf{x}; \boldsymbol{\beta})$. What do you gain and what might you lose?
Curved boundaries come for free, but is the log-likelihood still concave?



Question 1 in a picture: separable data, infinitely many perfect boundaries, and the likelihood keeps preferring steeper ones.

Some of these we take up next lecture; the rest are for you to think.