Regular Linear Regression

$$\underline{y} = \underline{X}\,\underline{\beta}\ ,\quad y \in [-\infty, \infty]$$

What if y is categorical?

$$\underline{y} \in \{0, 1\}$$

Applications

Pass/fail test

tumor/normal tissue

disease / not

for independent decision variables, no change in method

for dependent response variables, requires very different framework

$\Rightarrow$ logistic regression

$$\underline{x} \longrightarrow \boxed{\text{linear model}} \longrightarrow \underset{\substack{\uparrow \\ \text{continuous}}}{\overset{[-\infty,\infty]}{t}} \longrightarrow \boxed{\text{link function}} \longrightarrow \underset{\text{interpreted as a probability}}{(0, 1)}$$

What is our link function?

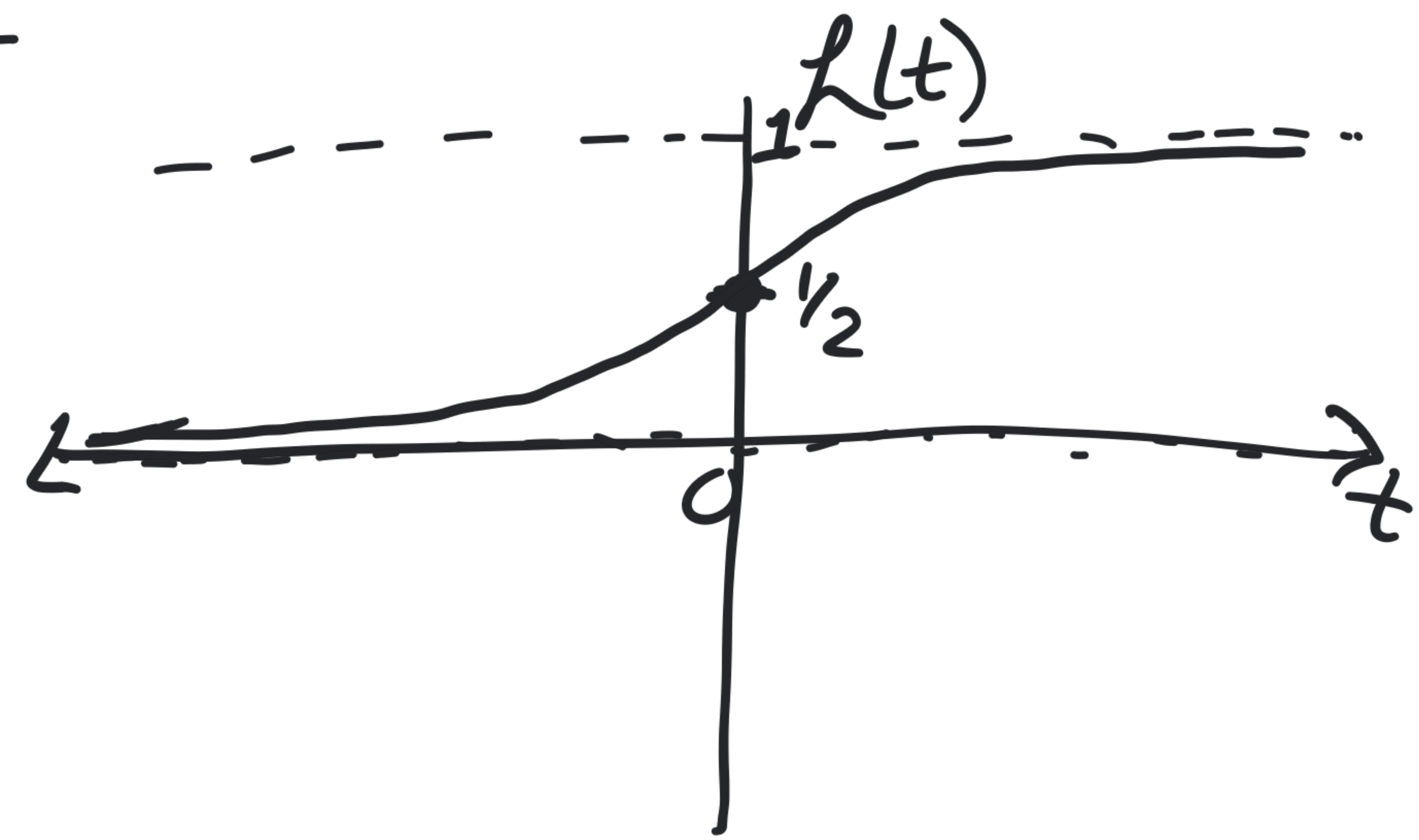
Logistic function (logistic regression) (one of many link functions)

$$\mathcal{L}(t) = \frac{1}{1+e^{-t}}$$

$$\mathcal{L}(0) = \frac{1}{1+e^{0}} = \frac{1}{1+1} = \frac{1}{2}$$

$$\mathcal{L}(-\infty) = \frac{1}{1+e^{+\infty}} \longrightarrow 0$$

$$\mathcal{L}(\infty) = \frac{1}{1+\underbrace{e^{-\infty}}_{0}} \longrightarrow \frac{1}{1+0} = 1$$



$t = \begin{pmatrix}\text{linear}\\ \text{model}\end{pmatrix} = 0 \Rightarrow \mathbb{P}(t) = 1/2$

$\begin{pmatrix}\text{linear}\\ \text{model}\end{pmatrix} > 0 \Rightarrow y = 1$

$\begin{pmatrix}\text{linear}\\ \text{model}\end{pmatrix} < 0 \Rightarrow y = 0$

Interpreting coefficients of Logistic Regression

output is Probability

Input $\longrightarrow$ link function $\longrightarrow$ P

$\Rightarrow$ coefficients tell us about Δt, not ΔP!

Let's talk about odds:

P: long-run fraction of successes

$P = 0.15$, ~15% of <u>many</u> trials will succeed.

odds: $\dfrac{P}{1-P}$ $\quad \dfrac{\leftarrow \text{p of success}}{\leftarrow \text{p of failure}}$ $\quad P + (1-P) = 1$

Team has $P = 0.2$ of winning

they have $\dfrac{0.2}{1-0.2} = \dfrac{0.2}{0.8} = \dfrac{1}{4}$ odds

What are the odds for a logistic function

$$P = \frac{1}{1+e^{-t}}$$

$$\text{odds} = \frac{P}{1-P} = \frac{\cancel{1/1+e^{-t}}}{1 - 1/1+e^{-t}} \frac{\cancel{(1+e^{-t})}}{(1+e^{-t})}$$

$$= \frac{1}{(1+e^{-t}) - 1}$$

$$= \frac{1}{e^{-t}} = e^{t}$$

Linear model $\beta_0 + \beta_1 x$

odds of response $= 1$ are $e^{\beta_0 + \beta_1 x}$

For Huntington's Example:

$\text{odds}([\text{CAGs}])$

$= e^{-10+0.4[\text{CAGs}]}$

odds(20 CAGs)

$= 0.14 : 1$

odds(35 CAGs)

$= 55 : 1$

Odds Ratio

$$\frac{odds(x+1)}{odds(x)}$$

fold increase in odds with a unit increase in the input.

For Huntington's, fold increase of getting disease given 1 more CAG.

$$OR = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = \frac{\cancel{e^{\beta_0}}\,\cancel{e^{\beta_1 x}}\,e^{\beta_1}}{\cancel{e^{\beta_0}}\,\cancel{e^{\beta_1 x}}} = e^{\beta_1}$$

$$t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$

$OR_{x_1} = e^{\beta_1}$ unit Δ in x_1

$OR_{x_2} = e^{\beta_2}$ unit Δ in x_2

For Huntington's:

$$OR = e^{\beta_1} = e^{0.4} = 1.5$$