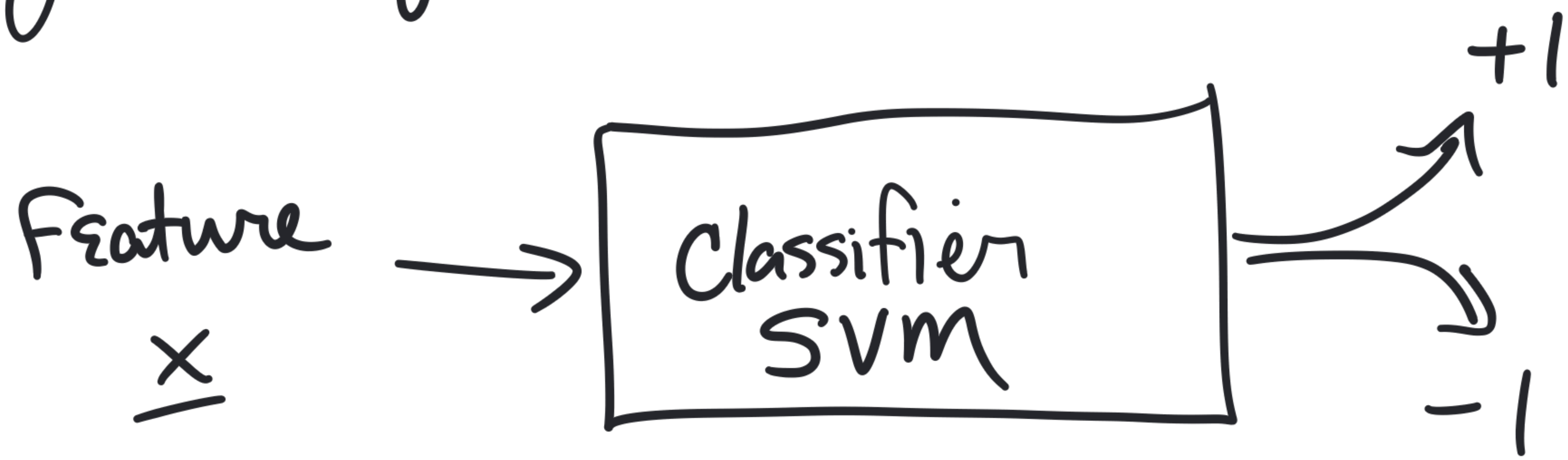
Logistic Regression

BIOE 210

Binary Classification



Problems to using a linear model

1. Classification is discrete
2. Bounds

Solution: Logistic Regression

We predict $P(y=1)$, but probabilities $[0,1]$

$$\text{odds}(y) = \frac{P(y=1)}{P(y=0)}$$

$$P(y=1) + P(y=0) = 1 \Rightarrow P(y=0) = 1 - P(y=1)$$

$$\text{odds}(y) = \frac{P(y=1)}{1-P(y=1)} \in [0, \infty)$$

$$P(y=1) = \frac{\text{odds}(y)}{1+\text{odds}(y)} \in [0, 1]$$

$$\log(\text{odds}(y)) \in (-\infty, \infty)$$

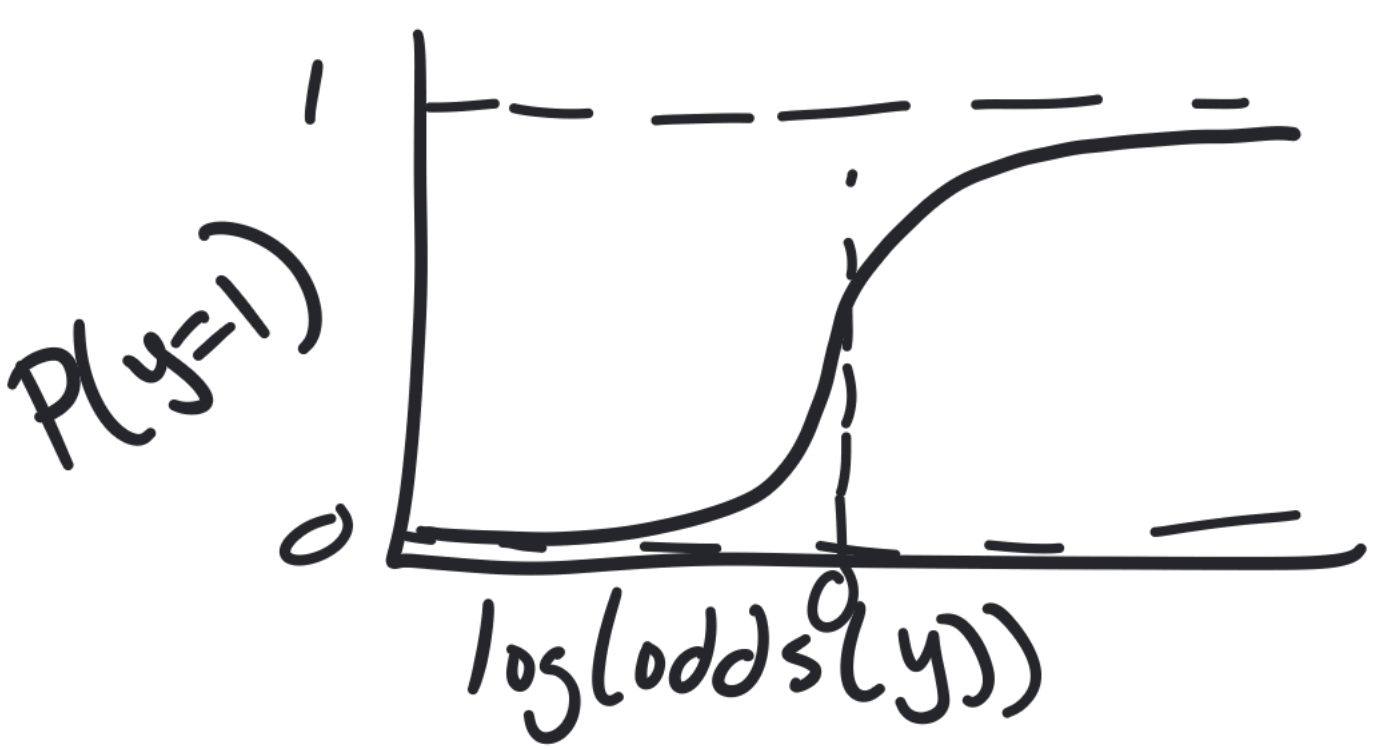
Logistic Regression

$$\log(\text{odds}(y)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

$$\text{odds}(y) = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n} \equiv e^t$$

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

$$\underline{x} \rightarrow \underset{\beta_0+\beta_1 x_1+\cdots}{\text{linear model}} \rightarrow t \rightarrow \underbrace{\frac{1}{1+e^{-t}}}_{\text{Link function}} \rightarrow P(y=1)$$



Huntingtin (*HTT*)

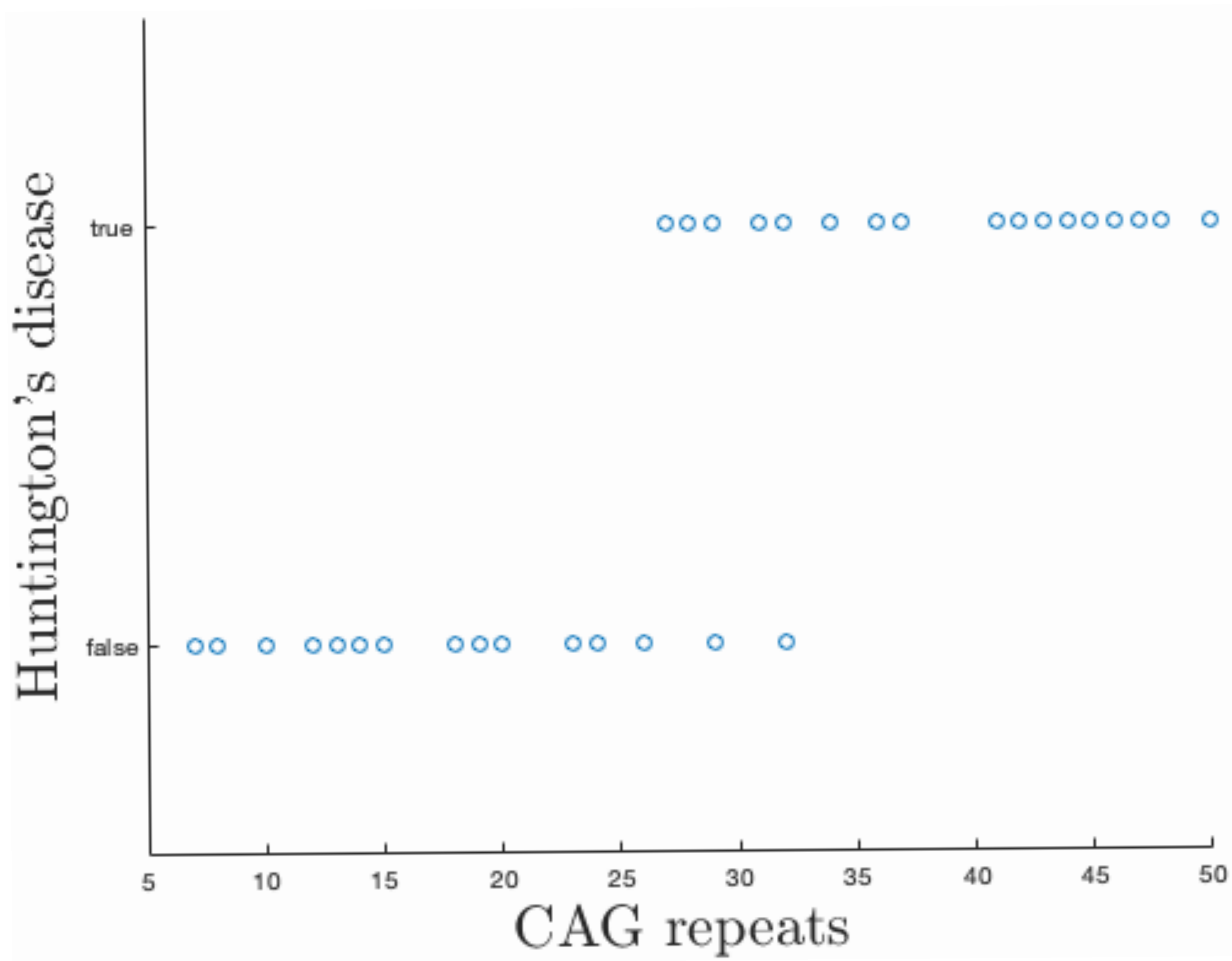
```
Leu Lys Ser Phe Gln Gln ... Gln Gln Gln Gln Pro
ctc aag tcc ttc cag cag ... cag cag caa cag ccg
```

# of CAG Repeats	Disease Outcome
< 28	Not affected.
28-35	Increased risk.
36-40	Affected; some offspring affected.
> 40	Affected; all offspring affected.

Source: Walker FO. Huntington's disease. *The Lancet*. 2007: **369**, (9557), 218–228

$$P(\text{disease}) = f([\text{CAGs}])$$

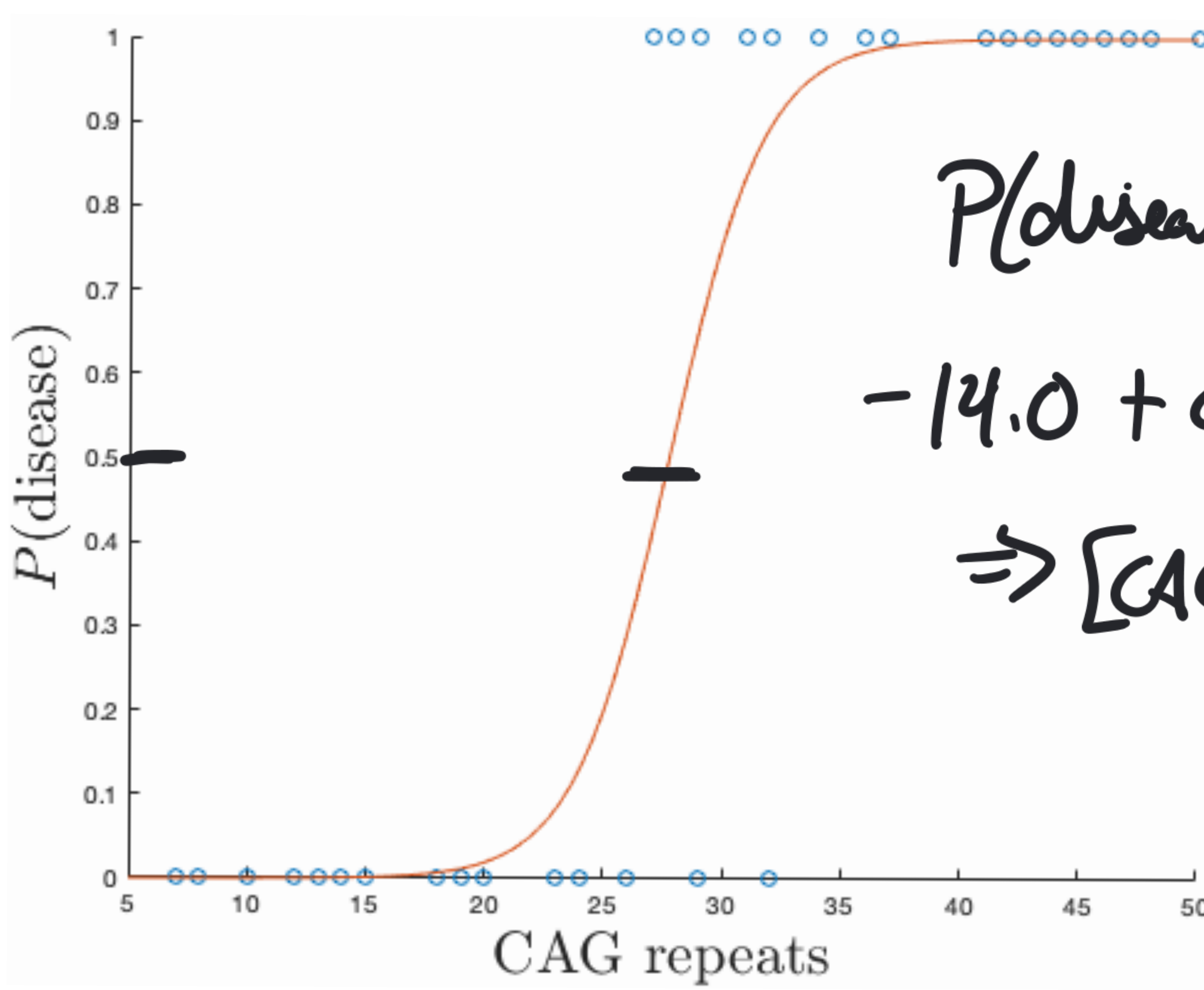
$$\log(\text{odds}(\text{disease})) = \beta_0 + \beta_1[\text{CAGs}]$$



$$\log(\text{odds}(\text{disease})) = \beta_0 + \beta_1 [\text{CAGs}]$$

$$= -14.0 + 0.51[\text{CAGs}]$$

$$P(\text{disease}) = \frac{1}{1 + e^{-14.0 + 0.51[\text{CAGs}]}}$$



$$P(\text{disease}) = 1/2$$

$$-14.0 + 0.51[\text{CAGs}] = 0$$

$$\Rightarrow [\text{CAGs}] = 14.0/0.51 \approx 28$$

How do we interpret the β's ?

$$\text{odds ratio}(x_i) = \frac{\text{odds}(x_i + 1)}{\text{odds}(x_i)}$$

$$\text{odds ratio}([CAGs]) = \frac{\text{odds}([CAGs]+1)}{\text{odds}([CAGs])}$$

$$= \frac{e^{\beta_0 + \beta_1([CAGs]+1)}}{e^{\beta_0 + \beta_1 [CAGs]}}$$

$$= \frac{\cancel{e^{\beta_0}}\,\cancel{e^{\beta_1[CAGs]}}\,e^{\beta_1}}{\cancel{e^{\beta_0}} + \cancel{e^{\beta_1[CAGs]}}} = e^{\beta_1}$$

$$\text{odds ratio}(x_i) = \frac{\text{odds}(x_i+1)}{\text{odds}(x_i)} = e^{\beta_i}$$