GLMs for Binary Data

2/22/22

Response variable: $Y = \begin{cases} 1 & \text{w/prob. } \pi \\ 0 & \text{w/prob. } 1-\pi \end{cases}$

Goal: Model π using covariates $x_1, x_2, \dots, x_k$.

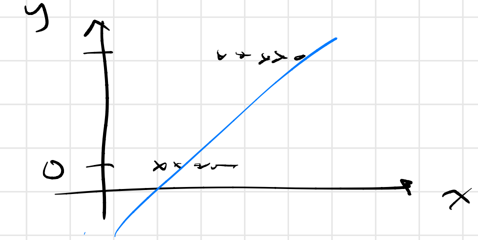
Note: $E(Y) = \pi$ $0 \le \pi \le 1$

① $Y \sim Bin(1, \pi)$ ← Random component

$\eta = \alpha + \beta x$ ← Systematic component

$g(\pi) = \pi$ ← link function

↳ Linear probability model: $\pi(x) = \alpha + \beta x$



Interpret:

α: Our probability (π) when $x = 0$.

Not predicted value of Y when $x=0$ — why?

→ $Y \in \{0, 1\}$

β: Change in π for a 1 unit increase in x.

Issues with this model: $-\infty < \alpha + \beta x < \infty$

but $0 \le \pi \le 1$

Framingham Ex:

$\hat{\pi}(x) = \underbrace{-0.18897}_{} + \underbrace{0.00377x}_{}$

The estimated probability of having a heart attack increases by 0.00377 for each 1 mmHg increase in SBP.

not useful
- extrapolation (sbp = 0)
- neg. prob.

A 1 mmHg increase in sys. blood pressure is associated with a 0.00377 increase in estimated prob. of a heart attack.

② Logistic regression : "Logit link"

Random: $Y \sim \text{Bin}(1, \pi)$

Systematic: $\eta = \alpha + \beta x$

Link: $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$

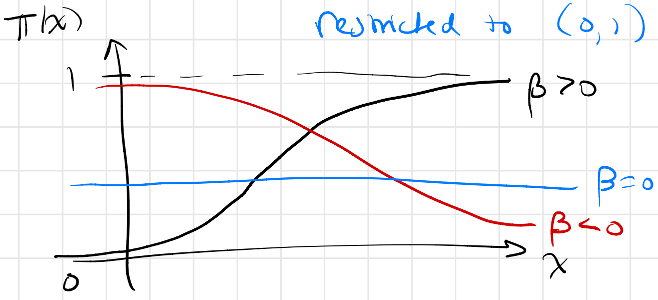
"log odds"

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x$$

$$\Leftrightarrow \quad \frac{\pi}{1-\pi} = e^{\alpha+\beta x} \quad \leftarrow \text{odds}$$

$$\Leftrightarrow \quad \pi = \frac{e^{\alpha+\beta x}}{\underbrace{1+e^{\alpha+\beta x}}_{}} \quad \leftarrow \text{risk}$$

restricted to (0, 1)



$$y = \frac{e^x}{1+e^x}$$

is called a logistic curve

Interpret α & β?

α : The probability that $Y=1$ when $X=0$ is $\frac{e^{\alpha}}{1+e^{\alpha}}$.

The odds when $x=0$ are e^{α}.

β : Odds ratio $x+1$ to x

$$\frac{\dfrac{\pi(x+\overset{10}{\textcircled{1}})}{1-\pi(x+\overset{10}{\textcircled{1}})}}{\dfrac{\pi(x)}{1-\pi(x)}} = \frac{e^{\alpha+\beta(x+\overset{10}{\textcircled{1}})}}{e^{\alpha+\beta x}} = \underline{e^{\overset{10}{}\beta}}$$

From Ex: $$\log\left(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right) = -3.01 + 0.0166x$$

Exponentiate coefs: $e^{-3.01} = 0.0494$

$e^{0.0166} = 1.0167$ ← estimated odds ratio for $x+1$ to x

% increase: 1.67%

A one mmHg increase in SBP is associated with a <u>predicted</u> 1.67% increase in <u>odds</u> of CHD.

What about a <u>10</u> mmHg increase in SBP?

$\exp(10\cdot\hat{\beta}) = 1.1805 \longrightarrow 18\%$

Predicted probability for x:

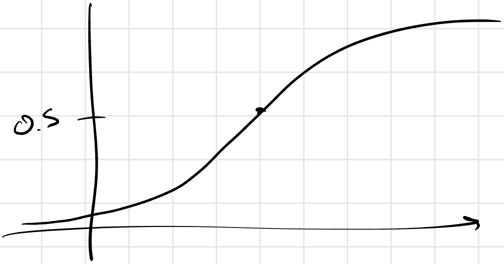
$$\hat{\pi}(x) = \frac{\exp(-3.009 + 0.0166x)}{1 + \exp(-3.009 + 0.0166x)}$$

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$ ← rate of change in $\pi(x)$ depends on value of x

$$\frac{d\pi(x)}{dx} = \beta\pi(x)(1-\pi(x))$$ ← Slope of tangent line to logistic curve at x

↙

Largest when $\pi(x) = 0.5$



↳ $\Rightarrow 0.5 = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$

$\Leftrightarrow \log\left(\frac{0.5}{1-0.5}\right) = \alpha+\beta x$

$\Leftrightarrow \boxed{x = \frac{-\alpha}{\beta}}$

Median effective level

From Ex: $\frac{-(-3.009)}{0.0166} = 181.33$ mmHg

↳ predict prob. of CHD is 0.5

"Hypertension"

SBP ≥ 140 → $x = 140$

$$\hat{\beta}\,\hat{\pi}(x)(1-\hat{\pi}(x)) = 0.0037$$

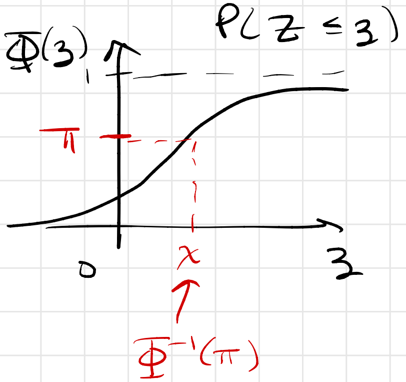
→ Predicted prob. of CHD for a 1 mmHg increase in SBP around 140 mmHg increases by 0.0037.

③ Probit regression:

Random: $Y \sim \text{Bin}(1, \pi)$

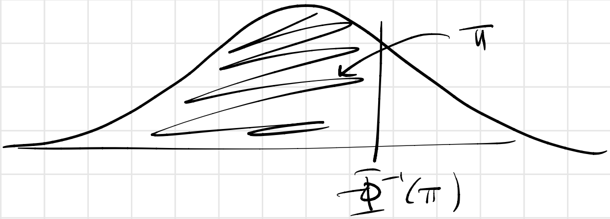
Link: $g(\pi) = \Phi^{-1}(\pi)$

Systematic: $\alpha + \beta x$



From Ex: $\Phi^{-1}(\hat{\pi}(x)) = -1.852 + 0.010x$

$\Phi^{-1}(\pi) \rightarrow$ Z-score that has area π under std. normal curve below



$$\hat{\pi}(x) = \Phi(-1.852 + 0.010x)$$

$$= P(\underset{\substack{\nearrow \\ \text{std. normal} \\ \text{rv.}}}{Z} \le -1.852 + 0.010x)$$

Conf. interval for: OR BP = 140 to BP = 120:

$$\frac{e^{\alpha+\beta(140)}}{e^{\alpha+\beta(120)}} = e^{20\beta}$$

① CI for $\beta \rightarrow (L, U)$

② $(e^{20L}, e^{20U}) \rightarrow$ CI OR.