Agenda

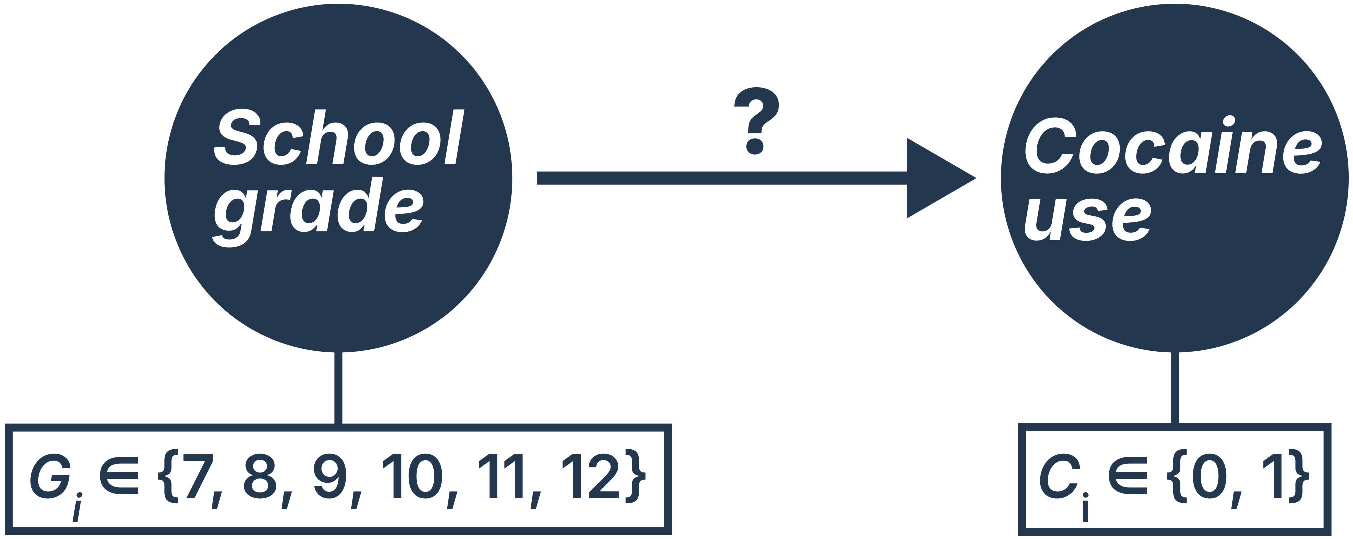
Parsimony & overfitting

1. Administrative
2. Cocaine use among adolescents
3. The inverse logit transformation
4. Starting simple: intercept-only logistic regression
5. ***Hands on:*** *Estimating logistic regression using MCMC in R*

Cocaine use among adolescents

(The trouble with binary outcomes)

COCAINE USE AMONG ADOLESCENTS



Why not use a standard linear regression?

$$C_i \sim \mathrm{Norm}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta G_i$$

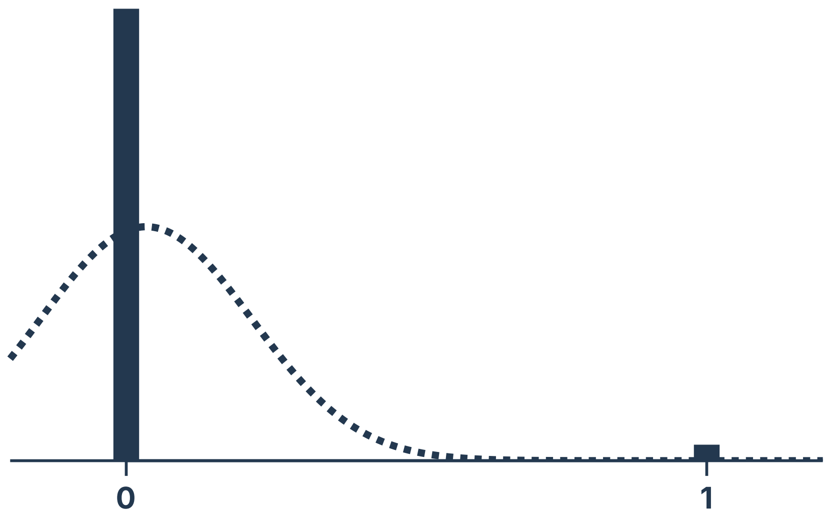
GAUSSIAN MODEL FOR BINARY DATA?

Why not use a standard linear regression?

Wrong support

Normal distribution has a support of (-∞,∞), but we know the outcome variable takes on only two values.

Bad intuitive "fit"

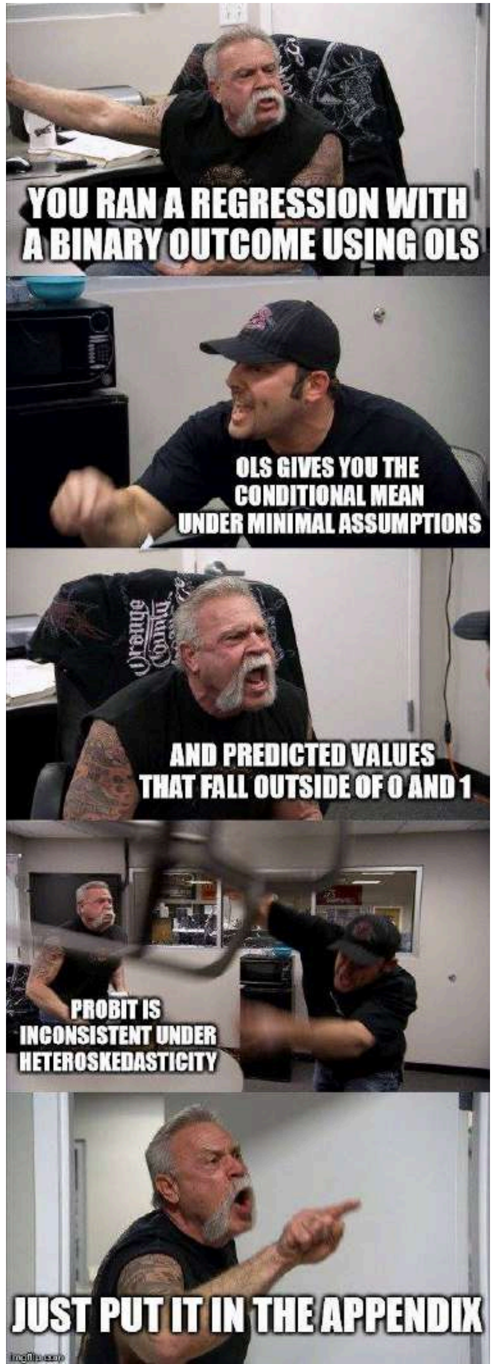


Interpretation

Under some circumstances, results can be interpreted as proportions or probabilities, but this can lead to predicted values less than zero or more than one.

GAUSSIAN MODEL FOR BINARY DATA?

Why not use a standard linear regression?

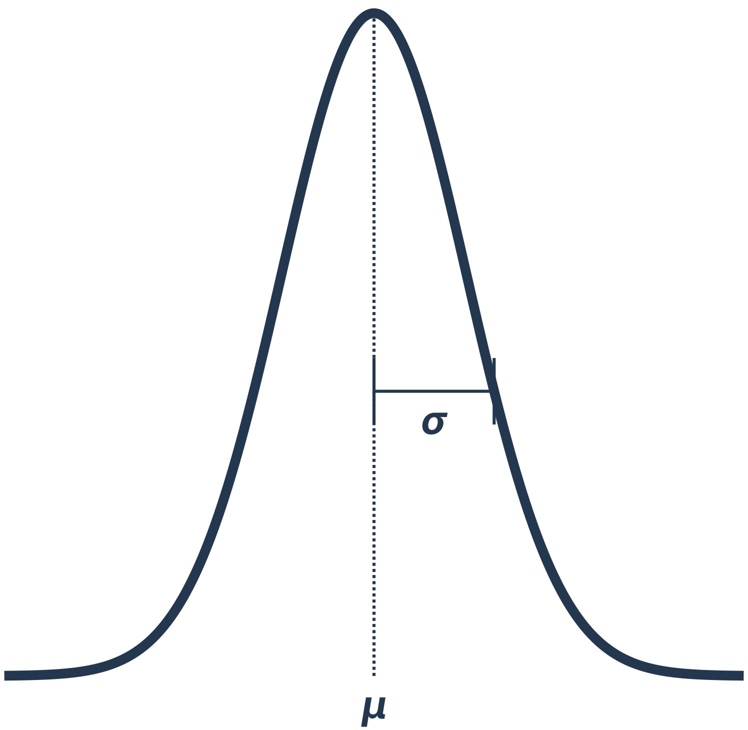


GAUSSIAN VS. BERNOULLI

Gaussian (normal) distribution

$\mathrm{Norm}(\mu, \sigma)$

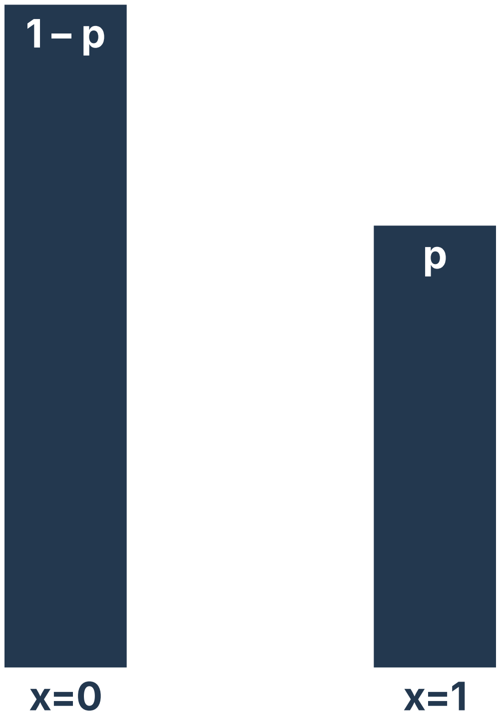
$$\Pr(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Bernoulli distribution

$\mathrm{Bernoulli}(p) = \mathrm{Binomial}(1, \mathrm{p})$

$$\Pr(x|p) = \begin{cases} 1-p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases}$$



LOGISTIC REGRESSION MODEL

Replace Norm(μ, σ) with Bernoulli(p)

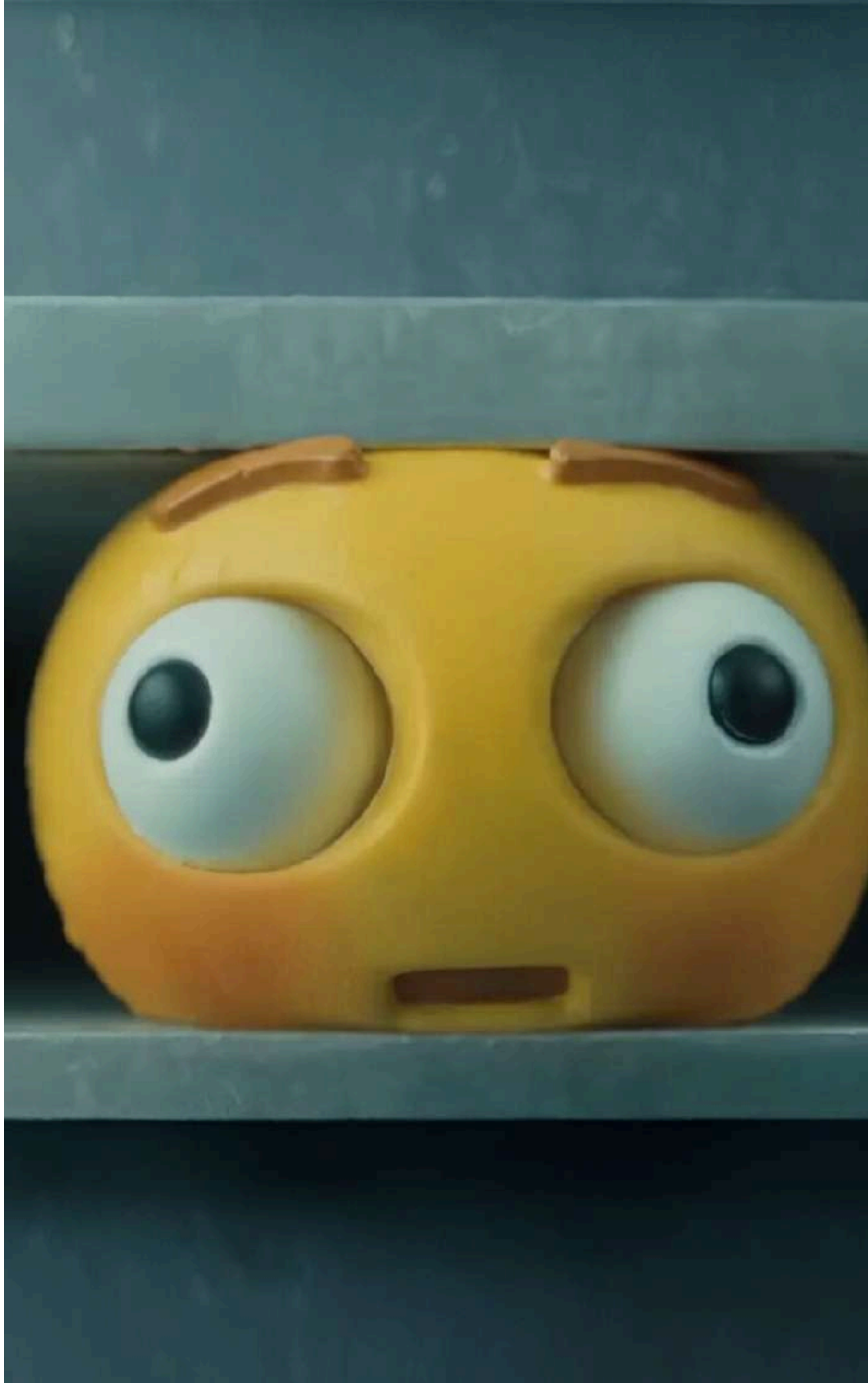
$$C_i \sim \mathrm{Bernoulli}(p_i)$$

$$f(p_i) = \alpha + \beta G_i$$

But now we need a "link" function

With normal distribution, μ could take on any value, but p is restricted to [0,1]

The inverse logit transfor-mation



INVERSE LOGIT TRANSFORMATION

Logit function

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Takes values between 0 and 1, and turns them into values between -∞ and ∞.

Inverse logit function

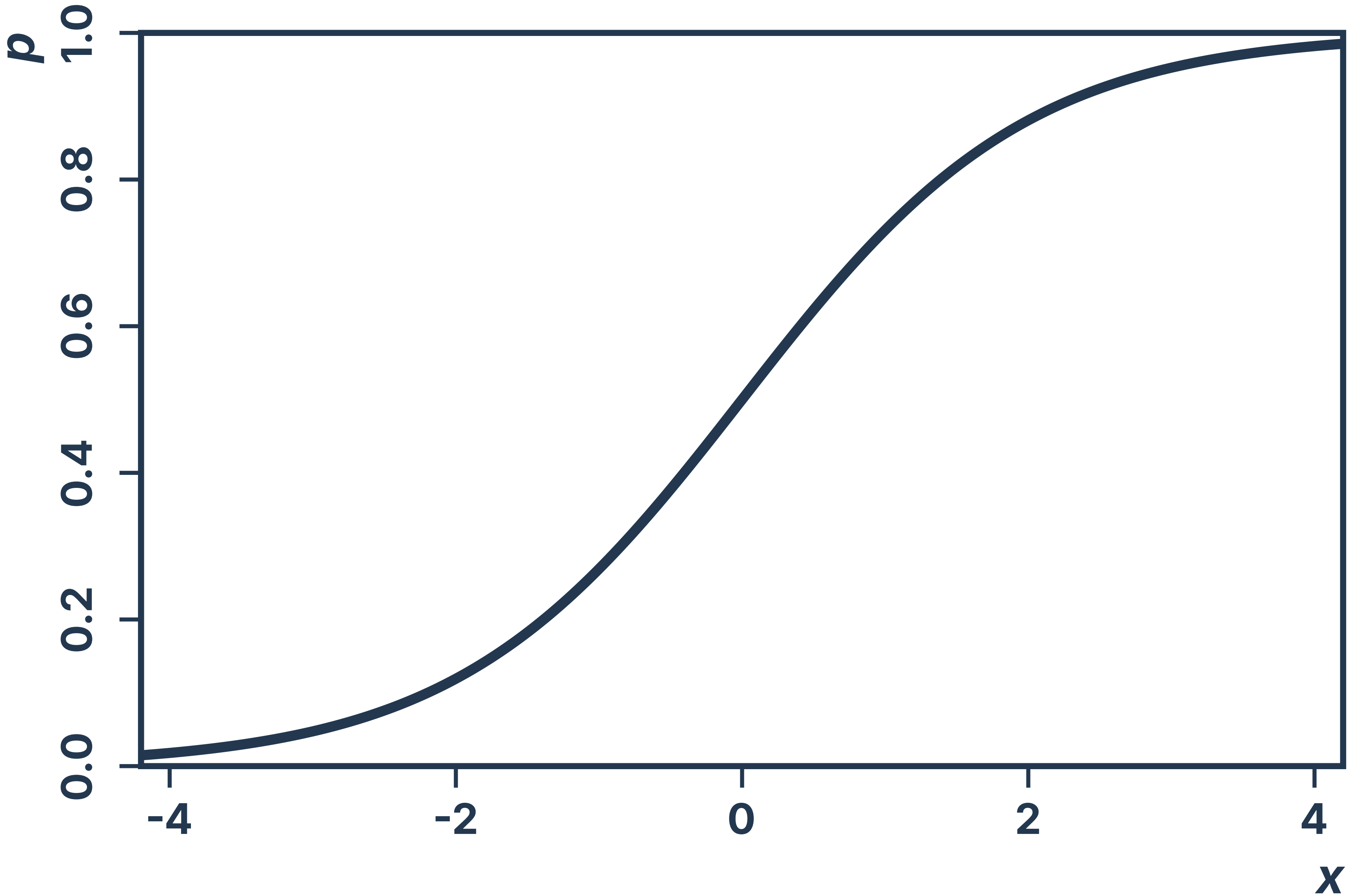
(aka 'logistic')

$$\text{logit}^{-1}(x) = \text{logistic}(x) = \frac{e^x}{e^x+1} = \frac{1}{1+e^{-x}}$$

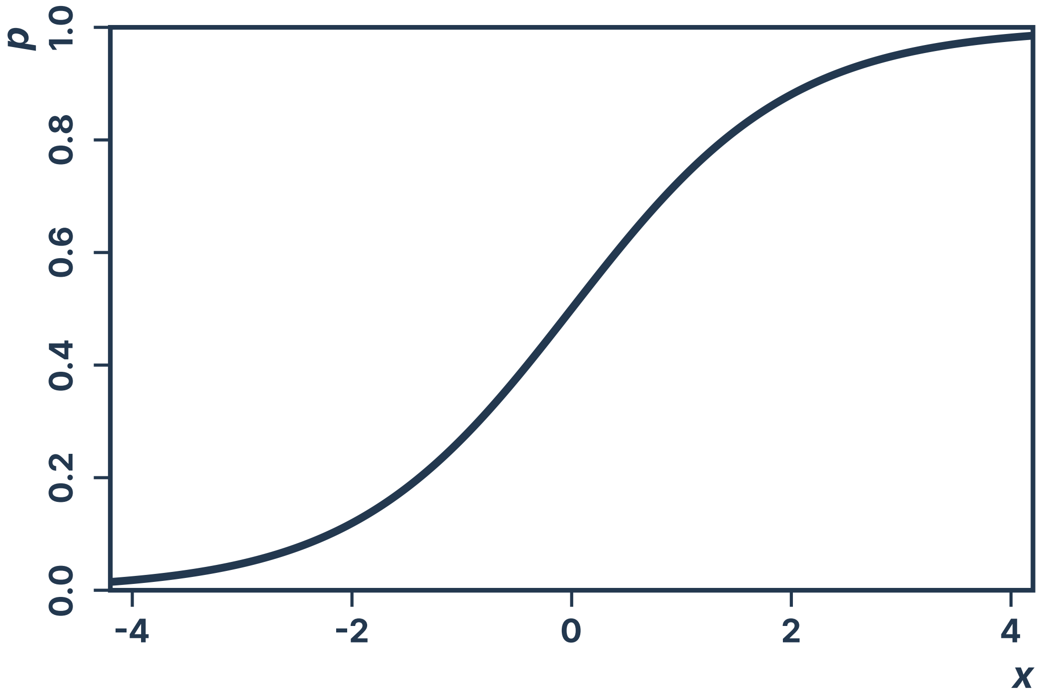
Takes values between -∞ and ∞, and turns them into values between 0 and 1.

$$\begin{aligned} C_i &\sim \text{Bernoulli}(p_i) \\ \text{logit}(p_i) &= \alpha + \beta G_i \end{aligned} \iff \begin{aligned} C_i &\sim \text{Bernoulli}(p_i) \\ p_i &= \text{logit}^{-1}(\alpha + \beta G_i) \end{aligned}$$

INVERSE LOGIT TRANSFORMATION



INVERSE LOGIT TRANSFORMATION



$$C_i \sim \mathrm{Bernoulli}(p_i)$$
$$\mathrm{logit}(p_i) = \alpha + \beta G_i$$

x	$\mathrm{logit}^{-1}(x)$
-2	0.119
-0.5	0.119
0	0.119
0.5	0.119
2	0.119

INTERCEPT-ONLY LOGISTIC MODEL

$$
\begin{aligned}
C_i &\sim \text{Bernoulli}(p_i) \\
\text{logit}(p_i) &= \alpha
\end{aligned}
$$

Why this model instead of the model we built in the first week of class?

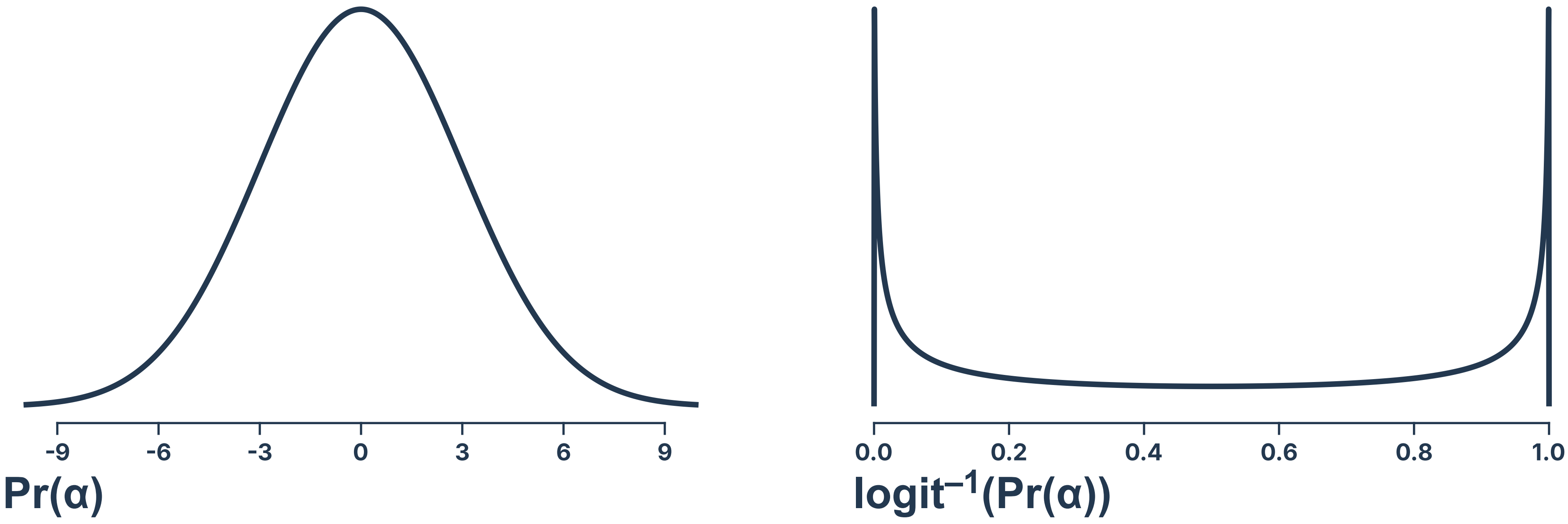
$$
\begin{aligned}
\text{Count}(C) &\sim \text{Binom}(n, p) \\
p &\sim \text{Unif}(0, 1)
\end{aligned}
$$

Logistic regression allows us to include explanatory covariates.

$$
\begin{aligned}
C_i &\sim \text{Bernoulli}(p_i) \\
\text{logit}(p_i) &= \alpha \\
\\
\alpha &\sim \text{Norm}(0, ???)
\end{aligned}
$$

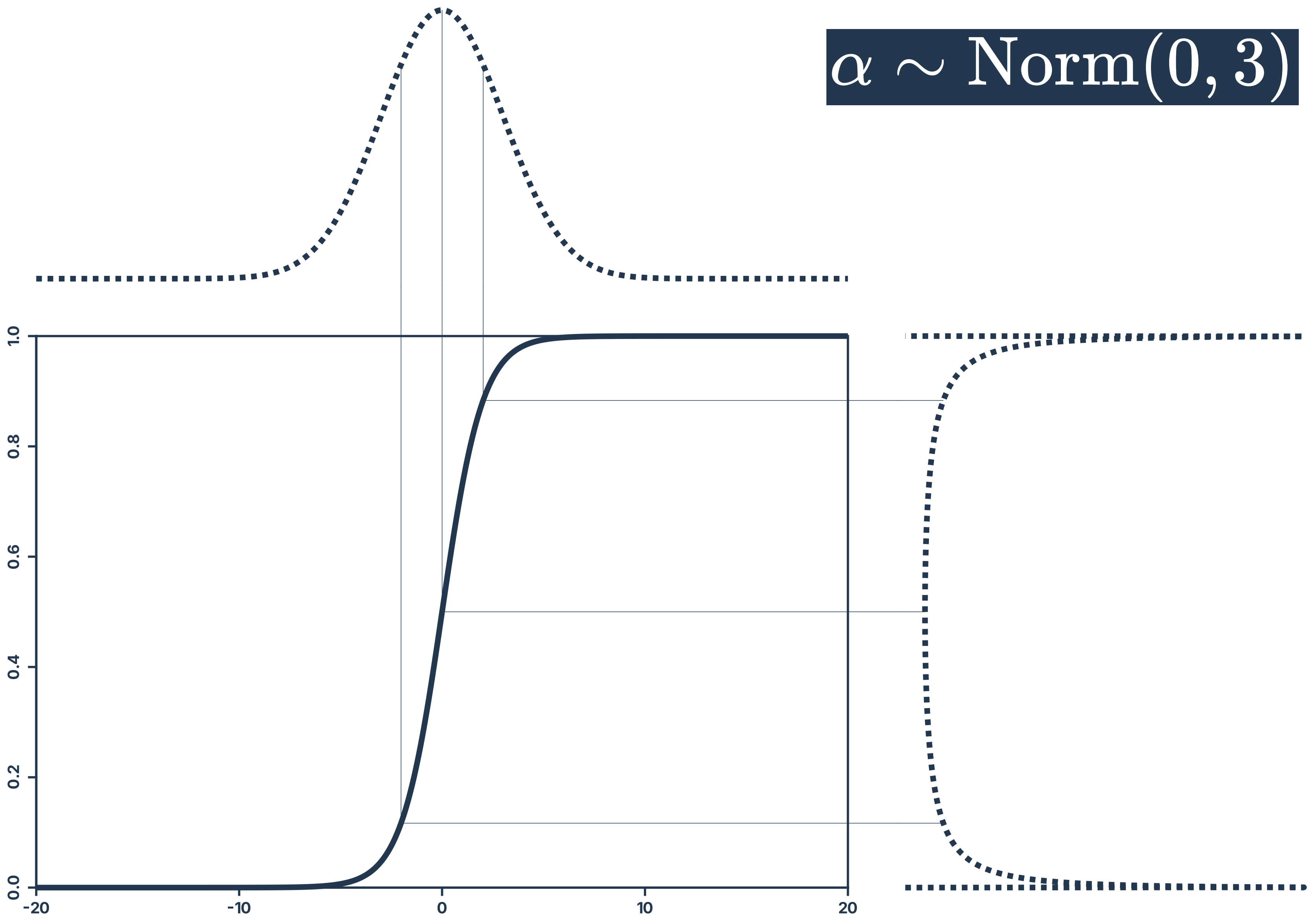
PRIORS IN LOGISTIC REGRESSION

$$\alpha \sim \mathrm{Norm}(0,3)$$



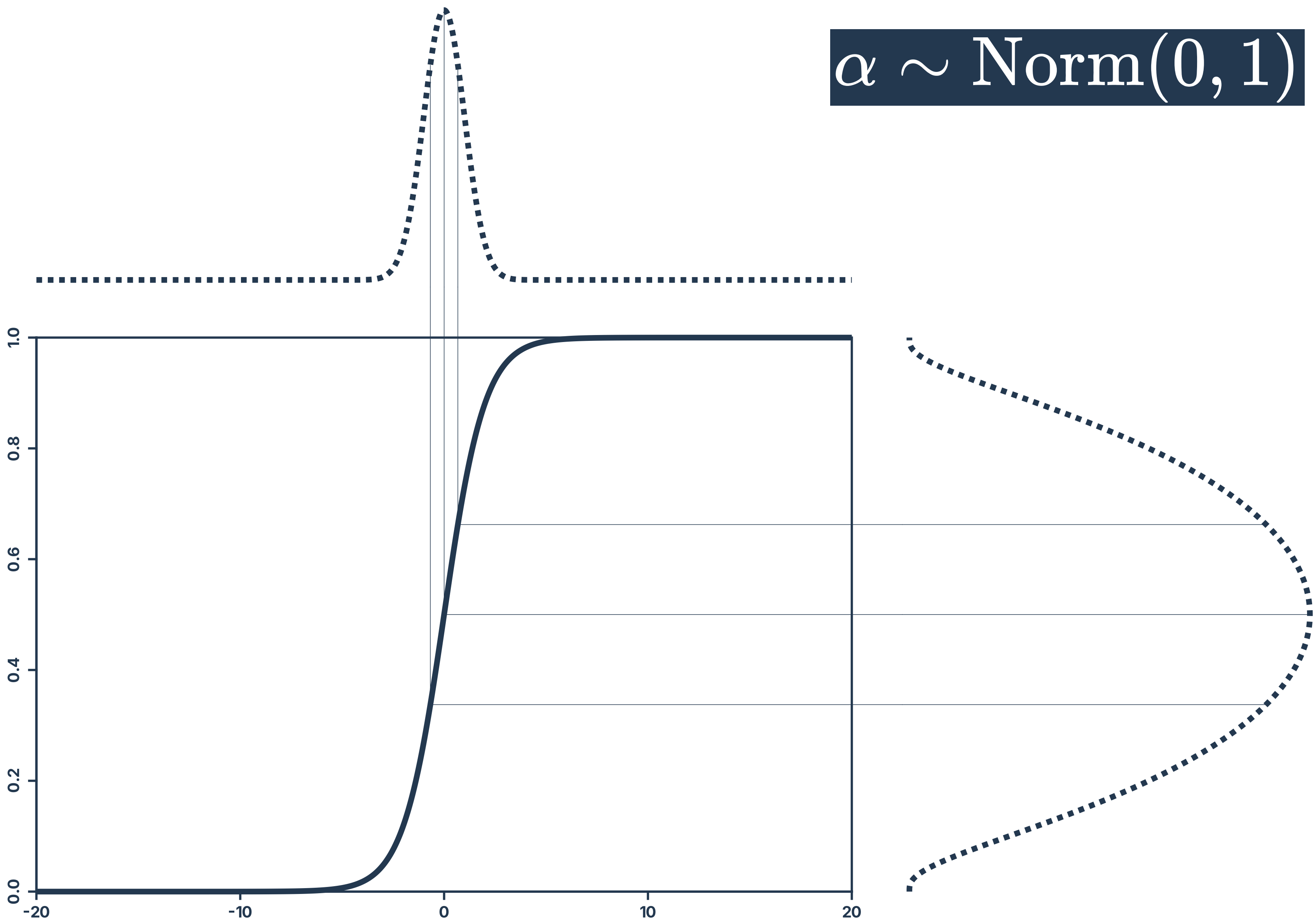
PRIORS IN LOGISTIC REGRESSION

$$\alpha \sim \mathrm{Norm}(0, 3)$$



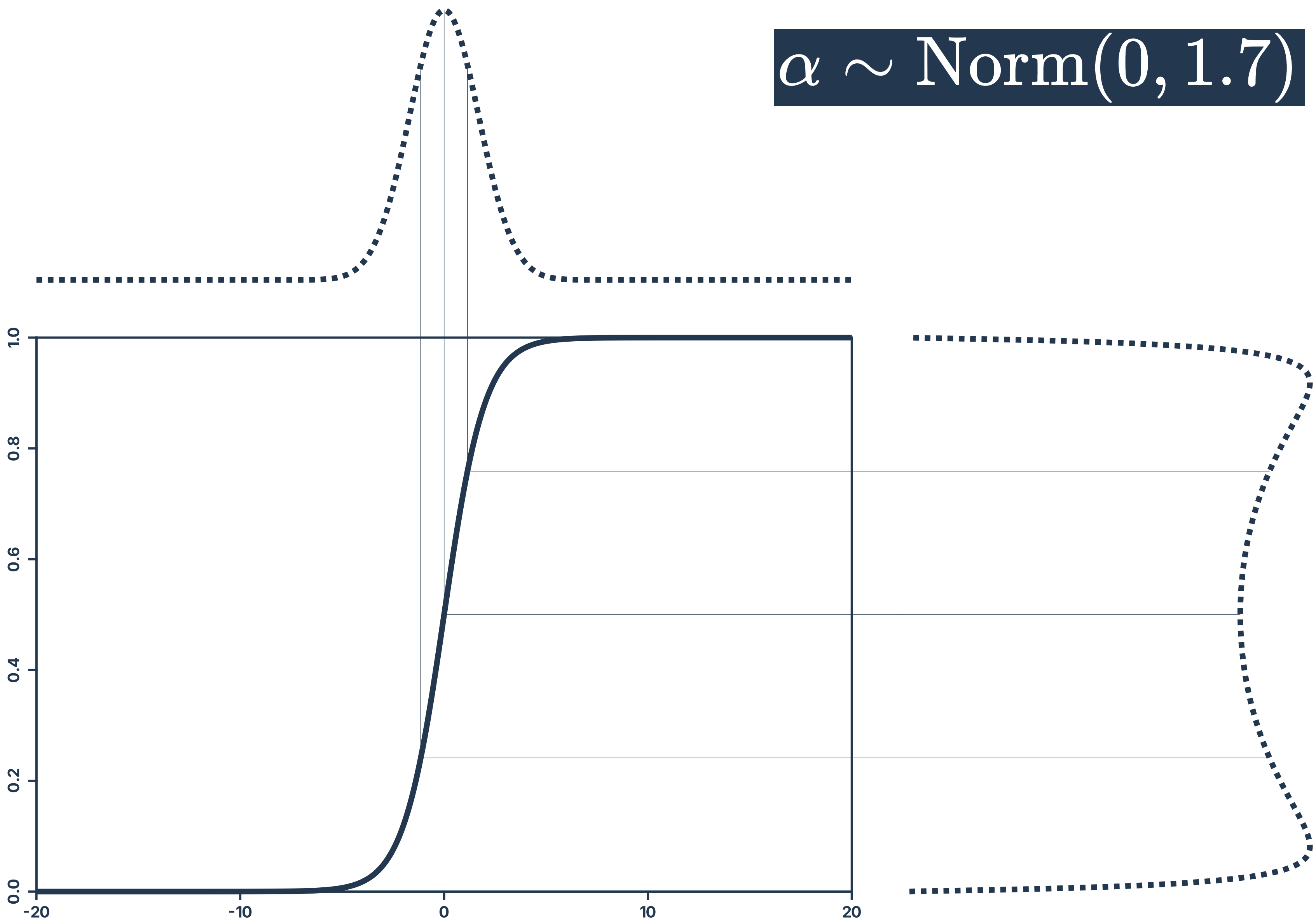
PRIORS IN LOGISTIC REGRESSION

$$\alpha \sim \mathrm{Norm}(0, 1)$$



PRIORS IN LOGISTIC REGRESSION

$\alpha \sim \mathrm{Norm}(0, 1.7)$



INTERCEPT-ONLY LOGISTIC REGRESSION

$$
\begin{aligned}
C_i &\sim \text{Bernoulli}(p_i) \\
\text{logit}(p_i) &= \alpha \\
\\
\alpha &\sim \text{Norm}(0, 1.7)
\end{aligned}
$$

	Median	**95% C.I.**
α	-3.34	(-3.48, -3.20)
$\exp(\alpha)$	0.036	(0.031, 0.041)
$\text{logit}^{-1}(\alpha)$	0.034	(0.030, 0.039)