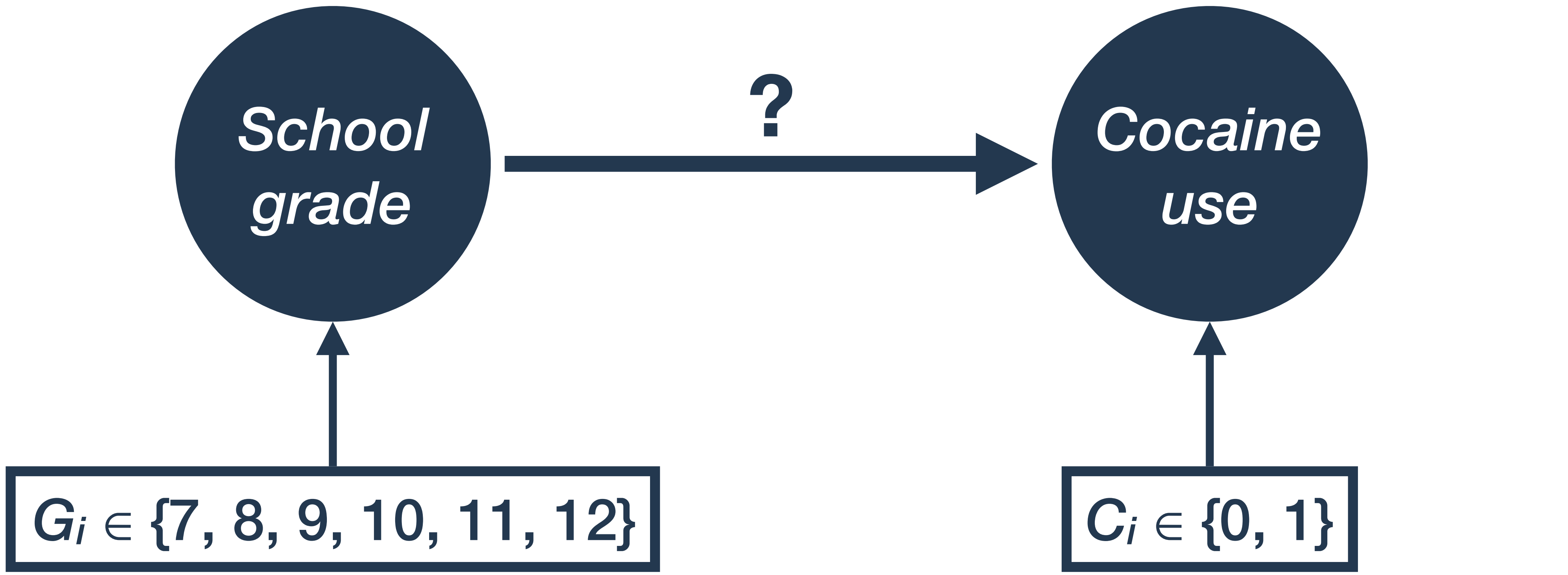
Agenda

1. The trouble with binary outcomes
2. Binomial and Bernoulli distributions
3. Logistic link function
4. Intercept-only logistic regression

Cocaine use among adolescents



Why not use a standard linear regression?

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

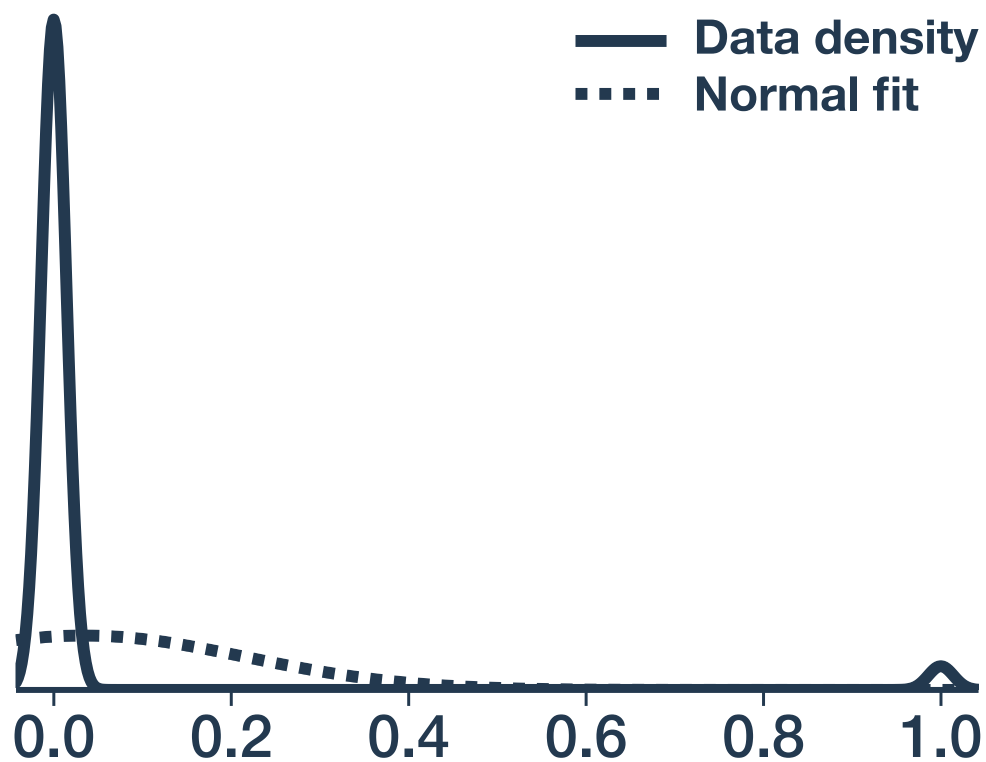
Normal model of binary data

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"



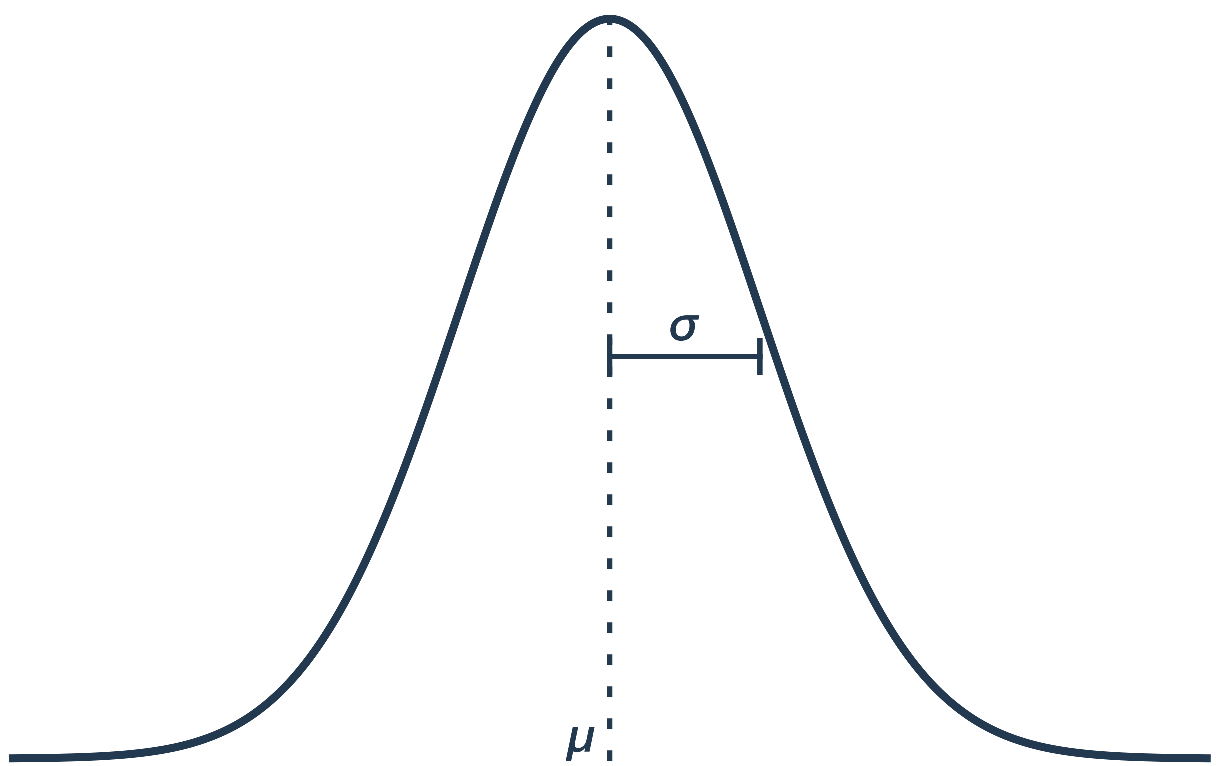
Interpretation

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.

Normal versus Bernoulli

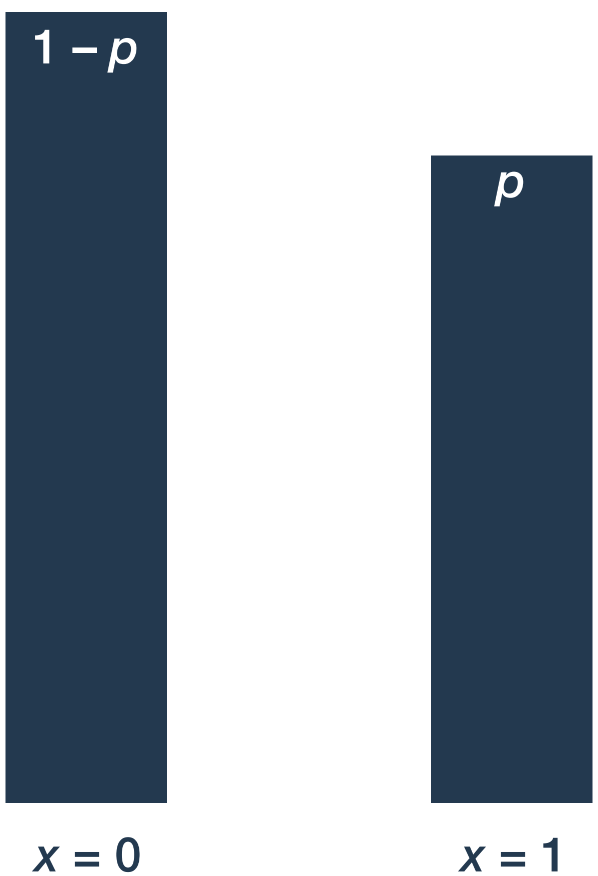
Normal distribution

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



Bernoulli distribution

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

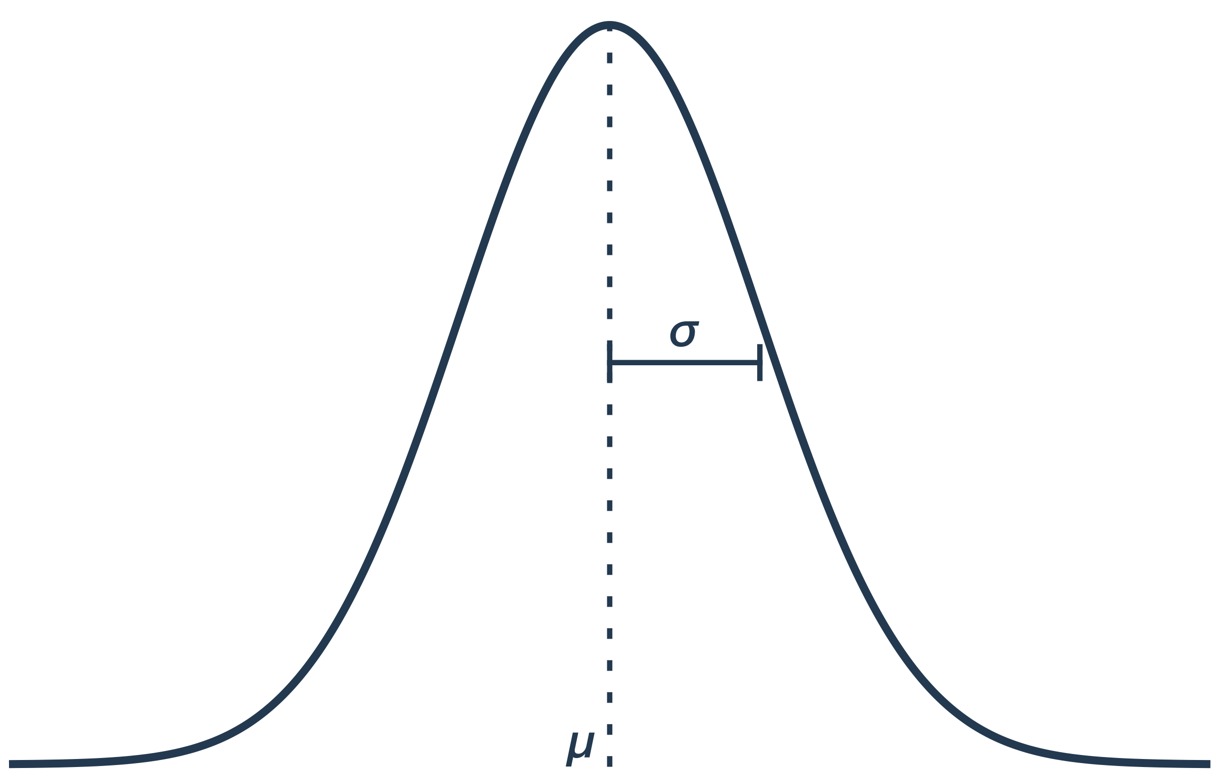


Normal versus Bernoulli

Normal distribution

Norm(μ, σ)

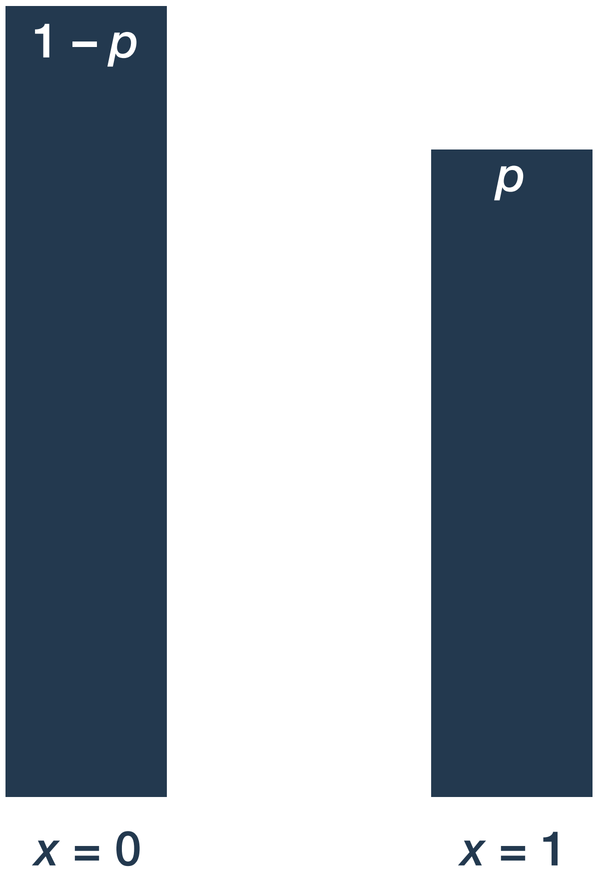
$$\Pr(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Bernoulli distribution

Bernoulli(p) = Binomial(1,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)

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

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

But now we need a "link function"

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

Logistic transformation

Logit function

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

Takes values between 0 and 1, and
returns values between -∞ and ∞.

Logistic function

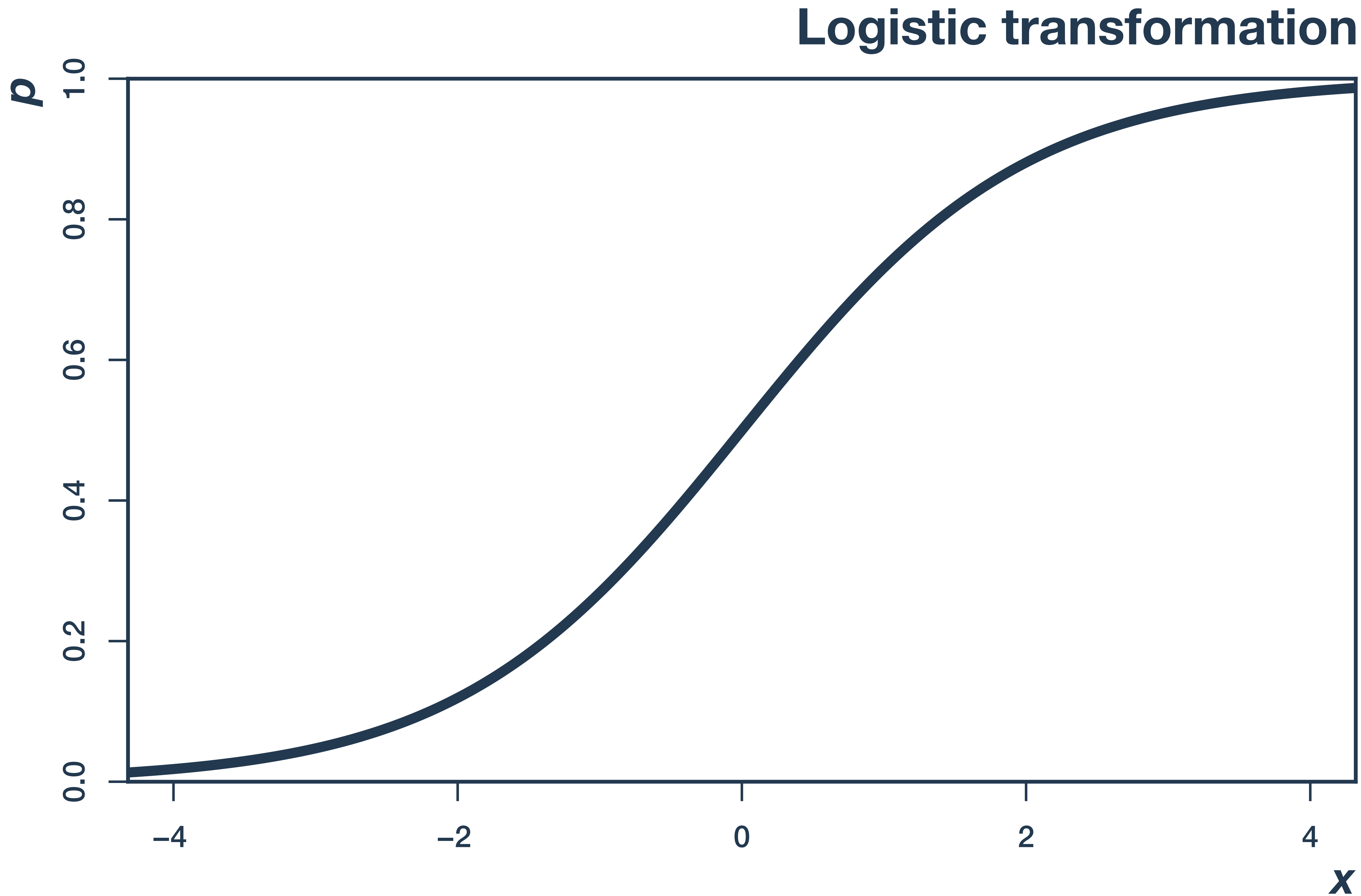
$$\mathrm{logistic}(x) = \frac{e^x}{e^x+1} = \frac{1}{1+e^{-x}}$$

Takes values between -∞ and ∞, and
returns values between 0 and 1.

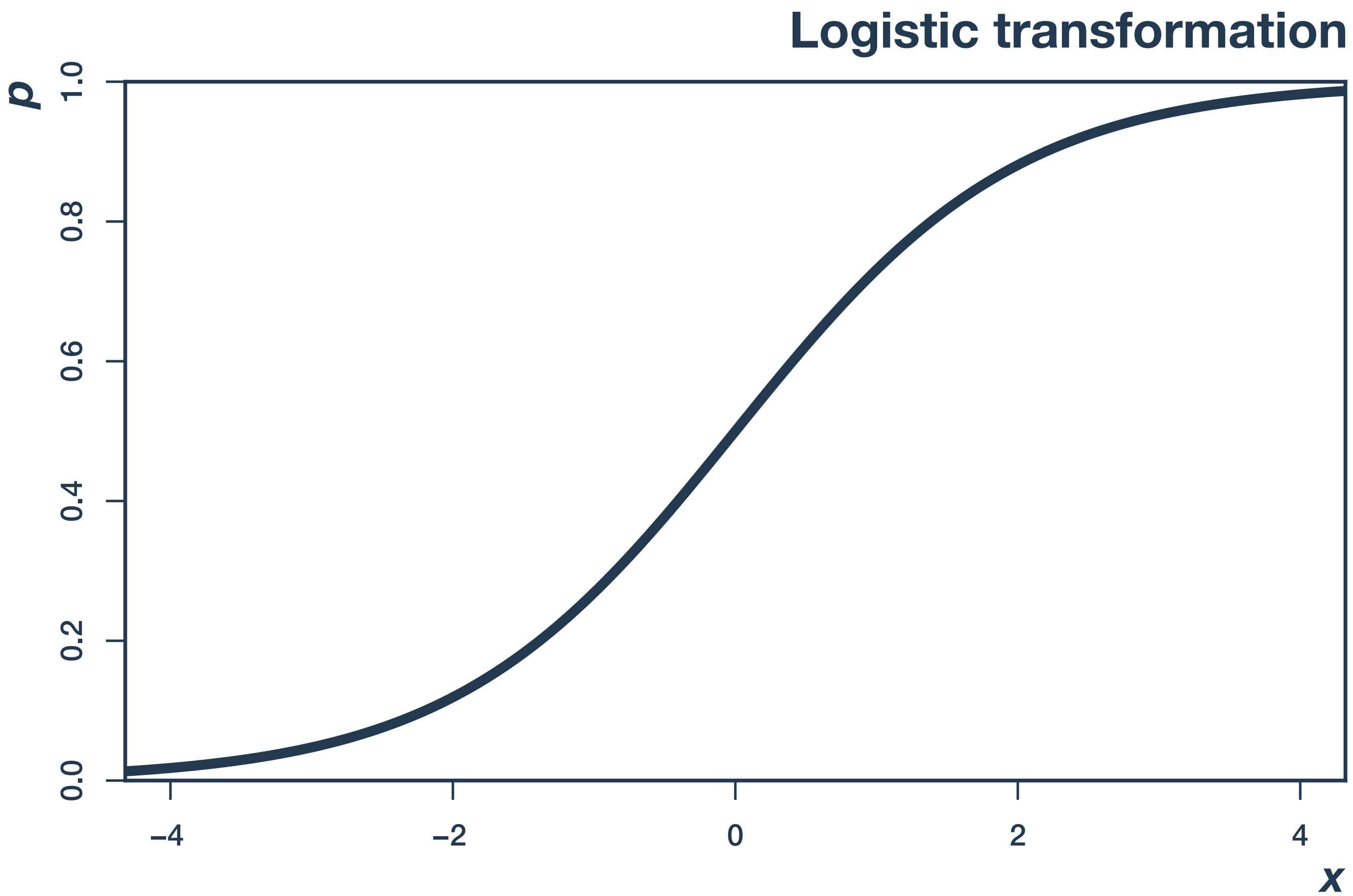
$$C_i \sim \mathrm{Bernoulli}(p_i)$$
$$\mathrm{logit}(p_i) = \alpha + \beta G_i$$

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

Logistic transformation



Logistic transformation



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

x	logistic(x)
-2	0.119
-0.5	0.378
0	0.500
0.5	0.622
2	0.881

Intercept-only logistic model

$$C_i \sim \text{Bernoulli}(p_i)$$
$$\text{logit}(p_i) = \alpha$$

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

$$Y \sim \text{Binom}(n, p)$$
$$p \sim \text{Unif}(0, 1)$$

Logistic regression allows us to include explanatory covariates.

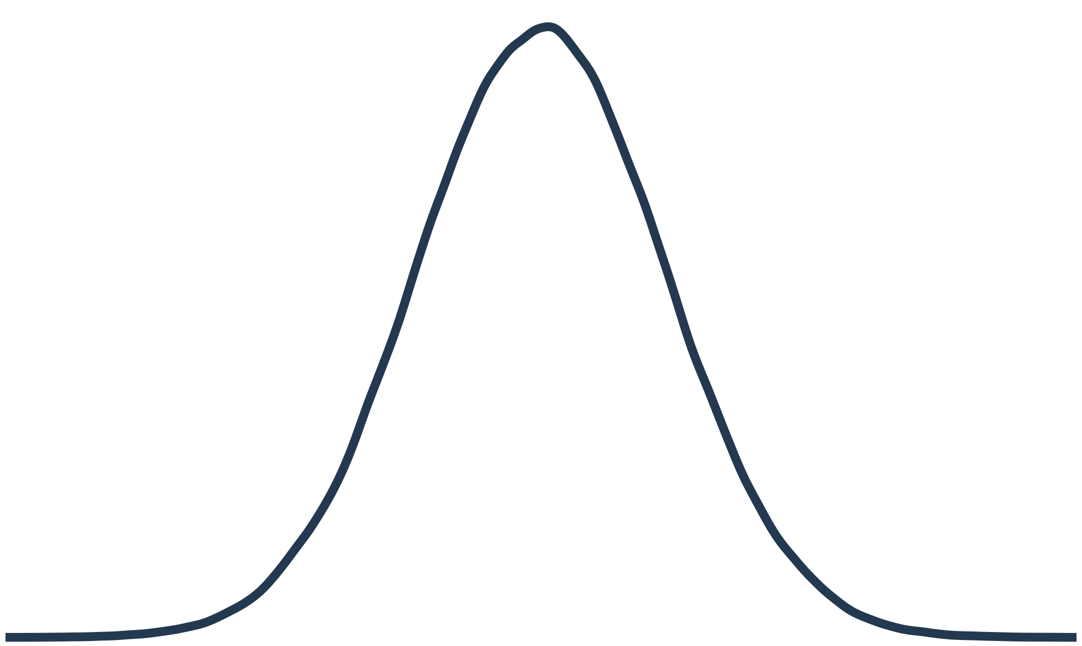
Priors in logistic regression

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

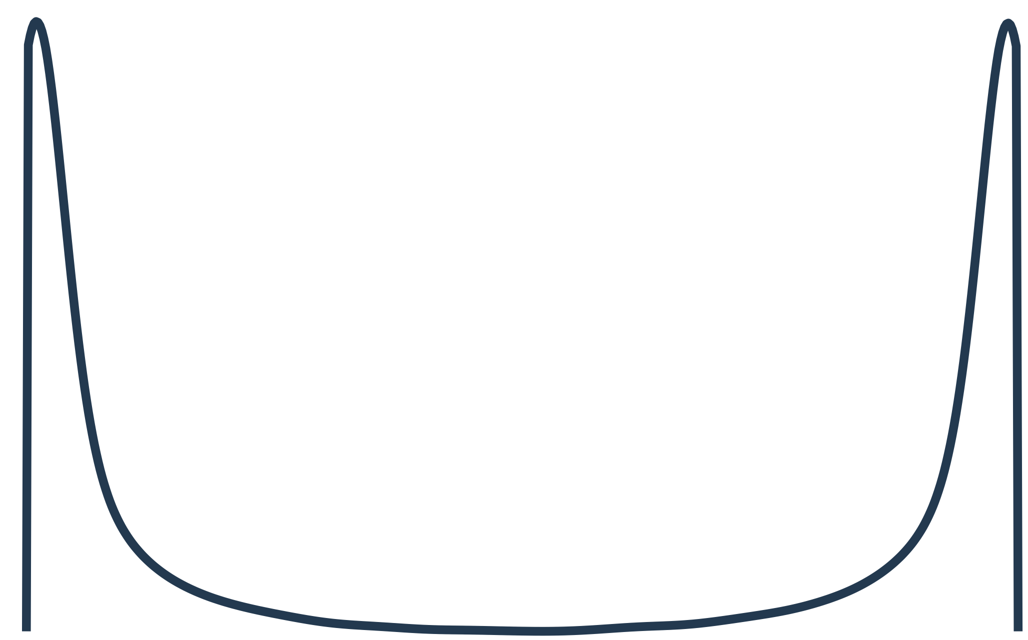
$$a \sim \text{Norm}(0, ?)$$

Priors in logistic regression

$$\alpha \sim \mathrm{Norm}(0, 10)$$



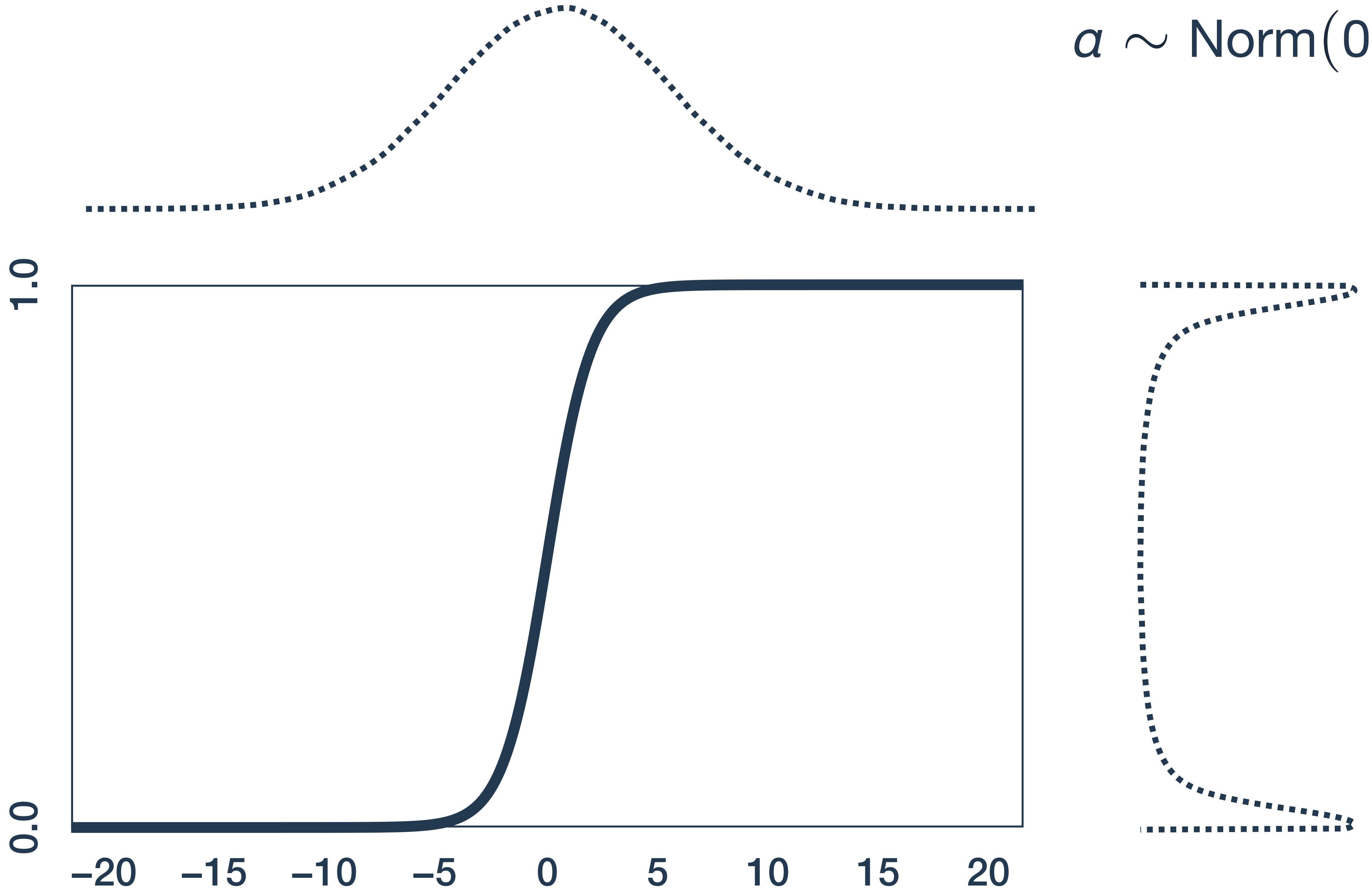
$\Pr(\alpha)$



$\mathrm{logistic}(\Pr(\alpha))$

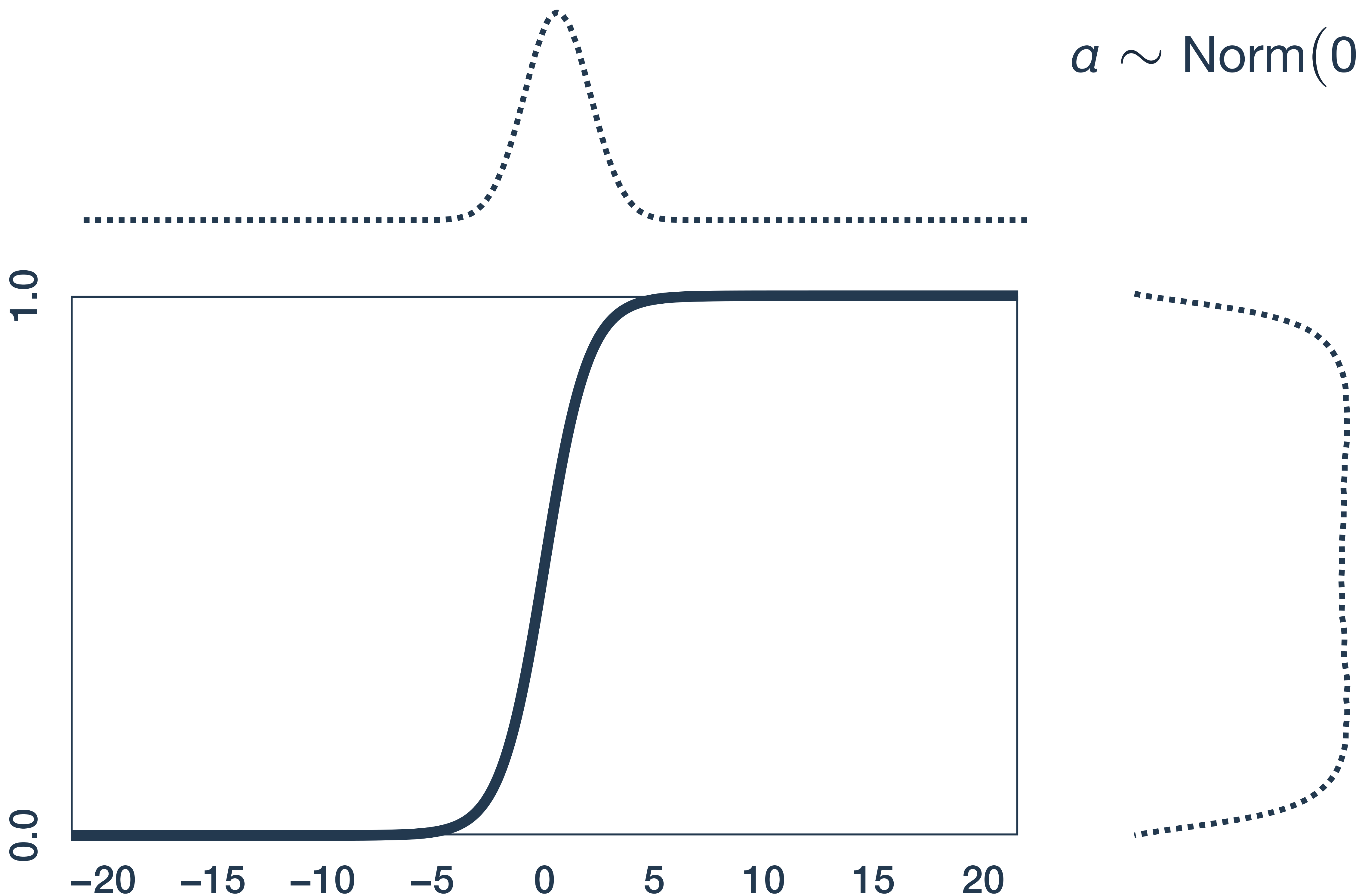
Priors in logistic regression

$$a \sim \mathrm{Norm}(0, 10)$$



Priors in logistic regression

$\alpha \sim \text{Norm}(0, 1.5)$



Intercept-only logistic model

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

$$\text{logit}(p_i) = \alpha$$

	Mean	Std. Dev.
α	-3.334	0.068

n = 6,404

$$\alpha \sim \text{Norm}(0, 1.5)$$

exp(-3.334) = 0.0357 (odds)

logistic(-3.334) = 0.0344 (probability)