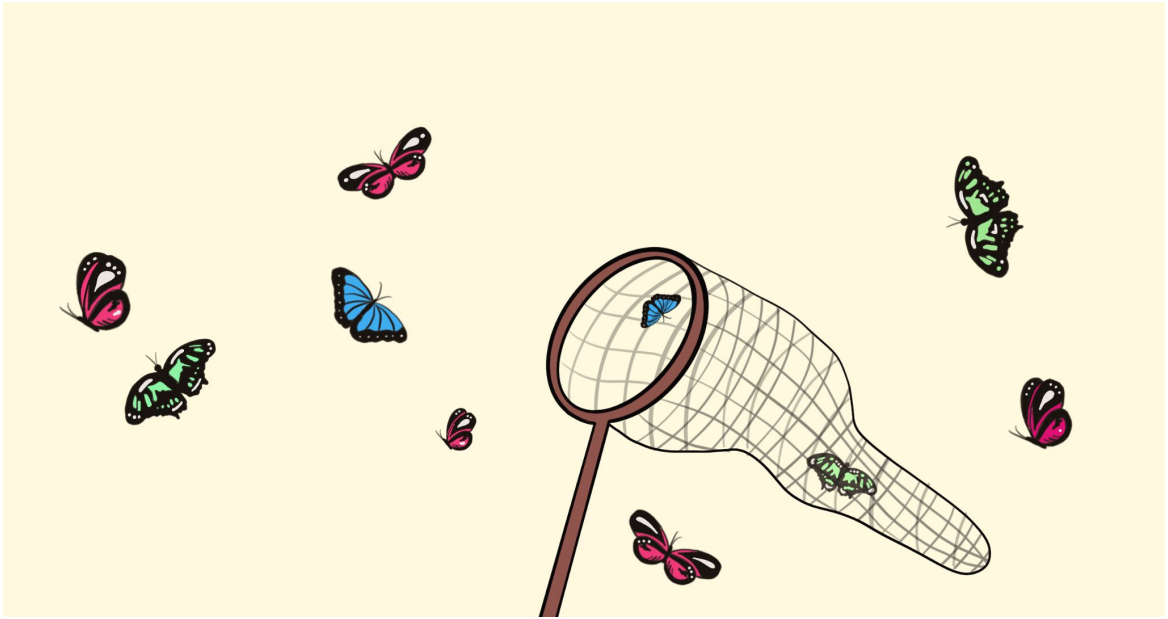
Week 7 Lecture 1:
Logistic regression

EDS 222: *Statistics for Environmental Data Science*

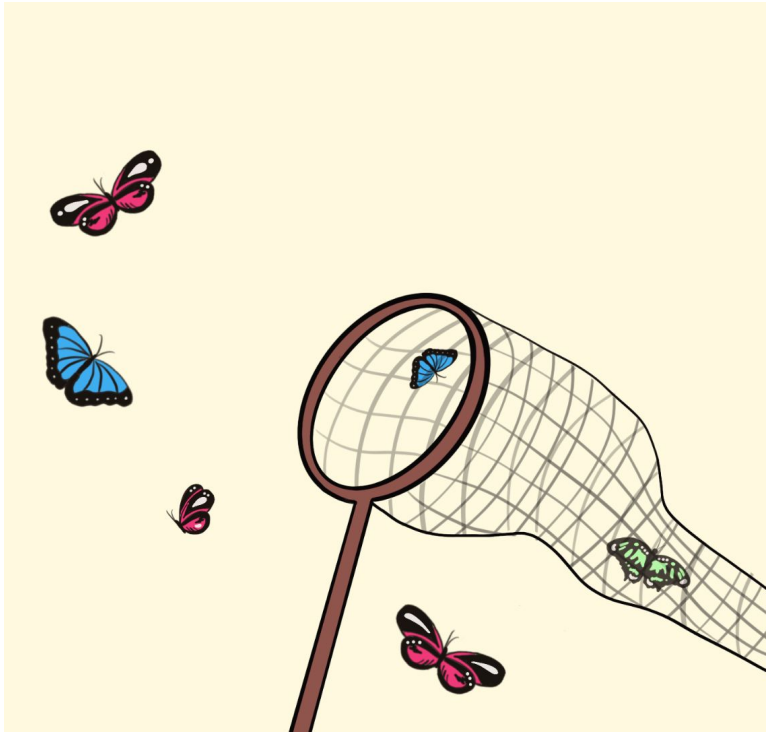


The color of drinking water

DAVID SWITZER AND MANUEL P. TEODORO

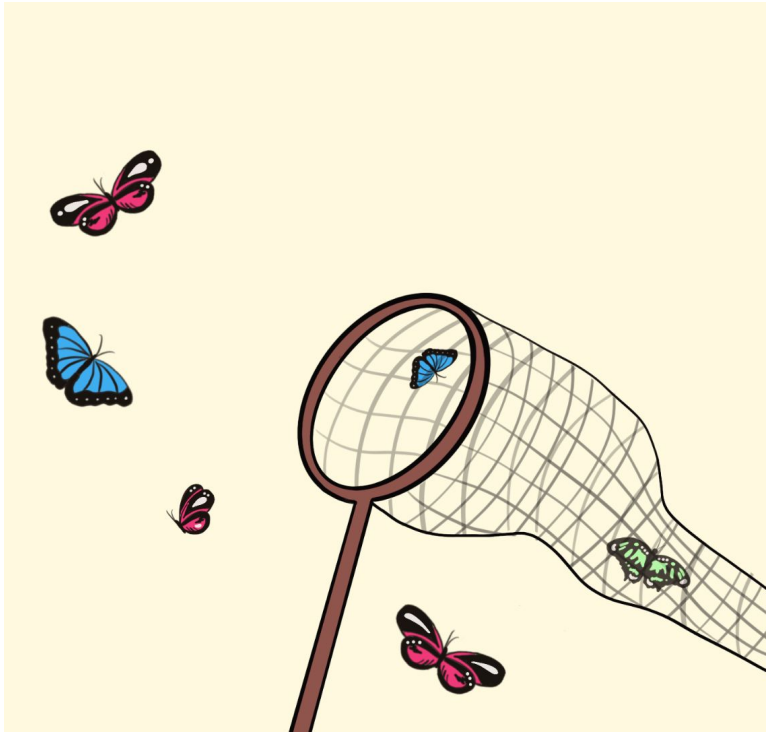
Today’s agenda

- ➔ Revisit hemes 1 and 2
- ➔ Logit link function
- ➔ Exploring coefficients

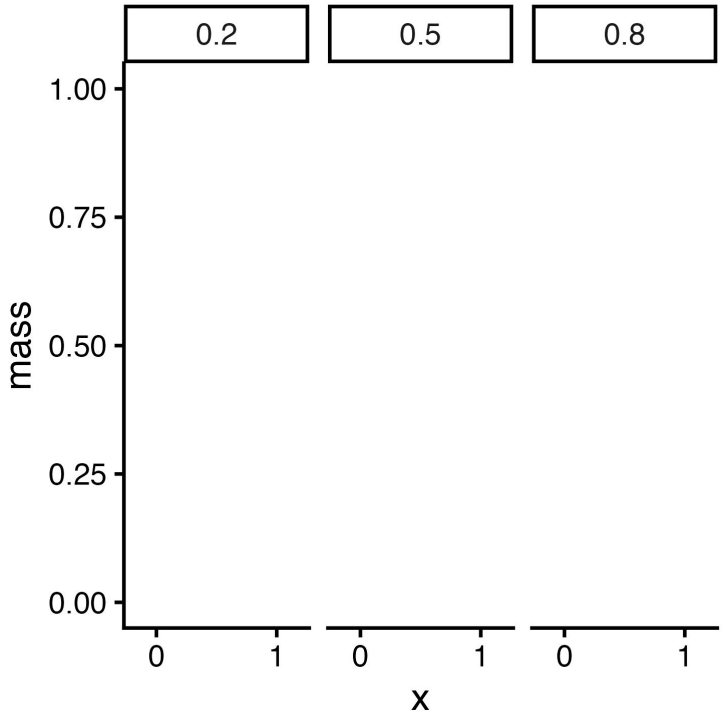


Today’s agenda

- ➔ **Revisit themes 1 and 2**
- ➔ Logit link function
- ➔ Exploring coefficients



Theme 1: binomial variables

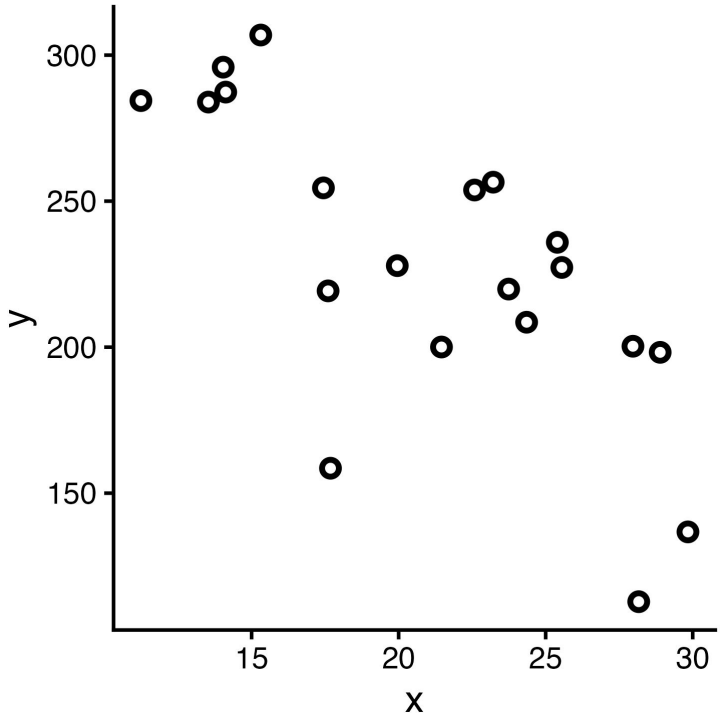


Fill in the PMFs to the left

Each facet should represent a Binomial variable with n=1 and p indicated by the facet header

Compare with a peer

Theme 1: statistical notation



In statistical notation, write the model you would use to find the association between x and y.

Theme 2: DAGs

Consider the following variables in the context of Switzer and Teodoro (2018) and draw a DAG describing what YOU think the causal relationships between them are.

State (categorical)

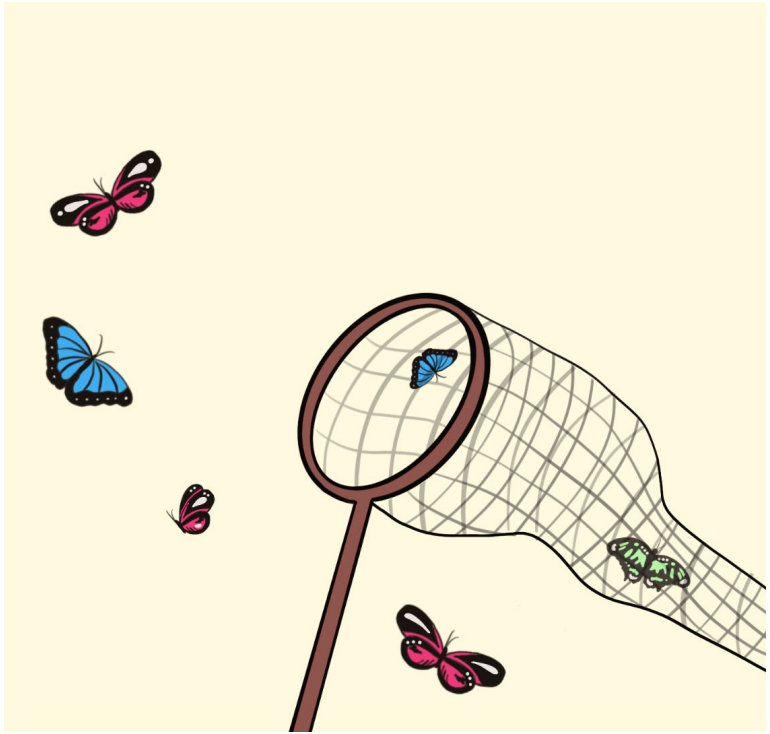
Violation (binary)

PctHisp (continuous)

PctPov (continuous)

Today’s agenda

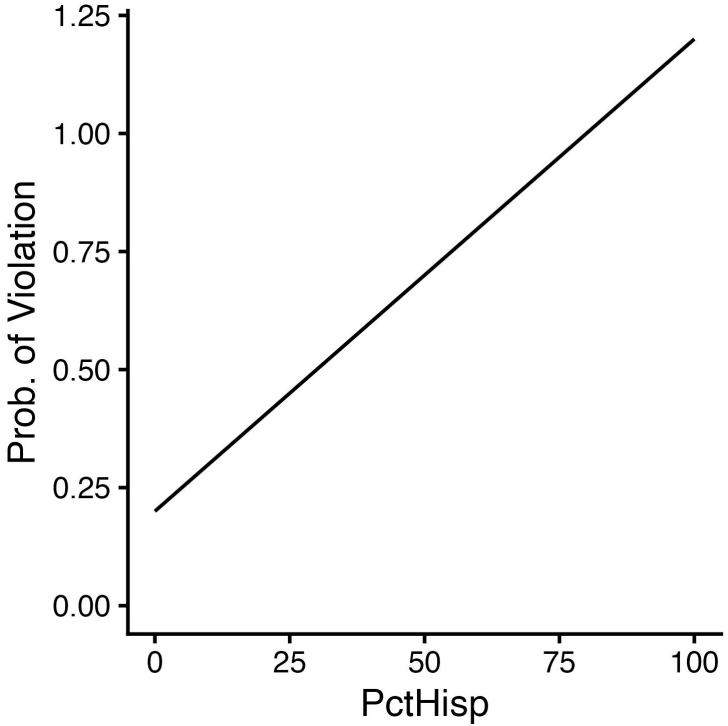
- ➔ Revisit themes 1 and 2
- ➔ **Logit link function**
- ➔ Exploring coefficients



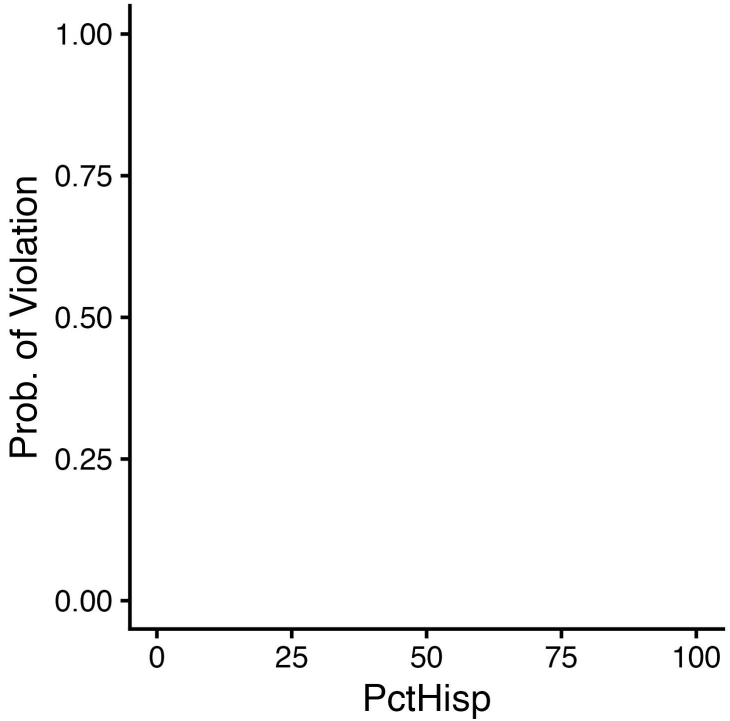
Why didn’t this work?

$$\begin{aligned} \text{violation} &\sim Binomial(1, p) \\ p &= \beta_0 + \beta_1 \text{PctHisp} \end{aligned}$$

Constrained parameters

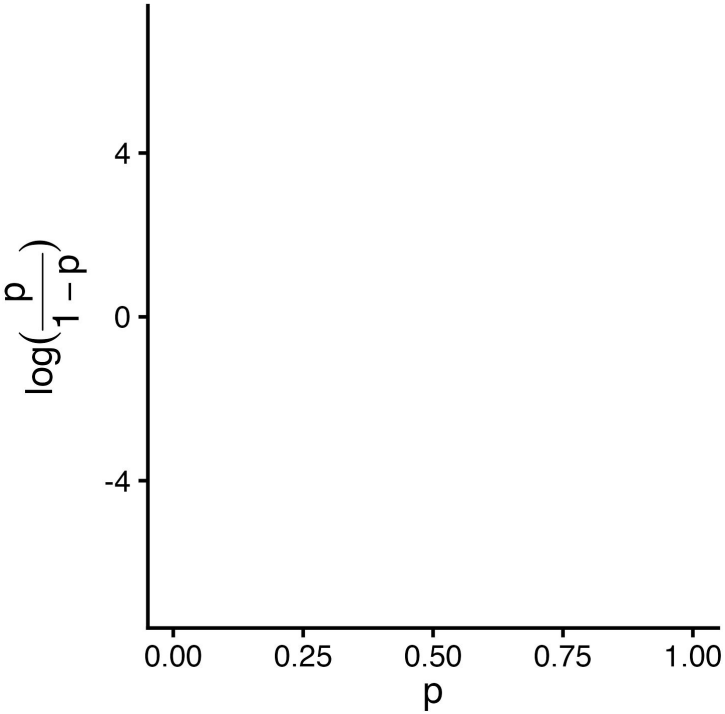


What does p look like?



Draw what you think the line for **p** (Prob. of Violation) should look like, given a positive association with PctHisp and the constraints placed on **p**

Meet the *logit* function



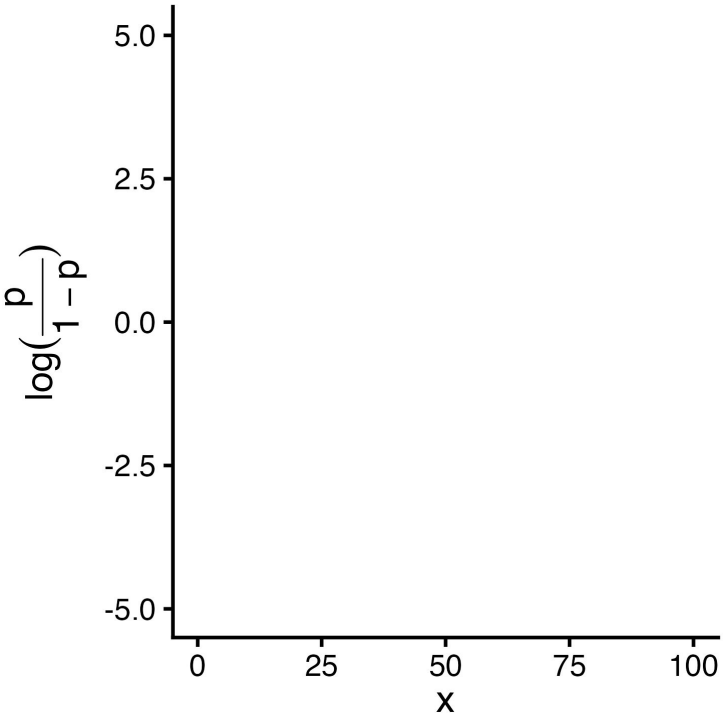

```
p <- seq(0.001,
         0.999,
         length.out = 1e4)
logit_p <- log(p / (1 - p))
ggplot(tibble(p, logit_p),
       aes(p, logit_p)) +
  geom_line()
```

logit in statistical notation

How does the *logit* function help?

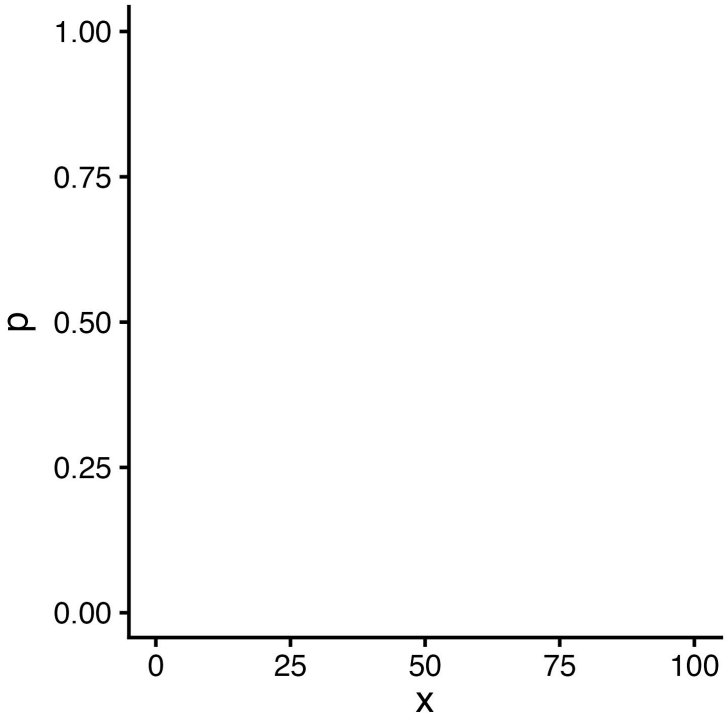
$$\begin{aligned} \text{violation} &\sim Binomial(1, p) \\ p &= \beta_0 + \beta_1 \text{PctHisp} \end{aligned}$$

x and logit(p) is linear




```
x <- seq(0,
         100,
         length.out = 1e4)
beta0 <- -5
beta1 <- 0.1
logit_p <- beta0 + beta1 * x
ggplot(tibble(x, logit_p),
       aes(x, logit_p)) +
  geom_line()
```

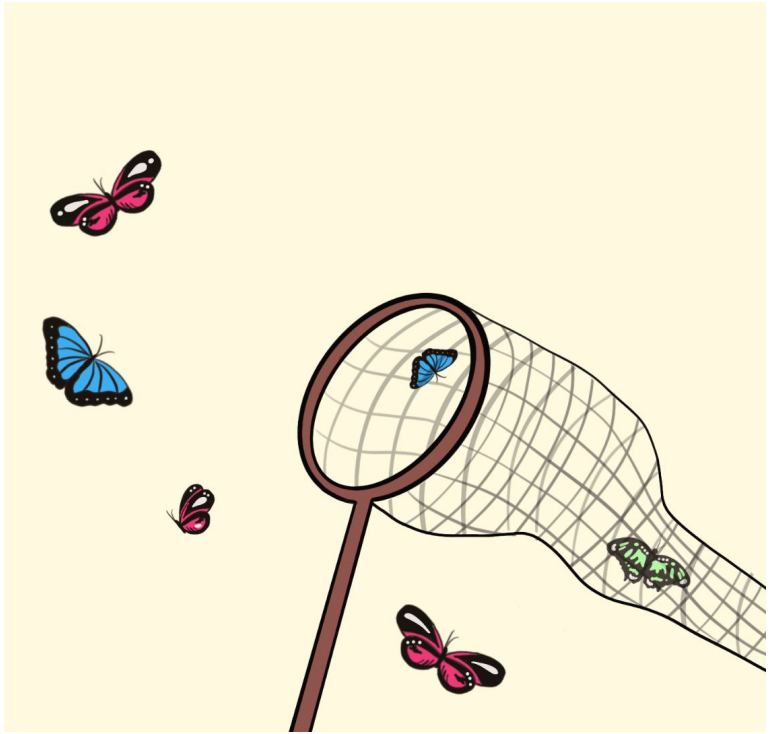
x and p is not




```
p <- exp(logit_p) /
  (1 + exp(logit_p))
ggplot(tibble(x, p),
       aes(x, p)) +
  geom_line()
```

Today’s agenda

- ➔ Revisit themes 1 and 2
- ➔ Logit link function
- ➔ **Exploring coefficients**



What □s reflect your hypothesis?

```
PctHisp <- seq(0, 100, length.out = 1e3)
beta0 <- PICK_ME
beta1 <- CHOOSE_ME
logit_p <- beta0 + beta1 * PctHisp
p <- exp(logit_p) / (1 + exp(logit_p))
ggplot(tibble(PctHisp, p),
       aes(PctHisp, p)) +
  geom_line()
```

$$\text{Violation} \sim Binomial(1, p)$$
$$logit(p) = \beta_0 + \beta_1 \text{PctHisp}$$

What combination of □s produce a line that resembles your hypothesis from earlier?

Interpreting □s

Increase $□_1$ by 0.5.

How did **p** change?

Flip the sign of $□_1$.

How did **p** change?

Decrease $□_0$ by 2.

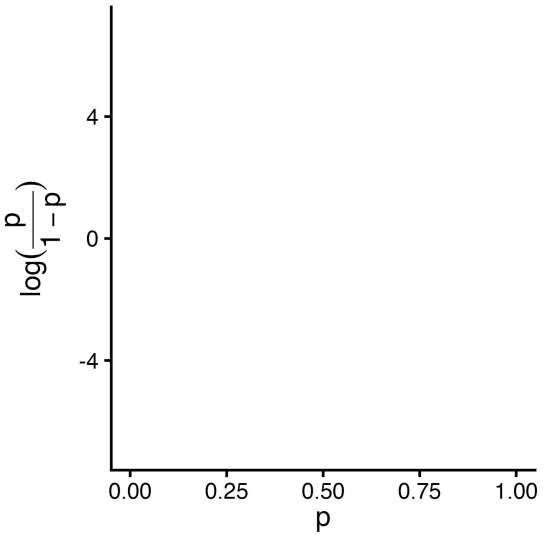
How did **p** change?

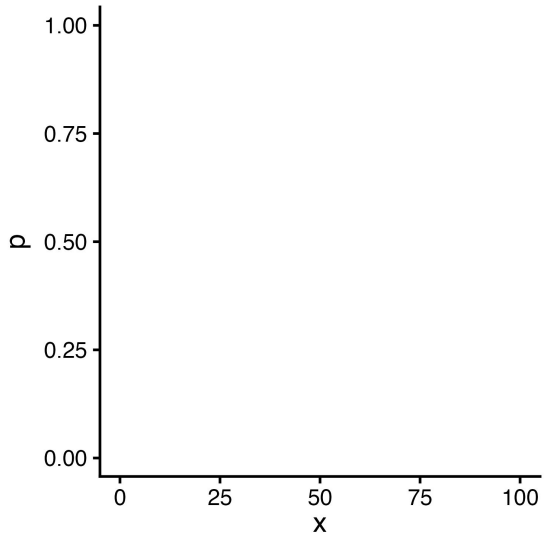
Flip the sign of $□_0$.

How did **p** change?

Recap

$$\text{violation} \sim Binomial(1, p)$$
$$p = \beta_0 + \beta_1 \text{PctHisp}$$





Next time

No lab tomorrow (Veteran's Day)

Wednesday will be like lab, come ready to code

Complete sections **Logistic model notation and R code** and **Read and explore the data** before class

Finish the lab outside of class and share your final answer on Slack to get an additional token