Logistic Regression (and other nonlinear models)

EDS 222

Tamma Carleton
Fall 2023

Announcements/check-in

- Assignment 03 pass/fail, due **today** (5pm)

Announcements/check-in

- Assignment 03 pass/fail, due **today** (5pm)
- Assignment 04 after we cover inference/uncertainty (likely assigned next week)

Announcements/check-in

- Assignment 03 pass/fail, due **today** (5pm)
- Assignment 04 after we cover inference/uncertainty (likely assigned next week)
- Final project proposals, due 11/10 (5pm)
    - More details in a few slides

Final project

Goal:

Apply **some of** the statistical concepts you have learned in this course to **answer an environmental data science question**.*

Final project

Goal:

Apply **some of** the statistical concepts you have learned in this course to **answer an environmental data science question**.*

Two parts:

Deliverable 1: Technical blog post. Some examples:

- G-FEED
- emLab
- MEDS '22, ex. 1
- MEDS '22, ex. 2
- MEDS '22, ex. 3

Final project

Goal:

Apply **some of** the statistical concepts you have learned in this course to **answer an environmental data science question**.*

Final project

Goal:

Apply **some of** the statistical concepts you have learned in this course to **answer an environmental data science question**.[*]

Two parts:

Deliverable 2: Three-minute in-class presentation during final exam slot (4-7pm, 12/12)

[*]: Your project *must* include concepts from the second half of the course.

Final project

Proposal:

Short paragraph (4-5 sentences) describing your proposed project. Motivate the question, describe possible data sources, suggest possible analyses.

Email Sandy your proposal at sandysum@ucsb.edu by 5pm on November 10th.

Final project

Full guidelines on our Resources Page

Some example topics:

- Are political views on climate change associated with recent natural disaster exposure?

Final project

Full guidelines on our Resources Page

Some example topics:

- Are political views on climate change associated with recent natural disaster exposure?
- Detecting trends in water quality for indigenous communities in Chile

Final project

Full guidelines on our Resources Page

Some example topics:

- Are political views on climate change associated with recent natural disaster exposure?
- Detecting trends in water quality for indigenous communities in Chile
- Spatial patterns of deforestation during COVID-19

Final project

Full guidelines on our Resources Page

Some example topics:

- Are political views on climate change associated with recent natural disaster exposure?
- Detecting trends in water quality for indigenous communities in Chile
- Spatial patterns of deforestation during COVID-19
- Are there gendered health effects of wildfire smoke?

Today

More on nonlinear relationships with linear regression models

Log-linear, log-log regressions

Today

More on nonlinear relationships with linear regression models

Log-linear, log-log regressions

Logistic regression

How do we model binary outcomes?

Nonlinear relationships in linear regression models

Nonlinear transformations

- Our linearity assumption requires that **parameters enter linearly** (*i.e.*, the β_k multiplied by variables)
- We allow nonlinear relationships between y and the explanatory variables x.

Example: Polynomials

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + u_i$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + u_i$$

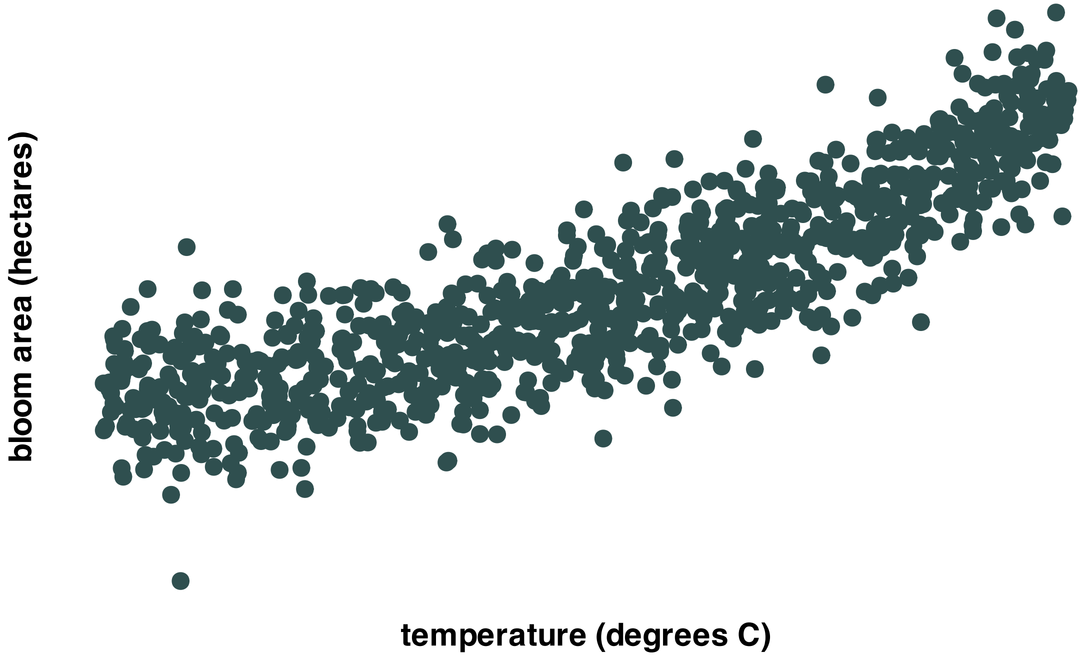
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + u_i$$

...

Polynomials

- Recall the relationship between **temperature** and **harmful algal blooms**:

$$area_i = \beta_0 + \beta_1 temperature_i + \beta_2 temperature_i^2 + u_i$$



Polynomials

Estimating polynomial regressions in `R`:

```
blooms_df = blooms_df %>% mutate(temp2 = temp^2)
summary(lm(area~temp+temp2, data=blooms_df))
#> 
#> Call:
#> lm(formula = area ~ temp + temp2, data = blooms_df)
#> 
#> Residuals:
#>     Min      1Q  Median      3Q     Max 
#> -12.597  -2.092  -0.142   1.995   9.487 
#> 
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)    
#> (Intercept)   0.0636     0.2925    0.22     0.83    
#> temp          0.6254     0.4401    1.42     0.16    
#> temp2         1.9212     0.1416   13.57   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> 
#> Residual standard error: 3.02 on 997 degrees of freedom
#> Multiple R-squared:  0.777,    Adjusted R-squared:  0.777
```

Other nonlinear-in-X regressions

- **Polynomials** and **interactions:** $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{2i} + \beta_4 x_{2i}^2 + \beta_5 (x_{1i} x_{2i}) + u_i$ (more on this today)
- **Exponentials** $\log(y_i) = \beta_0 + \beta_2 e^{x_{2i}} + u_i$
- **Logs:** $\log(y_i) = \beta_0 + \beta_1 x_{1i} + u_i$ (Today!)
- **Indicators** and **thresholds:** $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 \mathbb{I}(x_{1i} \geq 100) + u_i$

Other nonlinear-in-X regressions

- **Polynomials** and **interactions:** $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{2i} + \beta_4 x_{2i}^2 + \beta_5 (x_{1i} x_{2i}) + u_i$ (more on this today)
- **Exponentials** $\log(y_i) = \beta_0 + \beta_2 e^{x_{2i}} + u_i$
- **Logs:** $\log(y_i) = \beta_0 + \beta_1 x_{1i} + u_i$ (Today!)
- **Indicators** and **thresholds:** $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 \mathbb{I}(x_{1i} \geq 100) + u_i$

In all cases, the effect of a change in x on y will vary depending on your baseline level of x. This is not true with linear relationships!

Log-linear specification

You will frequently see logged* outcome variables with linear (non-logged) explanatory variables, *e.g.*,

$$\log(\mathbf{area}_i) = \beta_0 + \beta_1\,\mathbf{temperature}_i + u_i$$

This specification changes our interpretation of the slope coefficients.

Log-linear specification

You will frequently see logged[*] outcome variables with linear (non-logged) explanatory variables, *e.g.*,

$$\log(\mathbf{area}_i) = \beta_0 + \beta_1 \, \mathbf{temperature}_i + u_i$$

This specification changes our interpretation of the slope coefficients.

Interpretation

- A one-unit increase in our explanatory variable increases the outcome variable by approximately $\beta_1 \times 100$ percent.

- *Example:* If $\beta_1 = 0.03$, an additional degree of warming increases algal bloom area by approximately 3 percent.

[*]: When I say "log", I mean "natural log", i.e. $ln(x) = log_e(x)$.

Review: Percent changes

- What is a percent change again, anyway?

Review: Percent changes

- What is a percent change again, anyway?
- Local gasoline prices were $5/gallon, but last month increased by 12%. How much are they now?

Review: Percent changes

- What is a percent change again, anyway?
- Local gasoline prices were $5/gallon, but last month increased by 12%. How much are they now?

$$5(1 + 0.12) = 5 \times 1.12 = 5.6$$

Review: Percent changes

- What is a percent change again, anyway?

- Local gasoline prices were \$5/gallon, but last month increased by 12%. How much are they now?

$$\mathbf{5(1+0.12) = 5 \times 1.12 = 5.6}$$

Can also write this as

$$\mathbf{0.12 = \frac{5.6 - 5}{5}}$$

Review: Percent changes

- What is a percent change again, anyway?
- Local gasoline prices were \$5/gallon, but last month increased by 12%. How much are they now?

$$5(1+0.12) = 5 \times 1.12 = 5.6$$

Can also write this as

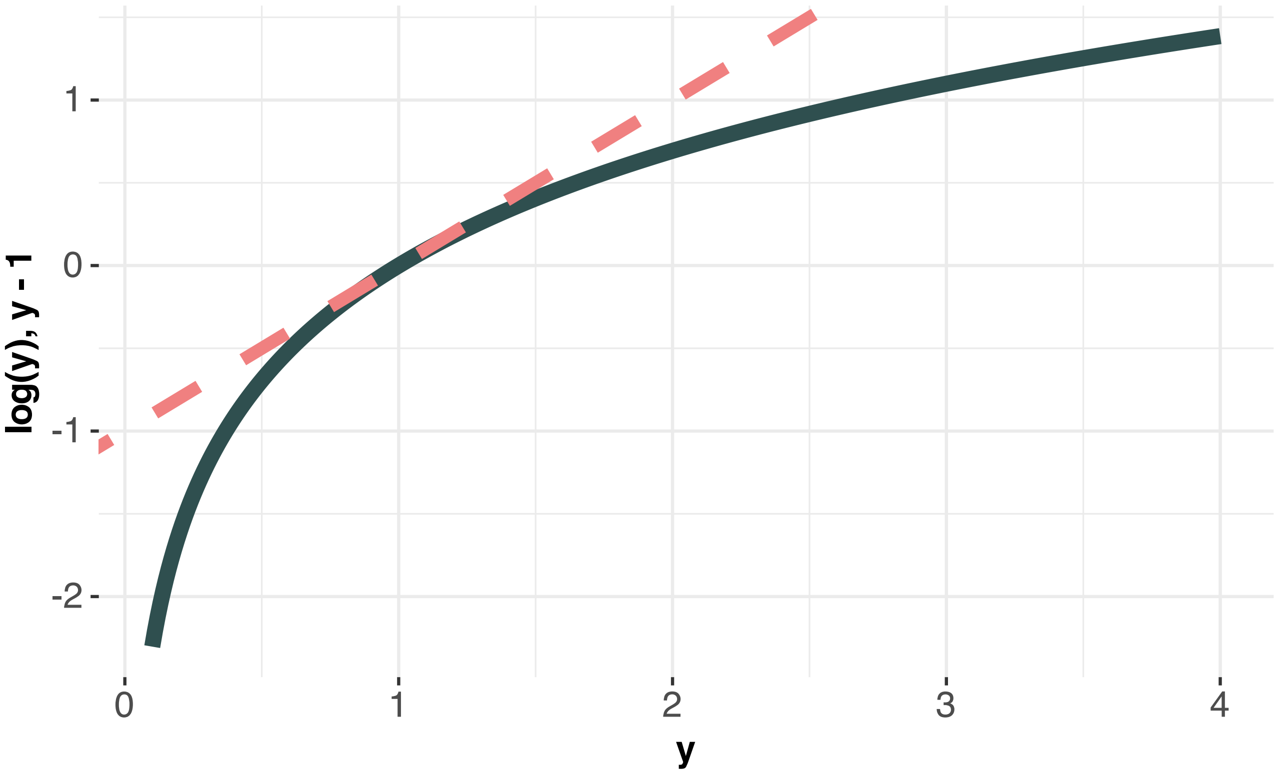
$$0.12 = \frac{5.6-5}{5}$$

Generally, we have that when y increases by r percent, our new value is $y(1+r)$.

$$r = \frac{y_2 - y_1}{y_1}$$

Log differences as percent changes?

Near $y = 1$, $log(y)$ is approximately slope 1, i.e. $log(y) \approx y - 1$



Log differences as percent changes?

Near $y = 1$, $log(y)$ is approximately slope 1, i.e. $log(y) \approx y - 1$

Therefore, $log(1 + r) \approx r$ **when r is small!** (so that you're still close to 1 on the x-axis)

Log differences as percent changes?

Near $y = 1$, $log(y)$ is approximately slope 1, i.e. $log(y) \approx y - 1$

Therefore, $log(1 + r) \approx r$ **when r is small!** (so that you're still close to 1 on the x-axis)

This lets us show that:

$$log(y(1 + r)) = log(y) + log(1 + r) \approx log(y) + r$$

So when we see $log(y)$ go up by r, we can say that represents an $r \times 100$ percent change in y!

Log differences as percent changes?

Near $y = 1$, $log(y)$ is approximately slope 1, i.e. $log(y) \approx y - 1$

Therefore, $log(1 + r) \approx r$ **when r is small!** (so that you're still close to 1 on the x-axis)

This lets us show that:

$$log(y(1 + r)) = log(y) + log(1 + r) \approx log(y) + r$$

So when we see $log(y)$ go up by r, we can say that represents an $r \times 100$ percent change in y!

For example: y is increased by 5% means y increases to $y(1.05)$. The log of y changes from $log(y)$ to approximately $log(y) + 0.05$. Increasing y by 5% is therefore (almost) equivalent to adding 0.05 to $log(y)$.

Log-linear specification

Back to our log-linear model

$$\log(y_i) = \beta_0 + \beta_1\, x_i + u$$

A one unit change in x causes a β_1 unit change in $log(y)$.

This is equivalent to a β_1 **percentage change** in y.

Log-linear specification

Because the log-linear specification comes with a different interpretation, you need to make sure it fits your data-generating process/model.

Does x change y in levels (*e.g.*, a 3-unit increase) or percentages (*e.g.*, a 10-percent increase)?

Log-linear specification

Because the log-linear specification comes with a different interpretation, you need to make sure it fits your data-generating process/model.

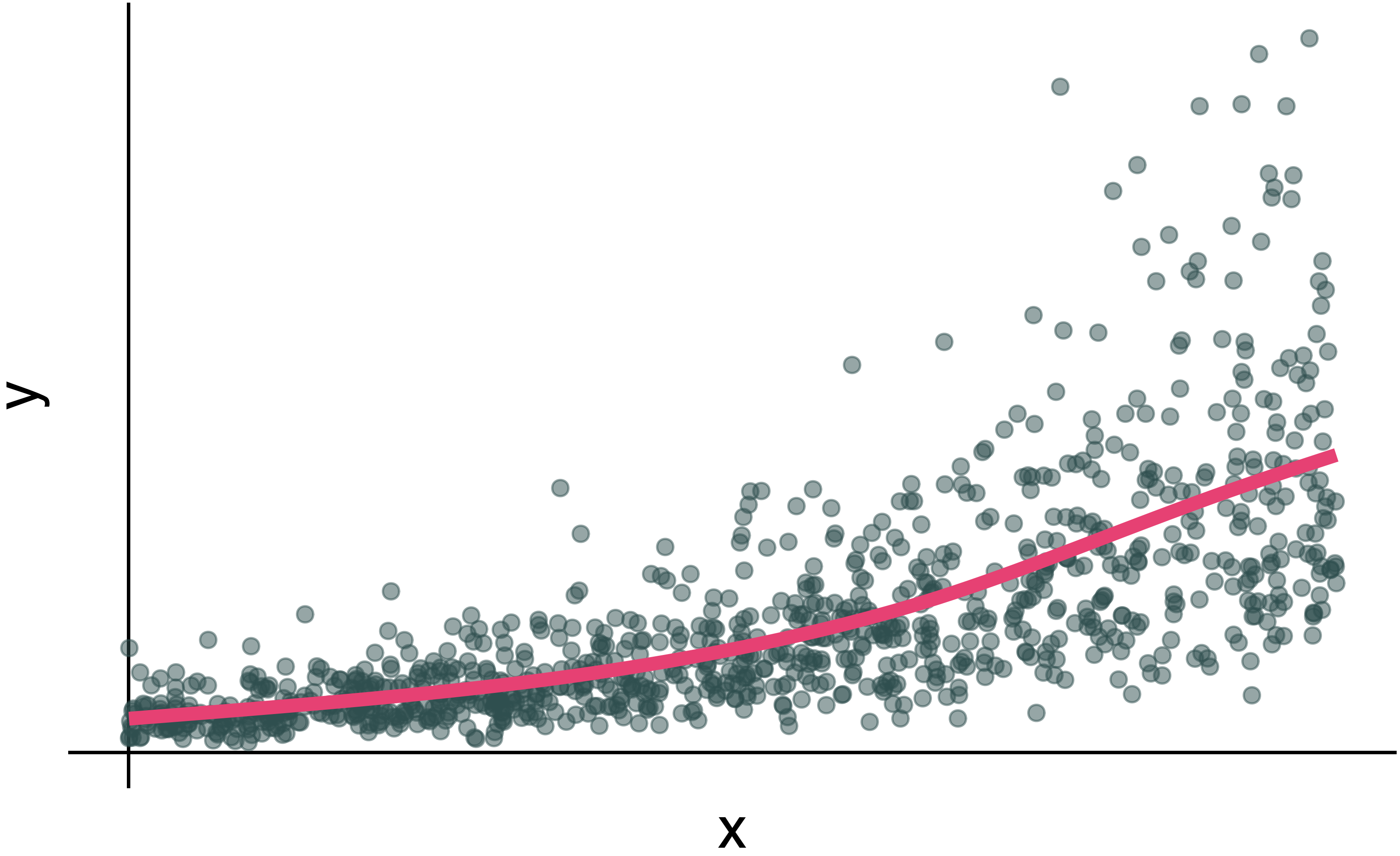
Does x change y in levels (*e.g.*, a 3-unit increase) or percentages (*e.g.*, a 10-percent increase)?

I.e., you need to be sure an exponential relationship makes sense:

$$\log(y_i) = \beta_0 + \beta_1 x_i + u_i \iff y_i = e^{\beta_0 + \beta_1 x_i + u_i}$$

Note: You are using linear regression to estimate a nonlinear-in-parameters relationship. This is the power of taking logs!

Log-linear specification



Log-log specification

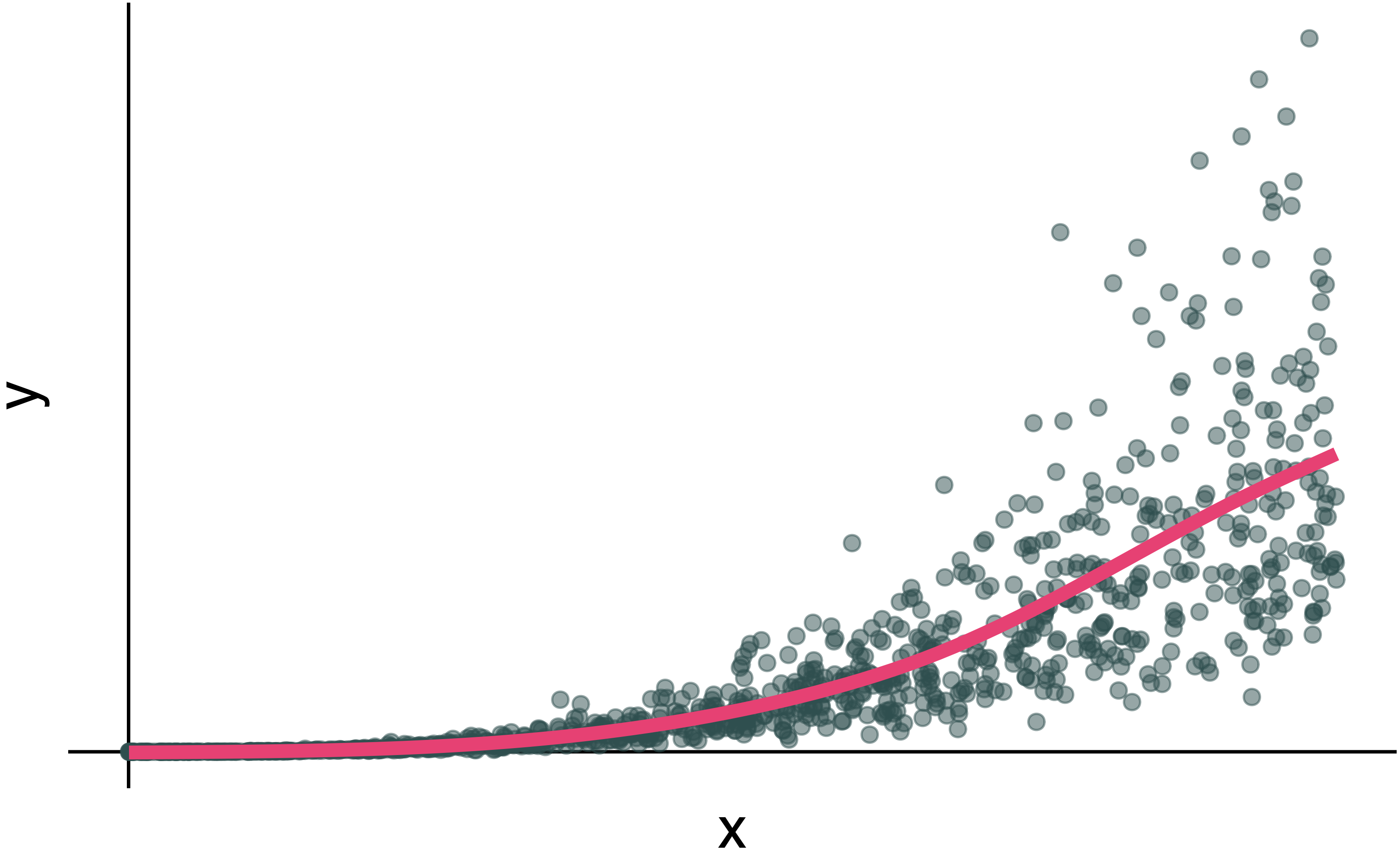
Similarly, log-log models are those where the outcome variable is logged *and* at least one explanatory variable is logged

$$\log(\log_i) = \beta_0 + \beta_1 \log(\text{temperature}_i) + u_i$$

Interpretation:

- A one-percent increase in x will lead to a β_1 percent change in y.
- Often interpreted as an "elasticity" in economics.

Log-log specification



Log-linear with a binary variable

Note: If you have a log-linear model with a binary indicator variable, the interpretation for the coefficient on that variable changes.

Consider:

$$\log(y_i) = \beta_0 + \beta_1 x_{1i} + u_i$$

for binary variable x_1.

The interpretation of β_1 is now

- When x_1 changes from 0 to 1, y will change by $100 \times \left(e^{\beta_1} - 1\right)$ percent.
- When x_1 changes from 1 to 0, y will change by $100 \times \left(e^{-\beta_1} - 1\right)$ percent.

When the approximation fails

The nice interpretation so far relies on the fact that near 1, $log(y) \approx y - 1$

- So, for example, $log(y(1 + r)) = log(y) + log(1 + r) \approx log(y) + r$

When the approximation fails

The nice interpretation so far relies on the fact that near 1, $log(y) \approx y - 1$

- So, for example, $log(y(1 + r)) = log(y) + log(1 + r) \approx log(y) + r$

What if r is large? E.g., r=0.8:

- $log(1 * (1.8)) = log(1) + log(1.8) = 0.59 \neq log(1) + 0.8 = 0.8$

When the approximation fails

The nice interpretation so far relies on the fact that near 1, $log(y) \approx y - 1$

- So, for example, $log(y(1+r)) = log(y) + log(1+r) \approx log(y) + r$

What if r is large? E.g., r=0.8:

- $log(1 * (1.8)) = log(1) + log(1.8) = 0.59 \neq log(1) + 0.8 = 0.8$

Exact percentage change (use for large predicted changes):

If $log(y) = \beta_0 + \beta_1 x + \varepsilon$, then the percentage change in y for a one unit change in x is:

$$\% \textbf{ change in y} = (e^{\beta_1} - 1) \times 100$$

When the approximation fails

The nice interpretation so far relies on the fact that near 1, $log(y) \approx y - 1$

- So, for example, $log(y(1+r)) = log(y) + log(1+r) \approx log(y) + r$

What if r is large? E.g., r=0.8:

- $log(1 * (1.8)) = log(1) + log(1.8) = 0.59 \neq log(1) + 0.8 = 0.8$

Exact percentage change (use for large predicted changes):

If $log(y) = \beta_0 + \beta_1 x + \varepsilon$, then the percentage change in y for a one unit change in x is:

$$\text{\% change in y} = (e^{\beta_1} - 1) \times 100$$

Note that e^x in `R` is `exp(x)`

When the approximation fails

Example: Suppose in $log(y) = \beta_0 + \beta_1 x + \varepsilon$, we estimate that $\hat{\beta}_1 = 0.6$

When the approximation fails

Example: Suppose in $log(y) = \beta_0 + \beta_1 x + \varepsilon$, we estimate that $\hat{\beta}_1 = 0.6$

This looks like a 1 unit change in x causes a 60% change in y. But the exact percentage change in y is:

- $(e^{0.6} - 1) \times 100 = 0.82 \times 100 \implies 82$ percent change in y
- Note that the imprecise approximation for large changes will always be biased *downwards*

When the approximation fails

Example: Suppose in $log(y) = \beta_0 + \beta_1 x + \varepsilon$, we estimate that $\hat{\beta}_1 = \mathbf{0.6}$

This looks like a 1 unit change in x causes a 60% change in y. But the exact percentage change in y is:

- $(e^{\mathbf{0.6}} - 1) \times \mathbf{100} = \mathbf{0.82} \times \mathbf{100} \implies \mathbf{82}$ percent change in y
- Note that the imprecise approximation for large changes will always be biased *downwards*

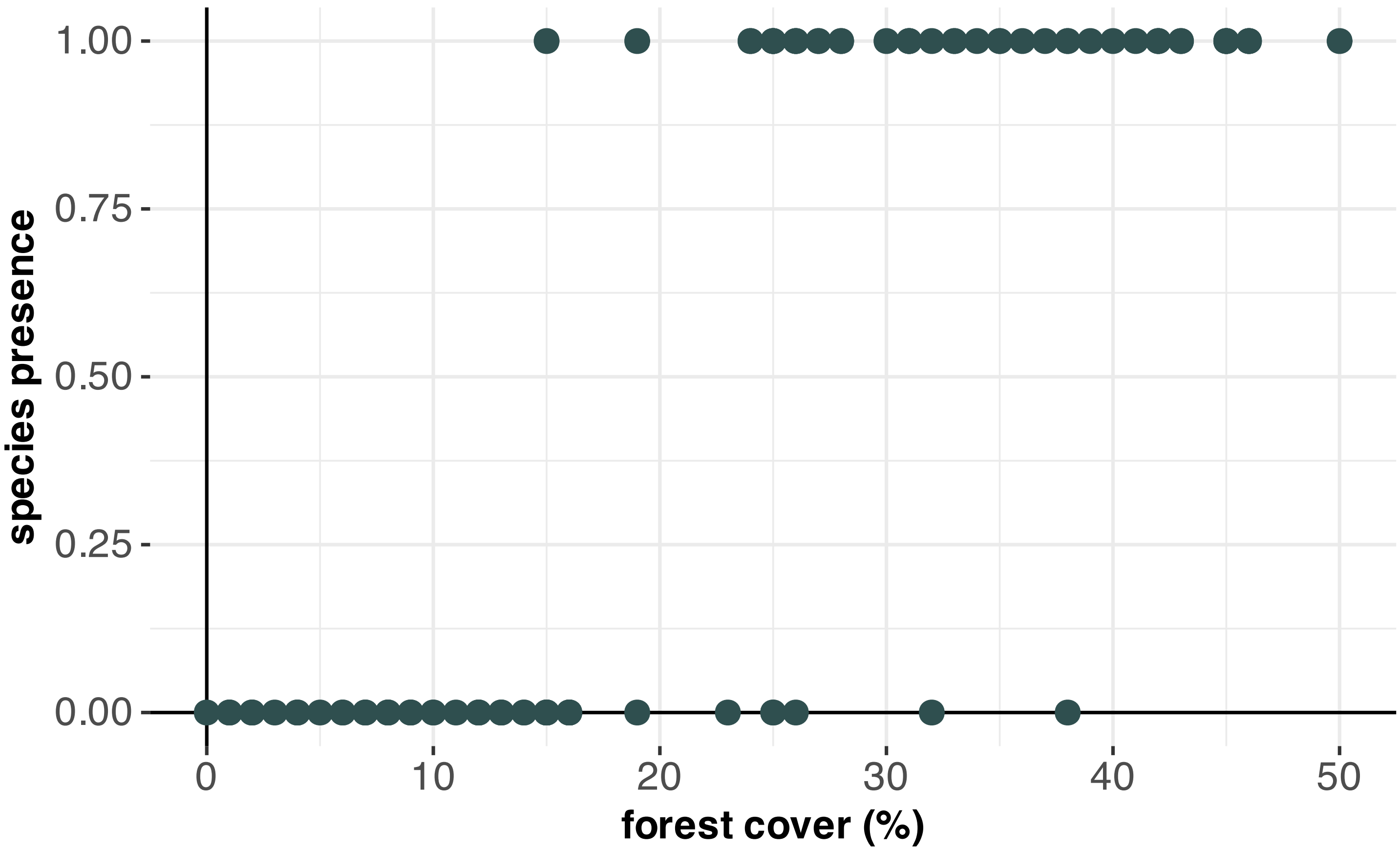
Can you just change units of x?

- Yes, mechanically you can do this and avoid the issues with approximation
- But think hard about your problem! You probably care about understanding the impacts of a meaningful increase in x, not a tiny increase in x

Logistic regression

Modeling binary outcomes

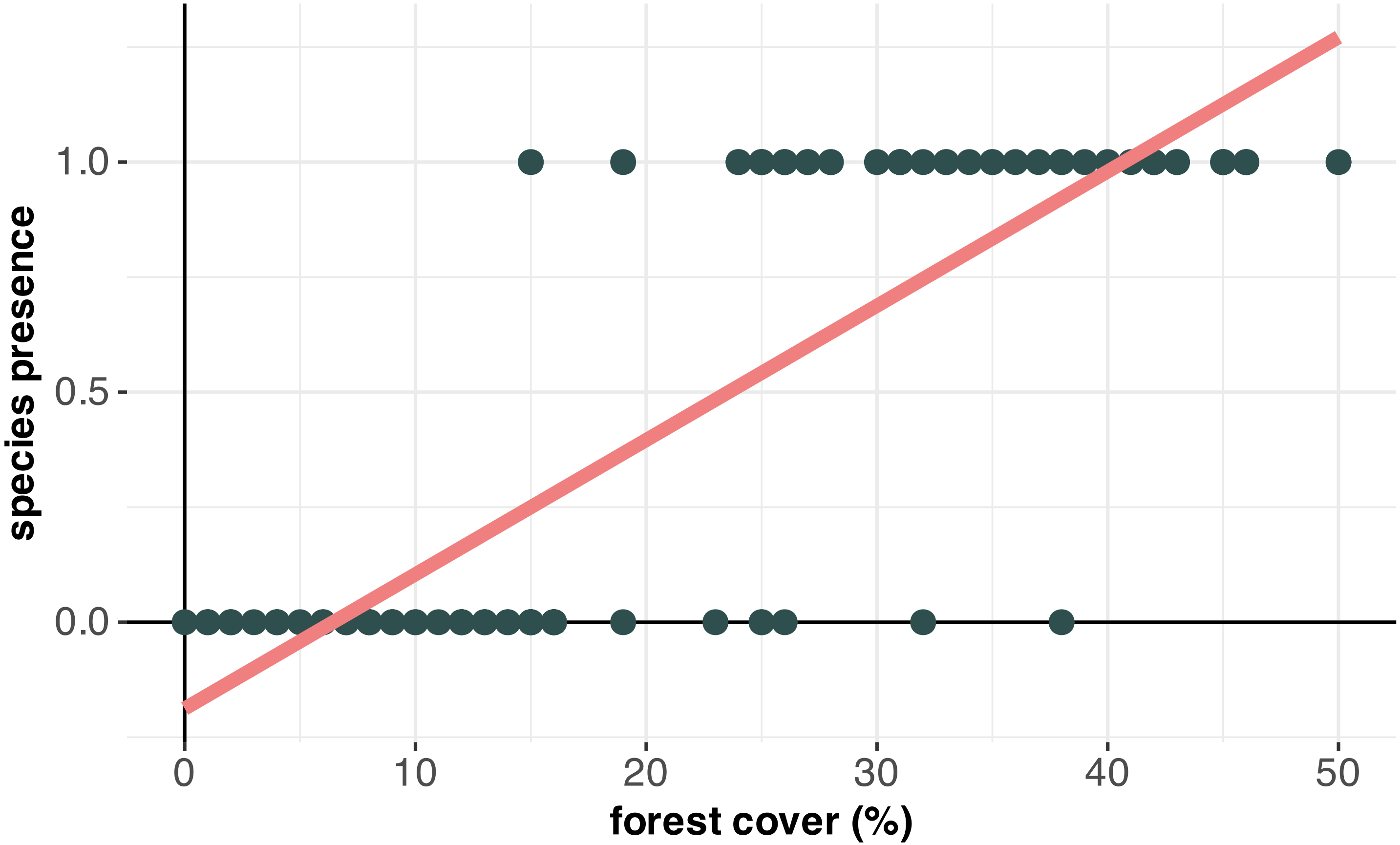
What do you do when your dependent variable takes on just two values?



Modeling binary outcomes

What's wrong with running our standard linear regression?

$$\text{species present}_i = \beta_0 + \beta_1 \text{forest cover}_i + \varepsilon_i$$



Modeling probabilities

- Our data take on the form $y_i = 1$ or $y_i = 0$

Modeling probabilities

- Our data take on the form $y_i = 1$ or $y_i = 0$
- For each individual i, there is some probability p_i they have $y_i = 1$, so probability $1 - p_i$ they have $y_i = 0$

Modeling probabilities

- Our data take on the form $y_i = 1$ or $y_i = 0$
- For each individual i, there is some probability p_i they have $y_i = 1$, so probability $1 - p_i$ they have $y_i = 0$
- We are interested in how a change in variable x changes the probability of $y_i = 1$
    - That is, **we model** p_i as a function of independent variables

Modeling probabilities

- Our data take on the form $y_i = 1$ or $y_i = 0$
- For each individual i, there is some probability p_i they have $y_i = 1$, so probability $1 - p_i$ they have $y_i = 0$
- We are interested in how a change in variable x changes the probability of $y_i = 1$
  - That is, **we model** p_i as a function of independent variables
- Basic idea: we need some transformation of the *probability* that lets us write:

$$transformation(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots$$

Modeling probabilities

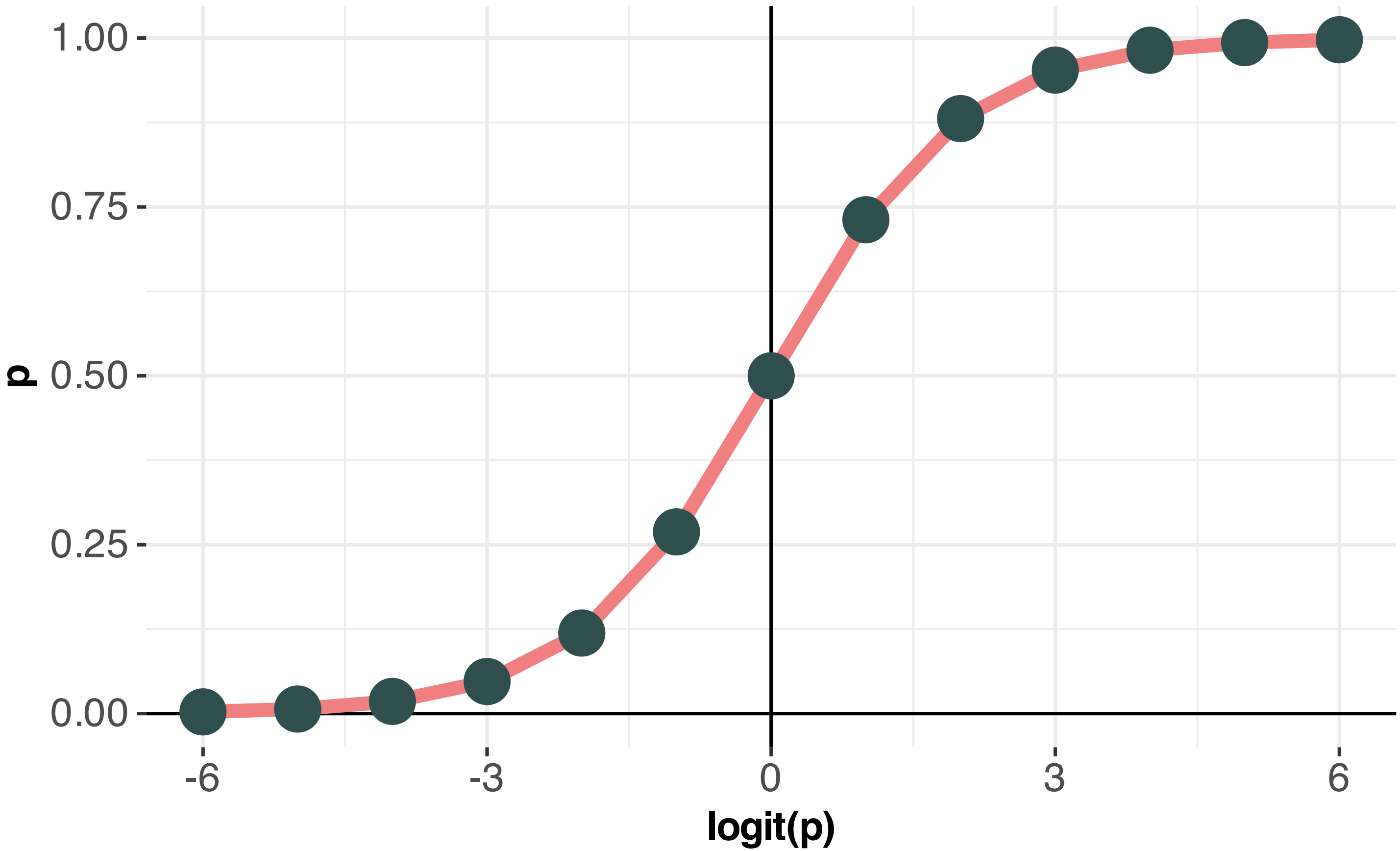
Basic idea: we need some transformation of the *probability* that lets us write:

$$transformation(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots$$

- We want this transformation to ensure that:
  - we can input a value between 0 and 1 and return a continuous variable (i.e., we want our outcome variable to be a continuous variable)
  - our predicted probabilities $\hat{p}_i$ (inverse of the transformation) will fall between 0 and 1

Logistic regression

The **logit function** is the most commonly used nonlinear transformation that ensures predicted probabilities between 0 and 1:



Logistic regression

The **logit function** is the most commonly used nonlinear transformation that ensures predicted probabilities between 0 and 1:

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

Logistic regression

The **logit function** is the most commonly used nonlinear transformation that ensures predicted probabilities between 0 and 1:

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

We can then write:

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots$$

The logit function is also called "log odds" because the "odds ratio" is the probability of success, p_i, divided by the probability of failure, $1 - p_i$

Logistic regression

The **logit function** is the most commonly used nonlinear transformation that ensures predicted probabilities between 0 and 1:

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

We can then write:

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots$$

The logit function is also called "log odds" because the "odds ratio" is the probability of success, p_i, divided by the probability of failure, $1 - p_i$

Because of the properties of the logit function (see last graph), this ensures we will generate predicted probabilities $\hat{p}_i$ that fall between 0 and 1.

Logistic regression

How do we estimate this regression?

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots$$

Logistic regression

How do we estimate this regression?

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots$$

- Can't use linear regression -- we don't have data on p_i! We only see $y_i = 1$ or $y_i = 0$
- We use what's called "maximum likelihood estimation" (alternatively, can use gradient descent)
  - Essentially, this asks: what combination of parameters $\beta_0, \beta_1, \ldots$ maximizes the likelihood that we would observe the data we have?
  - E.g., if you have high x_1 values coinciding with many $y_i = 1$ values, likely that β_1 is high and that p_i is high for observations with large x_1

Logistic regression

How do we estimate this regression?

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots$$

All you really need to know on estimation is...

- That we use `glm()` instead of `lm()` -- GLM for "generalized linear model"
- Interpreting coefficients is a lot more complicated! (next slide)

Interpreting logistic regression output

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots$$

- β_k: effect of a 1-unit change in x_k on the log-odds of $y = 1$ 🤔

Interpreting logistic regression output

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots$$

- β_k: effect of a 1-unit change in x_k on the log-odds of $y = 1$ 🤔

We need to transform our output to get predicted probabilities back!

$$\begin{aligned}
\log\left(\frac{p_i}{1-p_i}\right) &= b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i} \\
\frac{p_i}{1-p_i} &= e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}} \\
p_i &= (1-p_i)\, e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}} \\
p_i &= e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}} - p_i \times e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}} \\
p_i + p_i\, e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}} &= e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}} \\
p_i(1 + e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}}) &= e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}} \\
p_i &= \frac{e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}}}{1 + e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}}}
\end{aligned}$$

Interpreting logistic regression output

This means that if you run a regression with many independent variables, you need to plug your estimated $\hat{\beta}$'s *and* the values of all your x variables into this equation to get back a predicted probability for any individual:

$$p_i = \frac{e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}}}{1 + e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}}}$$

Interpreting logistic regression output

This means that if you run a regression with many independent variables, you need to plug your estimated $\hat{\beta}$'s *and* the values of all your x variables into this equation to get back a predicted probability for any individual:

$$p_i = \frac{e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}}}{1 + e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}}}$$

If you want to know the *effect* of changing just one variable x_j on the probability p_i, you need to compute:

$$p_i(x_j + 1) - p_i(x_j) = \frac{e^{b_0 + \cdots + b_j (x_{j,i} + 1) + \cdots + b_k x_{k,i}}}{1 + e^{b_0 + \cdots + b_j (x_{j,i} + 1) + \cdots + b_k x_{k,i}}} - \frac{e^{b_0 + \cdots + b_j x_{j,i} + \cdots + b_k x_{k,i}}}{1 + e^{b_0 + \cdots + b_j x_{j,i} + \cdots + b_k x_{k,i}}}$$

Interpreting logistic regression output

This means that if you run a regression with many independent variables, you need to plug your estimated $\hat{\beta}$'s *and* the values of all your x variables into this equation to get back a predicted probability for any individual:

$$p_i = \frac{e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}}}{1 + e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}}}$$

If you want to know the *effect* of changing just one variable x_j on the probability p_i, you need to compute:

$$p_i(x_j + 1) - p_i(x_j) = \frac{e^{b_0 + \cdots + b_j (x_{j,i} + 1) + \cdots + b_k x_{k,i}}}{1 + e^{b_0 + \cdots + b_j (x_{j,i} + 1) + \cdots + b_k x_{k,i}}} - \frac{e^{b_0 + \cdots + b_j x_{j,i} + \cdots + b_k x_{k,i}}}{1 + e^{b_0 + \cdots + b_j x_{j,i} + \cdots + b_k x_{k,i}}}$$

Note that this calculation depends on all the other x's! And it will vary with the baseline level of x_j

Logistic regression: Example

- Bertrand and Mullainathan (2003) study discrimination in hiring decisions
- Authors created many fake resumes, randomly assigning different characteristics (name, sex, race, experience, honors, etc.)

Logistic regression: Example

- Bertrand and Mullainathan (2003) study discrimination in hiring decisions
- Authors created many fake resumes, randomly assigning different characteristics (name, sex, race, experience, honors, etc.)
- **Outcome variable is binary:** Did the resume get a call back from a (real) potential employer?
    - Yes: $y_i = 1$
    - No: $y_i = 0$
- Manipulated first names to be those that are commonly associated with White or Black individuals
- Random study design allows estimation of the causal effect of race on callback probability

Logistic regression: Example

List of all 36 unique names along with the commonly inferred race and sex associated with these names.

first_name	**race**	**sex**	**first_name**	**race**	**sex**	**first_name**	**race**	**sex**
Aisha	Black	female	Hakim	Black	male	Laurie	White	female
Allison	White	female	Jamal	Black	male	Leroy	Black	male
Anne	White	female	Jay	White	male	Matthew	White	male
Brad	White	male	Jermaine	Black	male	Meredith	White	female
Brendan	White	male	Jill	White	female	Neil	White	male
Brett	White	male	Kareem	Black	male	Rasheed	Black	male
Carrie	White	female	Keisha	Black	female	Sarah	White	female
Darnell	Black	male	Kenya	Black	female	Tamika	Black	female
Ebony	Black	female	Kristen	White	female	Tanisha	Black	female
Emily	White	female	Lakisha	Black	female	Todd	White	male

Logistic regression: Example

Variables included in the data (all randomly assigned):

variable	**description**
`received_callback`	Specifies whether the employer called the applicant following submission of the application for the job.
`job_city`	City where the job was located: Boston or Chicago.
`college_degree`	An indicator for whether the resume listed a college degree.
`years_experience`	Number of years of experience listed on the resume.
`honors`	Indicator for the resume listing some sort of honors, e.g. employee of the month.

Logistic regression: Example

Variables included in the data (all randomly assigned):

variable	description
`military`	Indicator for if the resume listed any military experience.
`has_email_address`	Indicator for if the resume listed an email address for the applicant.
`race`	Race of the applicant, implied by their first name listed on the resume.
`sex`	Sex of the applicant (limited to only and in this study), implied by the first name listed on the resume.

Logistic Regression: example

- First, we estimate a single predictor: `race`
- `race` indicates whether the applicant is White or not (**Note:** `race` is also binary in this case!)
- We find:

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -2.67 + 0.44 \times \texttt{race\_White}$$

a. If a resume is randomly selected from the study and it has a Black associated name, what is the probability it resulted in a callback?

b. What would the probability be if the resume name was associated with White individuals?

Logistic regression: Example

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -2.67 + 0.44 \times \texttt{race\_white}$$

a. If a resume is randomly selected from the study and it has a Black associated name, what is the probability it resulted in a callback?

Logistic regression: Example

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -2.67 + 0.44 \times \texttt{race\_white}$$

a. If a resume is randomly selected from the study and it has a Black associated name, what is the probability it resulted in a callback?

Answer: If a randomly chosen resume is associated with a Black name, then `race_white` takes the value of 0 and the right side of the model equation equals -2.67. Solving for p_i gives $log(\frac{\hat{p}_i}{1-\hat{p}_i}) = -2.67 \implies \hat{p}_i = \frac{e^{-2.67}}{1+e^{-2.67}} = 0.065$.

Logistic regression: Example

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -2.67 + 0.44 \times \texttt{race\_white}$$

b. What would the probability be if the resume name was associated with White individuals?

Answer: If the resume had a name associated with White individuals, then the right side of the model equation is $-2.67 + 0.44 \times 1 = -2.23$. This translates into $\hat{p}_i = 0.097$.

Logistic regression: Example

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -2.67 + 0.44 \times \texttt{race\_white}$$

b. What would the probability be if the resume name was associated with White individuals?

Answer: If the resume had a name associated with White individuals, then the right side of the model equation is $\mathbf{-2.67 + 0.44 \times 1 = -2.23}$. This translates into $\boldsymbol{\hat{p}_i = 0.097}$.

Conclude: Being White increases the likelihood of a call back, by 3.2 percentage points.

Logistic regression: Example

Use the same process to compute predicted probabilities with multiple independent variables, you just have more calculations!

Logistic regression: Example

Use the same process to compute predicted probabilities with multiple independent variables, you just have more calculations!

For example, you might estimate:

$$\begin{aligned} log\left(\frac{p}{1-p}\right) \\ = -2.7162 - 0.4364 \times \texttt{job\_city}_{\texttt{Chicago}} \\ + 0.0206 \times \texttt{years\_experience} \\ + 0.7634 \times \texttt{honors} - 0.3443 \times \texttt{military} + 0.2221 \times \texttt{email} \\ + 0.4429 \times \texttt{race}_{\texttt{White}} - 0.1959 \times \texttt{sex}_{\texttt{man}} \end{aligned}$$

To predict callback probability for a White individual, you also need to know job location, experience, honors, military experience, whether they have an email, race, and sex!

Logistic regression: Example

For example, you might estimate:

$$
\begin{aligned}
&log\left(\frac{p}{1-p}\right) \\
&= -2.7162 - 0.4364 \times \texttt{job\_city}_{\texttt{Chicago}} \\
&\quad + 0.0206 \times \texttt{years\_experience} \\
&\quad + 0.7634 \times \texttt{honors} - 0.3443 \times \texttt{military} + 0.2221 \times \texttt{email} \\
&\quad + 0.4429 \times \texttt{race}_{\texttt{White}} - 0.1959 \times \texttt{sex}_{\texttt{man}}
\end{aligned}
$$

Note: the effect of race on call back now varies based on all the other covariates!

- Try it: Effect of being white for Chicago male with 10 years experience, an email, no honors and no military experience *versus* a female with the same characteristics?

Multinomial logistic regression

What if your outcome variable is categorical, not binary?

Multinomial logistic regression

What if your outcome variable is categorical, not binary?

For example:

- Species
- Socioeconomic status
- Survey responses
- …

Multinomial logistic regression

What if your outcome variable is categorical, not binary?

For example:

- Species
- Socioeconomic status
- Survey responses
- ...

Multinomial logistic regression generalizes the binary logistic regression you've seen here to work for multiple outcome categories

- Model predicts the probability an individual will fall into each category
- Beyond the scope of this class, but not a far leap from what you've seen here (lots of online resources -- ask me if you're interested!)

Slides created via the R package **xaringan**.

Some slide components were borrowed from Ed Rubin's awesome course materials.