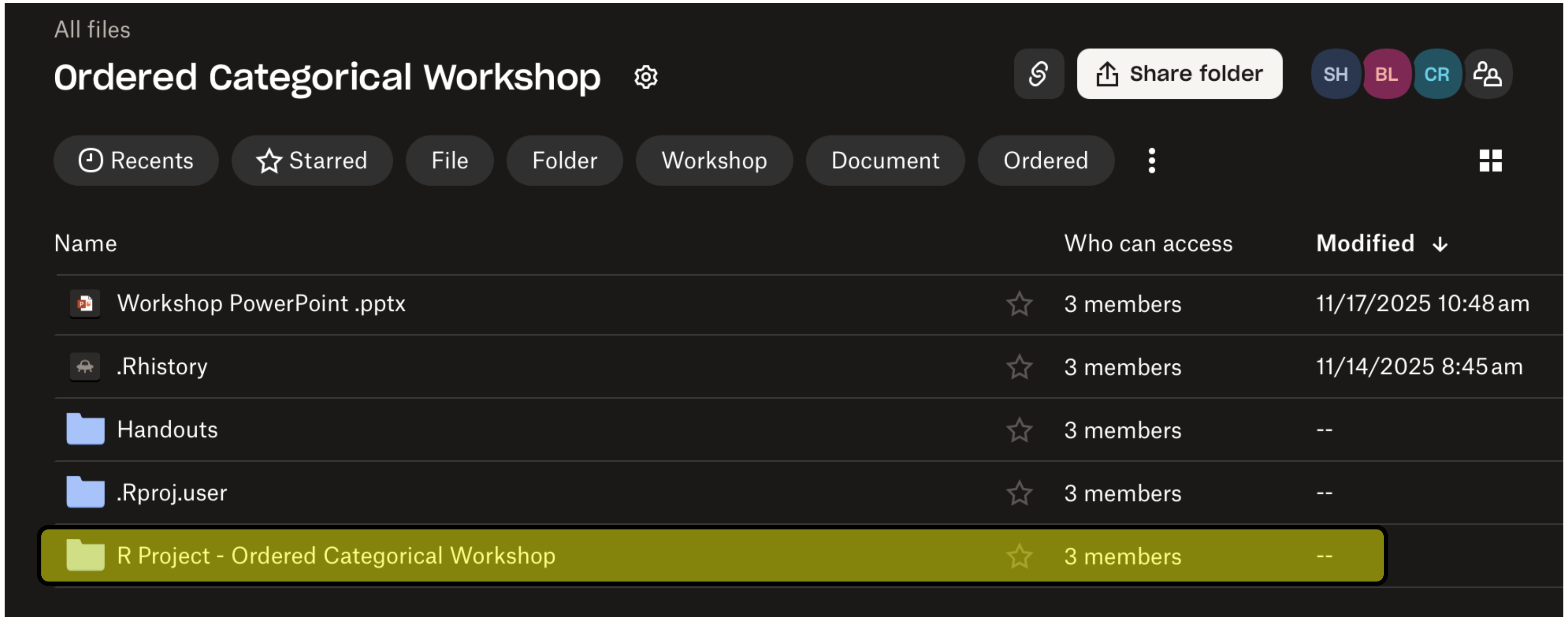
ORDERED CATEGORICAL VARIABLES AND HOW TO HANDLE LOGISTIC REGRESSION

CAROLINE ROBBINS
AND
SARAH HOLLAND

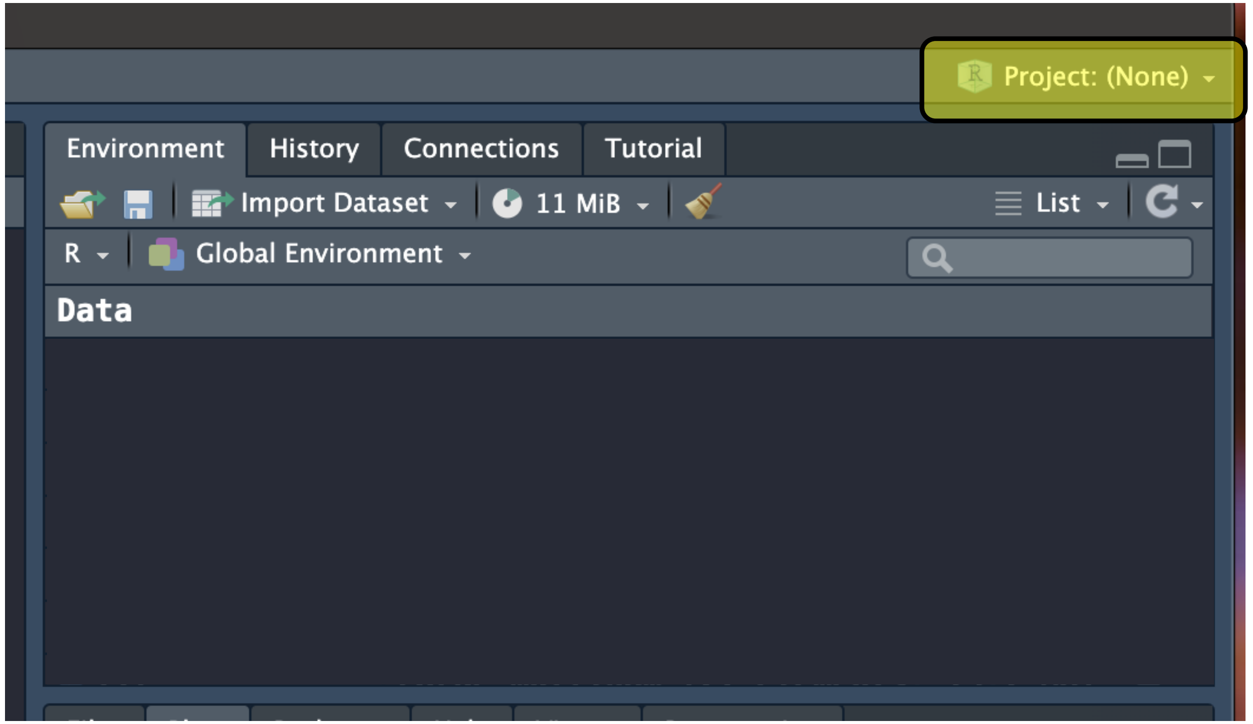
NOVEMBER 19TH, 2025 | POS 5747

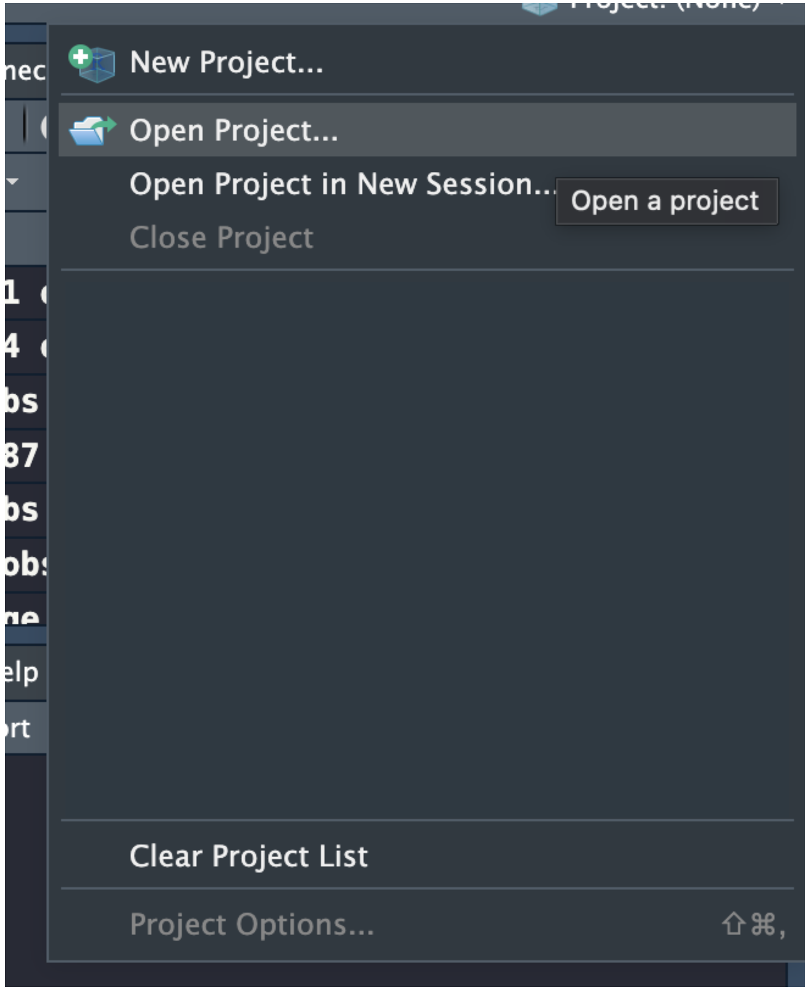
SET-UP

Please download this folder!

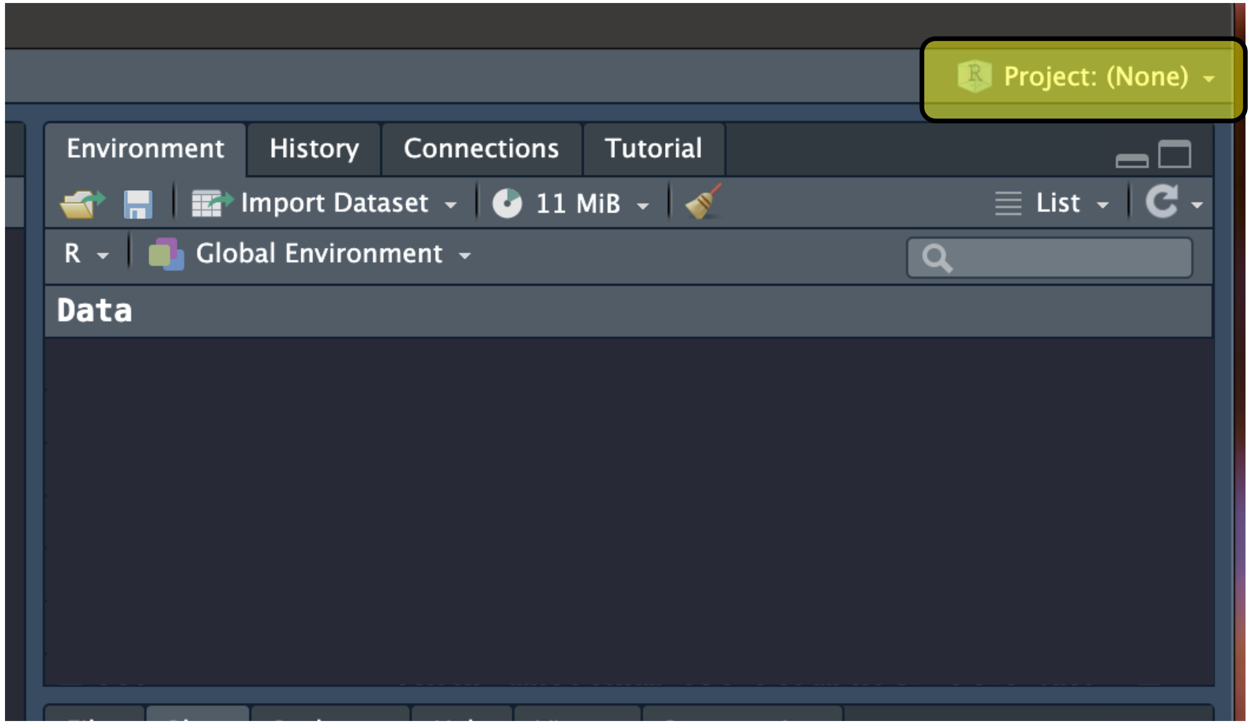


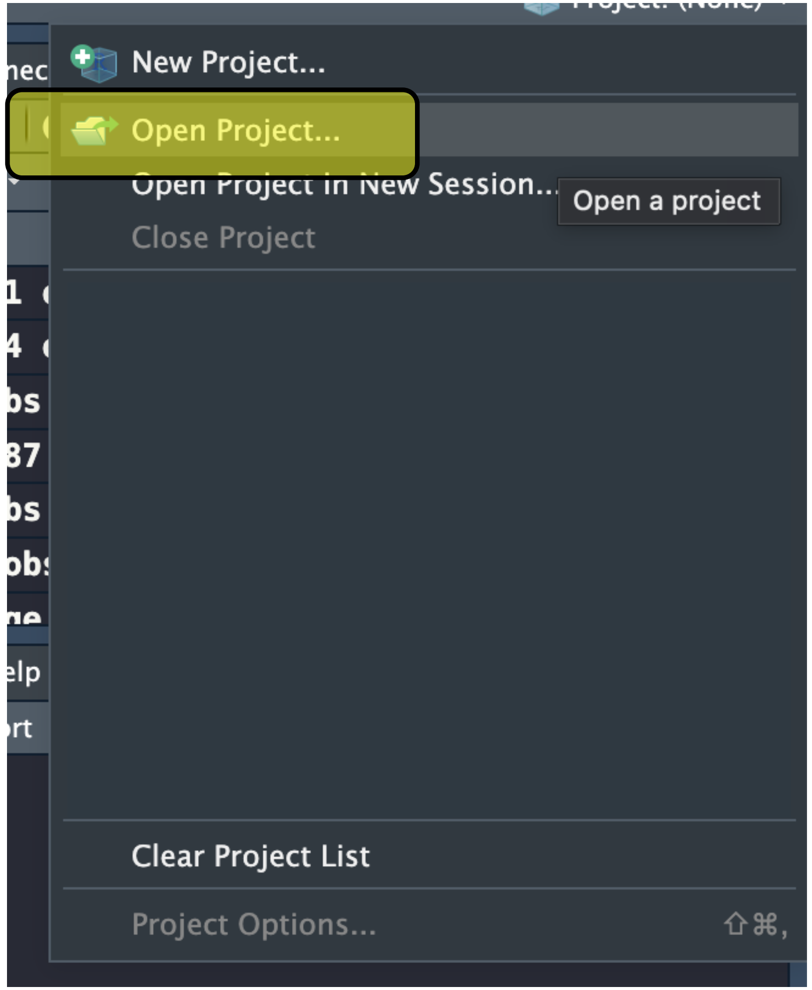
SET-UP: IN R STUDIO



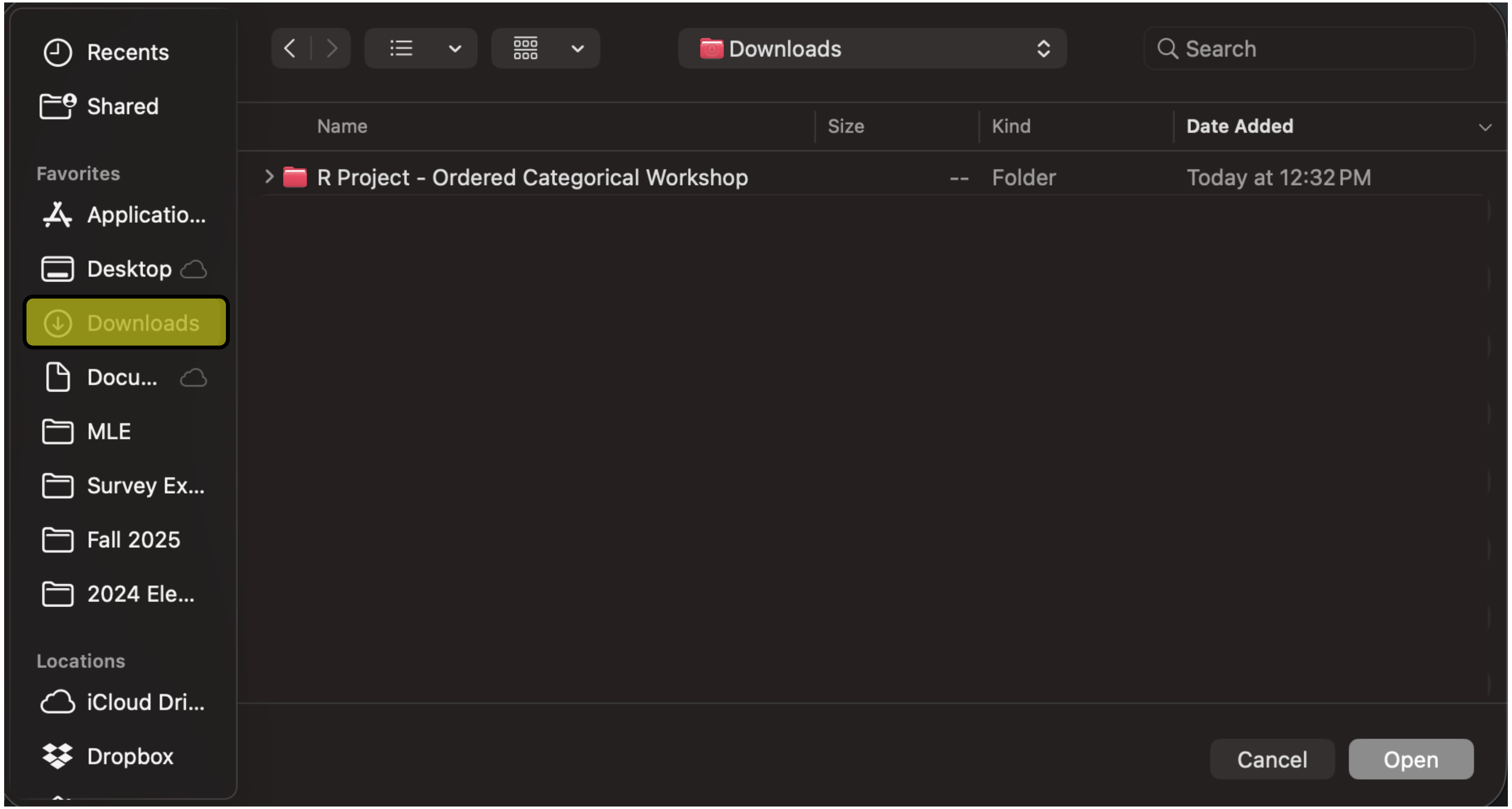


SET-UP: IN R STUDIO

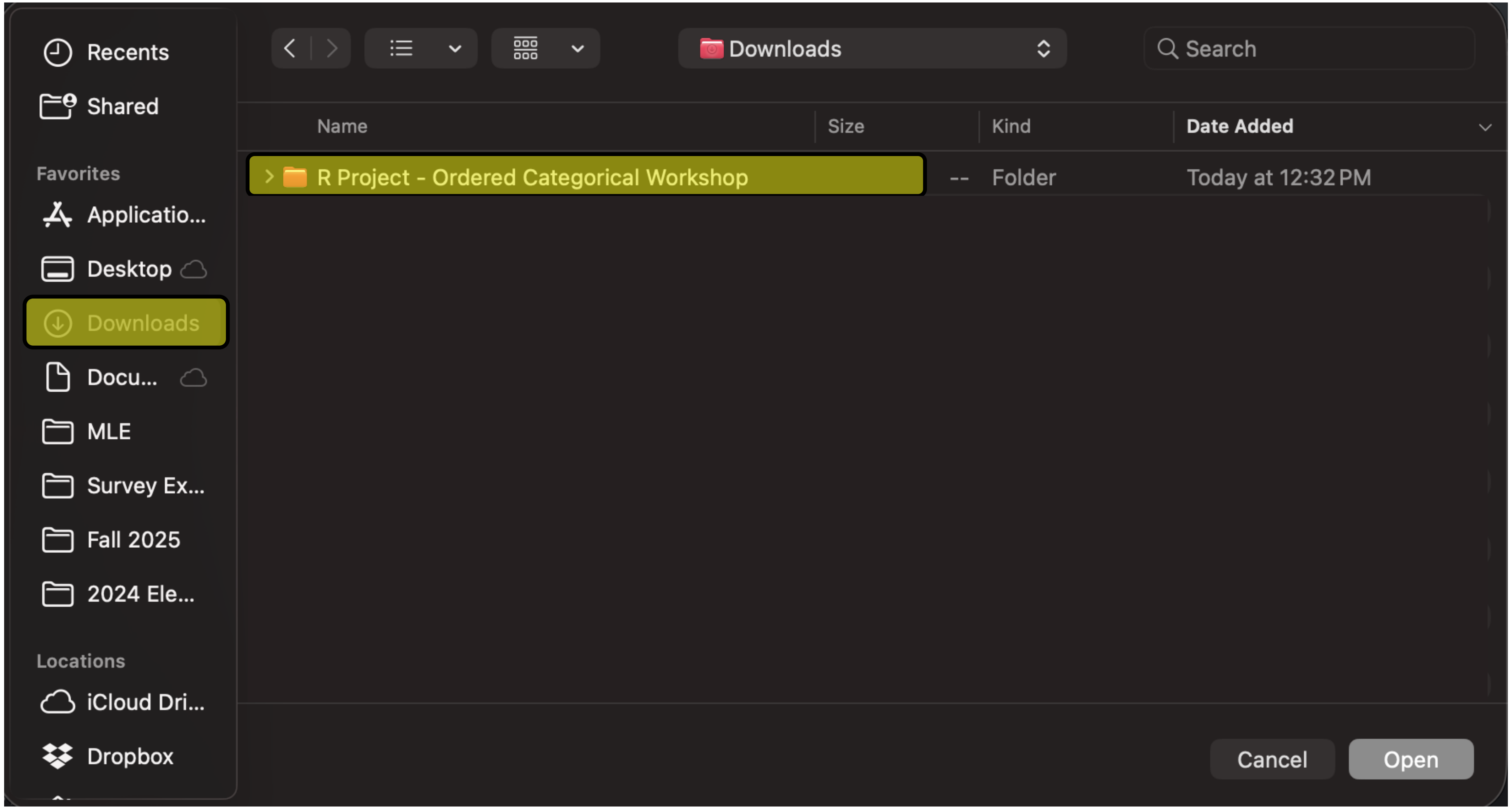




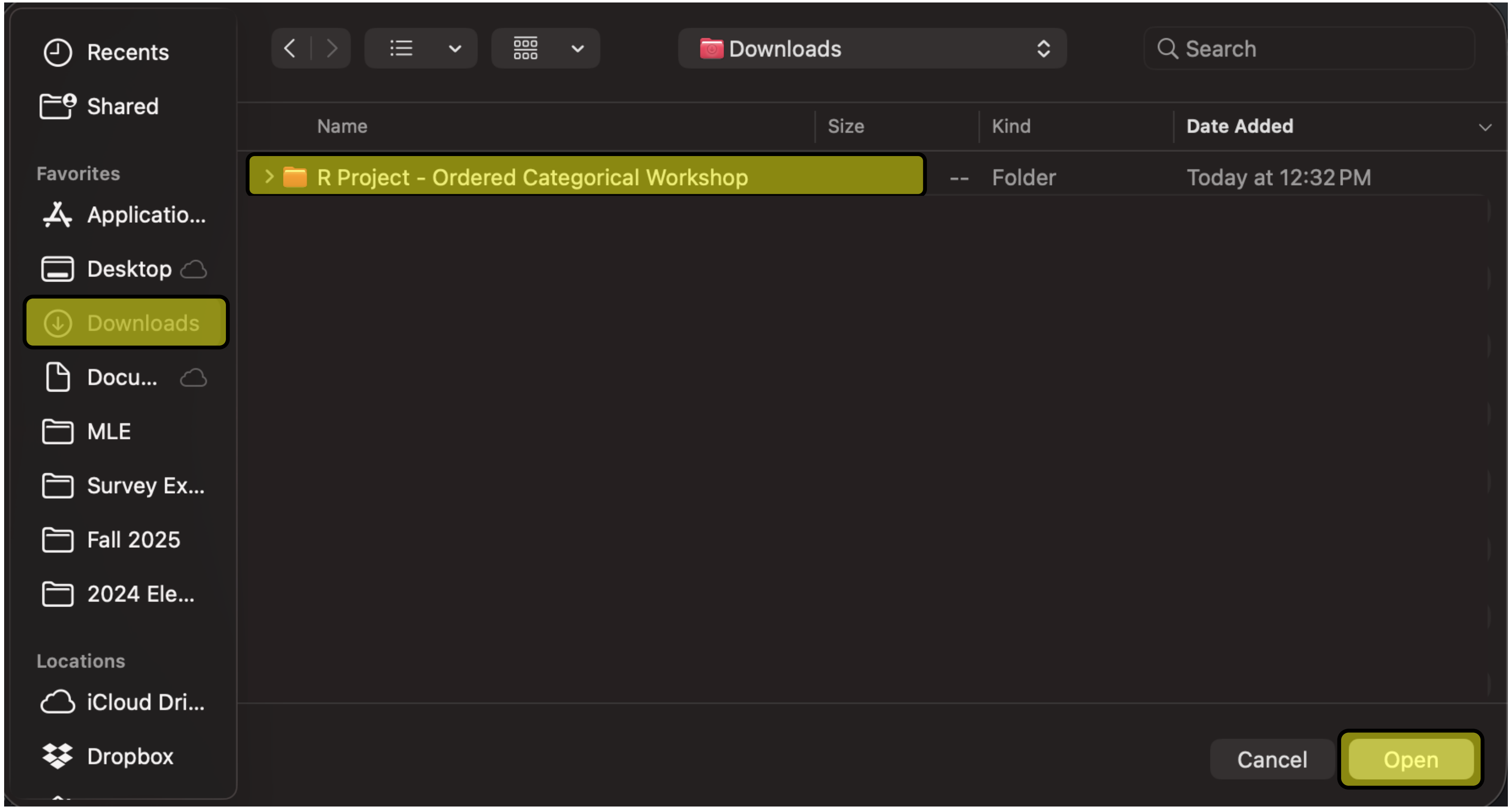
SET-UP: IN R STUDIO



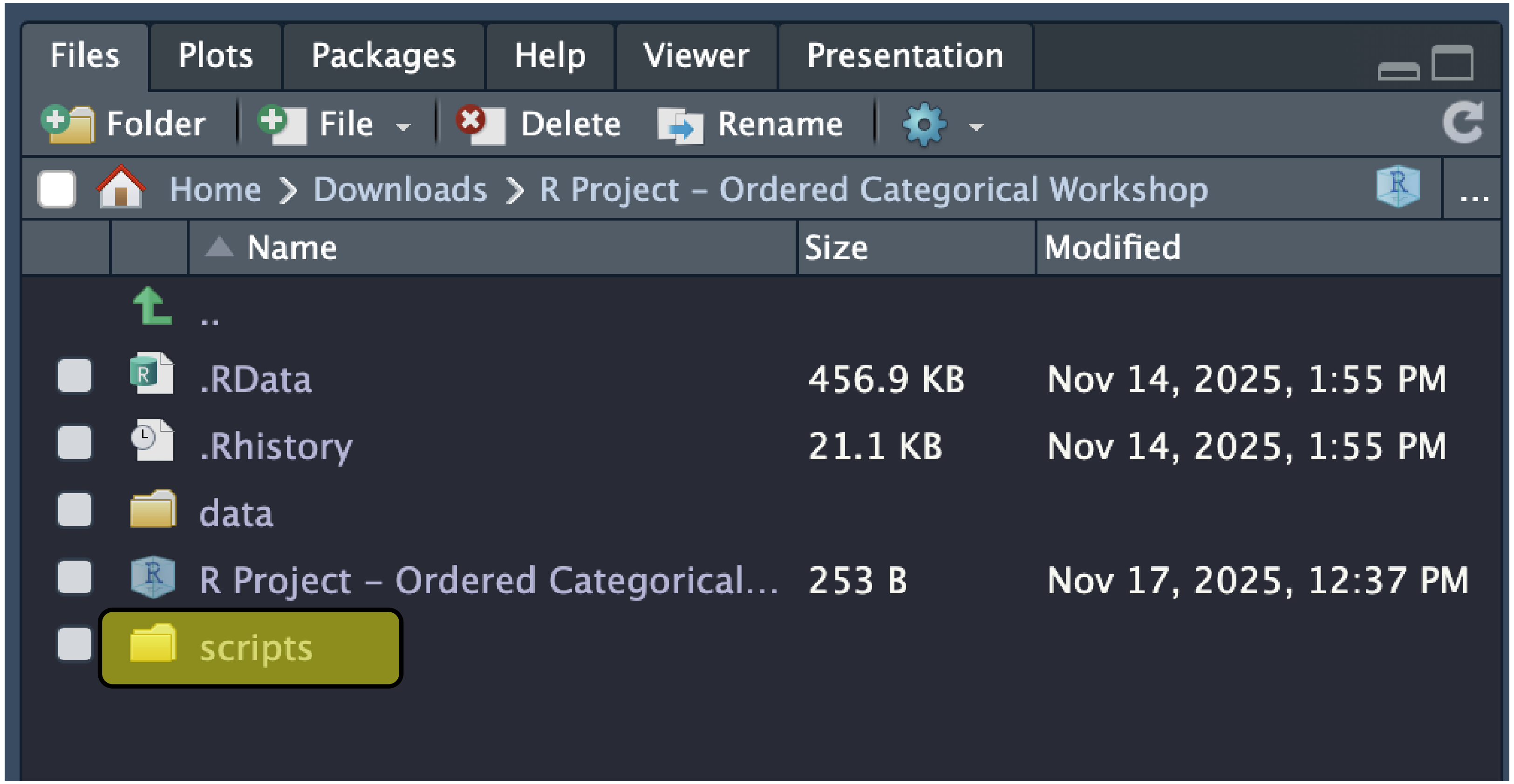
SET-UP: IN R STUDIO



SET-UP: IN R STUDIO



SET-UP: IN R STUDIO



OUTLINE

1. What are ordered categorical variables?

OUTLINE

1. What are ordered categorical variables?
2. What is logistic regression?

OUTLINE

1. What are ordered categorical variables?
2. What is logistic regression?
3. Example One: County Level Voter Confidence

OUTLINE

1. What are ordered categorical variables?
2. What is logistic regression?
3. Example One: County Level Voter Confidence
4. Example Two: Congressional Approval

SO, WHAT IN THE WORLD IS AN ORDERED CATEGORICAL VARIABLE?

WHAT IS AN ORDERED CATEGORICAL OUTCOME? (OCO)

Start with the last term first!

- **Outcome** - looking at **dependent variable(s)**

WHAT IS AN ORDERED CATEGORICAL OUTCOME? (OCO)

Start with the last term first!

- **Outcome** - looking at **dependent variable(s)**
- **Categorical** - these values represent different **groups**

WHAT IS AN ORDERED CATEGORICAL OUTCOME? (OCO)

Start with the last term first!

- **Outcome** - looking at **dependent variable(s)**
- **Categorical** - these values represent different **groups**
- **Ordered** - there is a **ranking** of these positions

WHAT IS AN ORDERED CATEGORICAL OUTCOME? (OCO)

Start with the last term first!

- **Outcome** - looking at **dependent variable(s)**
- **Categorical** - these values represent different **groups**
- **Ordered** - there is a **ranking** of these positions

So, what does it **mean** to have an ordered categorical outcome?

HOW TO THINK ABOUT OUTCOMES

Think in terms of **probabilities:**

HOW TO THINK ABOUT OUTCOMES

Think in terms of **probabilities**:

"How **likely** is it that I will end up in a **certain category**?"

WHAT IS AN ORDERED CATEGORICAL OUTCOME? (OCO)

Mathematically, we would represent these chances with a PMF (Probability Mass Function)

$$P(Y = k \mid \mathbf{x}) = \begin{cases} F(\tau_1 - \mathbf{x}^\top \boldsymbol{\beta}), & \text{if } k = 1 \\ F(\tau_k - \mathbf{x}^\top \boldsymbol{\beta}) - F(\tau_{k-1} - \mathbf{x}^\top \boldsymbol{\beta}), & \text{if } 1 < k < K \\ 1 - F(\tau_{K-1} - \mathbf{x}^\top \boldsymbol{\beta}), & \text{if } k = K \end{cases}$$

HOW TO THINK ABOUT OUTCOMES

Think in terms of **probabilities**:

"How **likely** is it that I will end up in a **certain category**?"

What sorts of categories might political scientists be interested in?

THINK LIKE A LIKERT

The most familiar type of OCO you may have encountered is a **Likert scale**!

THINK LIKE A LIKERT

The most familiar type of OCO you may have encountered is a **Likert scale**!

To what extent to you agree/disagree with the following statement?

1. Strongly Agree
2. Agree
3. Neither Agree nor Disagree
4. Disagree
5. Strongly Disagree

THINK LIKE A LIKERT

Let's break it down:

1. Is a Likert scale an **outcome**?

THINK LIKE A LIKERT

Let's break it down:

1. Is a Likert scale an **outcome**?

    - **Yes!** We want to know how the public feels about a policy - we measure it through how much they agree/disagree

THINK LIKE A LIKERT

Let's break it down:

1. Is a Likert scale an **outcome**? **Yes!**
2. Is a Likert scale categorical?

THINK LIKE A LIKERT

Let's break it down:

1. Is a Likert scale an **outcome**? **Yes!**
2. Is a Likert scale categorical?

    \- **Yes!** Instead of numerical values (like age increasing by each year), we have **bins of agreement**

    \- Don't let the numbers fool you! The numbers just **represent which group** the respondent **falls** in

THINK LIKE A LIKERT

Let's break it down:

1. Is a Likert scale an **outcome**? **Yes!**
2. Is a Likert scale categorical? **Yes!**
3. Is a Likert scale ordered?

THINK LIKE A LIKERT

Let's break it down:

1. Is a Likert scale an **outcome**? **Yes!**
2. Is a Likert scale categorical? **Yes!**
3. Is a Likert scale ordered?

   - **Yes!** The scale (in this case) runs from agree to disagree - with a **"ranking" of agreement**

APPLICATION

What might be some research questions in political science that might require OCO?

What about other disciplines?

Take a couple minutes and come up with some!
Write them down!

APPLICATION

Examples:

1. Does partisanship shape how confident voters feel that their ballots were counted accurately?

APPLICATION

Examples:

1. Does partisanship shape how confident voters feel that their ballots were counted accurately?
2. Does income or education increase the likelihood of "strongly agreeing" with democratic norms?

APPLICATION

Examples:

1. Does partisanship shape how confident voters feel that their ballots were counted accurately?
2. Does income or education increase the likelihood of "strongly agreeing" with democratic norms?
3. Do experiential treatments (ex. audit information, elite cues) shift respondents up or down an ordered attitude scale?

THINK LIKE A LIKERT

So, when we use a Likert scale, **what are we measuring?**

- In terms of regression, we want to know how likely it is that respondents fall into a certain category (i.e. agreement or disagreement)

THINK LIKE A LIKERT

So, when we use a Likert scale, **what are we measuring**?

- In terms of regression, we want to know how likely it is that respondents fall into a certain category (i.e. agreement or disagreement)

But what about OLS?

THINK LIKE A LIKERT

So, when we use a Likert scale, **what are we measuring**?

- In terms of regression, we want to know how likely it is that respondents fall into a certain category (i.e. agreement or disagreement)

But what about OLS?

- **An important note:** OCOs (like Likert scales) can only be used in OLS regressions when there is a strong assumption that the distance between categories is **equal**
  - Jump from Strongly disagree to Disagree = Disagree to Neither = Neither to Agree = Agree to Strongly agree

WHAT HAPPENS WHEN OUR CATEGORIES ARE NOT EQUALLY DISTANCED?

UNEQUAL OR UNKNOWN CATEOGRY DISTANCE

Two things to remember:

UNEQUAL OR UNKNOWN CATEOGRY DISTANCE

Two things to remember:

1. Ordered categorical variables are **not** continuous measures
   - Remember: we are separating respondents into separate bins

UNEQUAL OR UNKNOWN CATEOGRY DISTANCE

Two things to remember:

1. Ordered categorical variables are **not** continuous measures

    - Remember: we are separating respondents into separate bins

2. OLS/assumption of continuity can only occur if there **is equal distance** between measures

    - No equal distance complicates estimation

    - **Think**: If our categories were - Strongly agree, Neither agree nor disagree, Somewhat disagree, and Disagree, what problems might arise? How would we interpret coefficients in OLS?

WHAT IS A LOGISTIC REGRESSION

LEANING ON LOGISTIC

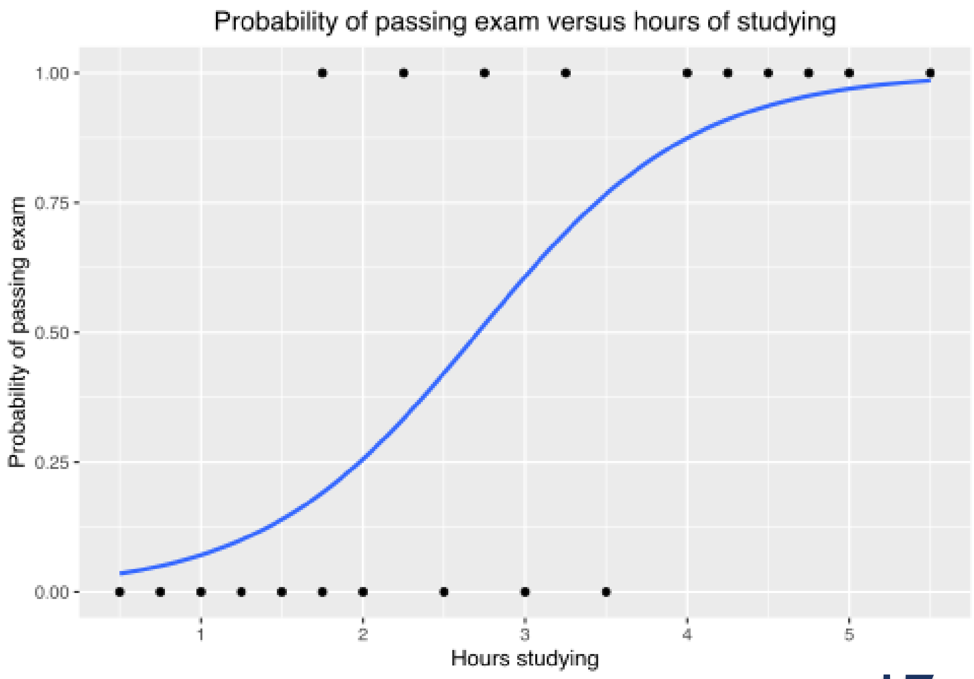
Logistic regression models are a great way to handle OCOs where our OLS assumptions are unable to be met

LEANING ON LOGISTIC

Logistic regression models are a great way to handle OCOs where our OLS assumptions are unable to be met

- Estimating the **probability** of an outcome
- logit(p) = ln(p/(1-p))
  - **p** = probability of the event occurring

We want to know the **odds**



LEANING ON LOGISTIC

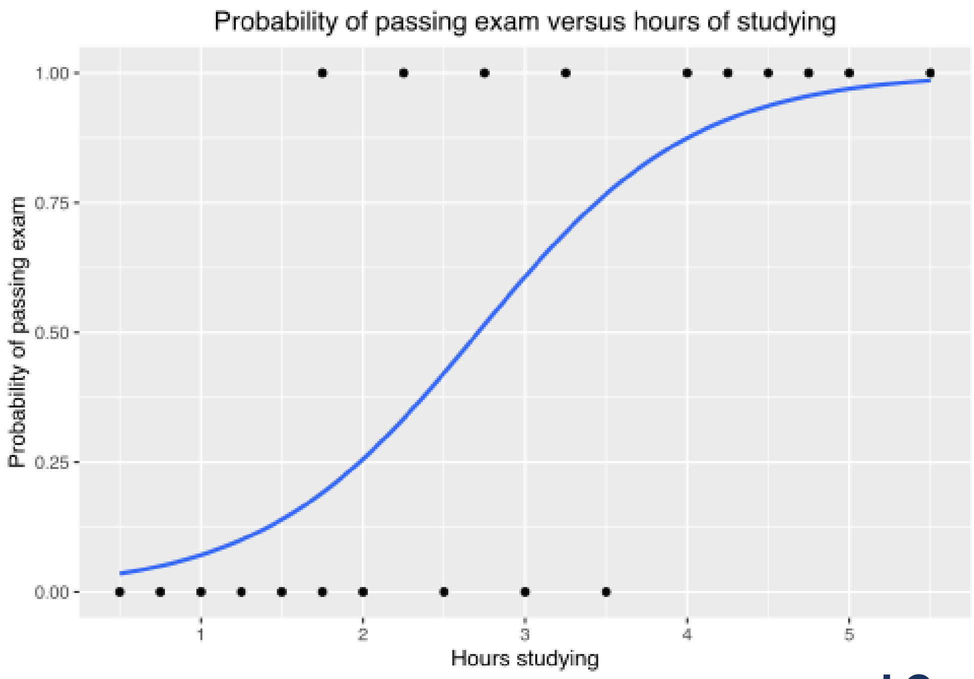
Why the odd S-curve?

- Keeps probabilities between 0 and 1
  - Linear models don't have this restriction!

Binary outcomes more common

- Predicting a Yes/No

What if we want to add more than one category?



LEANING ON LOGISTIC

To compare across multiple categories (think back to our 5-point scale), we must select a **reference category**

- Reference categories are a baseline
  - Think like a control group, but for outcomes
- Logit models are calculating a change in odds of receiving a particular outcome
  - Our variables are used to calculate how the odds of being in a particular category change relative to the odds of being in the reference category

LEANING ON LOGISTIC

But how do we interpret these outputs? (No longer a one-unit increase in X means coefficient-level increase in Y)

- We are finding the **probability** that our outcome is within a **set range of values**, **given** the levels of our independent variables

LEANING ON LOGISTIC

We are **not** particularly interested in **coefficient** values here

- Depending on the model, they can be extremely difficult to interpret

Instead, focus on the sign (+/-)

- **Positive** values mean the given combination of independent variables **increases the odds** of the respondent being in a **"higher"** category
- **Negative** values mean the combination means being in a **"higher" category is less likely**

BEYOND LOGITS

There are other methods of interpreting categorical outcomes:

- Probit models are very similar to logit models (normally distributed errors)
- Linear Probability Models (LPMs) are a cross between OLS and logits

For ease of understanding, **logits are the easiest to handle**

- Easy interpretation (sign as an odds-indicator)
- Errors are logistically distributed

POLR() IN R

`polr(formula, data = ...)`
runs an ordered logistic regression.

```
polr(formula,
     data = your data)
```

POLR() IN R

`polr(formula, data = ...)`
runs an ordered logistic regression.

- **`polr`** = Proportional Odds Logistic Regression

```
polr(formula,
        data = your data)
```

POLR() IN R

`polr(formula, data = ...)`
runs an ordered logistic regression.

- **`polr`** = Proportional Odds Logistic Regression
- **`Formula`** = tells R what you want to predict and which variables predict it
  - Example: outcome ~ x1 + x2 + x3

```
polr(formula,
     data = your data)
```

POLR ( ) IN R

`polr(formula, data = ...)`

runs an ordered logistic regression.

- **`polr`** = Proportional Odds Logistic Regression
- **`Formula`** = tells R what you want to predict and which variables predict it
  - Example: outcome ~ x1 + x2 + x3
- **`data`** = tells R where the variables come from

```
polr(formula,
     data = your data)
```

ANY QUESTIONS BEFORE SOME PRACTICE?

EXAMPLE ONE: COUNTY LEVEL

First, we are going to work with some data I collected from a survey I ran.

This survey was gauging how varying levels of information given to respondents about post-election audits altered individual levels of voter confidence.

Very confident, Somewhat confident, Not too confident, Not confident at all

EXAMPLE ONE: COUNTY LEVEL

1. Library your packages
2. Load your data

- Always be sure to look at your data!
  - Identify what your variables look like
  - Use `glimpse()`

```
load("data/survey_data.RData")
glimpse(survey_data)
```

```
Rows: 293
Columns: 13
$ age                <dbl> 76, 33, 41, 35, 30, 64, 68, 28, 80, 78, 63, 58, 55,…
$ gender             <fct> Female, Male, Male, Male, Female, Male, Female, Mal…
$ pre_treatment_conf <ord> Very confident, Somewhat confident, Somewhat confid…
$ county_conf        <ord> Very confident, Somewhat confident, Somewhat confid…
$ state_conf         <ord> Very confident, Somewhat confident, Very confident,…
$ nation_conf        <ord> Very confident, Somewhat confident, Somewhat confid…
$ over_conf          <ord> Very confident, Somewhat confident, Somewhat confid…
$ tr_group           <chr> "part_info", "control", "full_info_per", "full_info…
$ pid3               <fct> Democrat, Democrat, Independent, Democrat, Democrat…
$ ideology           <dbl> 2, 5, 5, 4, 5, 3, 6, 3, 4, 3, 6, 7, 4, 2, 5, 4, 4, …
$ pres_vote          <fct> Harris, Trump, Trump, Harris, Harris, Harris, Trump…
$ educ               <ord> "Professional degree (MD, DDS, JD)", "Bachelor's de…
$ race_ethnicity     <fct> "Asian", "Asian", "White", "Black", "White", "Asian…
```

EXAMPLE ONE: COUNTY LEVEL

Please run this function!

It prevents issues with different versions of R

```
get_cat_col <- function(obj) {
  possible <- c("group", "response",
"outcome", "predicted", "category")
  found <- intersect(possible,
names(obj))
  if (length(found) == 0) stop("No
category column found in object.")
  found[1]
}
```

EXAMPLE ONE: COUNTY LEVEL

Set up the ordered logistic regression:

- In **`MASS`**, use **`polr()`**
  - First value is your function (Y ~ X)
  - **`data = your data`**

- Call a summary to see the coefficient values, standard errors, and comparison statistics

```
model_polr_county <- polr(
   county_conf ~ pid3 +
ideology + pres_vote,
   data    = survey_data
)

summary(model_polr_county)
```

EXAMPLE ONE: COUNTY LEVEL

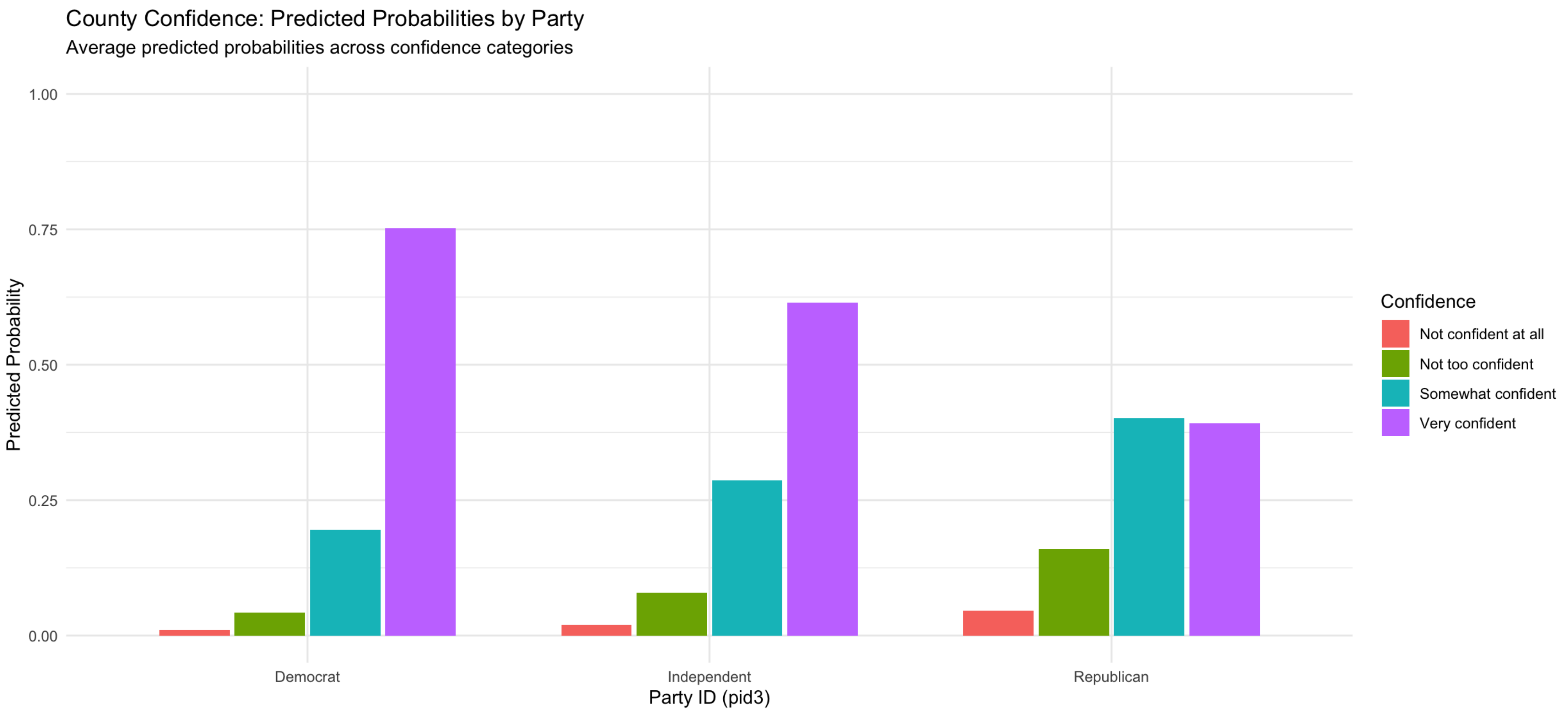
Look at marginal effects

- Using **`predictions()`** from **`marginaleffects()`** generates expected values for each data point

```
pred_county <-
predictions(model_polr_county)

glimpse(pred_county)   # predicted
probability for each category per
observation
```

How should we interpret these estimates?

EXAMPLE ONE: COUNTY LEVEL



EXAMPLE TWO: WORK WITH US!

Next, we are going to work with the Robbins et al. (Forthcoming) dataset!

This is data from an RIBC survey experiment! We were interested in understanding how the public reacts to Congress exposing covert foreign policy actions.
Overall, do you approve or disapprove of how {Respondent's In-Party} in Congress handled the situation?

Strongly approve, Somewhat approve, Slightly approve, Neither approve nor disapprove, Slightly disapprove, Somewhat disapprove, Strongly disapprove

EXAMPLE TWO: WORK WITH US!

1. Library your packages
2. Load your data

- Take a look at the outcome variable. What do you think the numbers represent?

```
robbins_data <- robbins_data %>%
  mutate(
    cong_overall = case_when(
      cong_overall == 3 ~ 1,
      cong_overall == 2 ~ 1,
      cong_overall == 1 ~ 1,
      cong_overall == 0 ~ 0,
      cong_overall == -1 ~ -1,
      cong_overall == -2 ~ -1,
      cong_overall == -3 ~ -1
    ),
    # Turn into an ordered factor: Disapproval <
Neutral < Approval

    cong_overall = factor(cong_overall, levels =
c(-1, 0, 1), ordered = TRUE)
  )

glimpse(robbins_data)
```

EXAMPLE TWO: WORK WITH US!

1. Set up your formula
- Using `polr()` ,

what arguments do you need?

2. Make sure to generate a summary!

```
f <-
```

```
summary(model)
```

EXAMPLE TWO: WORK WITH US!

1. Set up your formula
- Using `polr()` ,
what arguments do you need?

2. Make sure to generate a summary!

```
f <- cong_overall ~ failure
  + amplify + pid7
```

```
summary(model)
```

EXAMPLE TWO: WORK WITH US!

1. Set up your formula
- Using `polr()`, what arguments do you need?

2. Make sure to generate a summary!

```
f <- cong_overall ~ failure
  + amplify + pid7

model <-



summary(model)
```

EXAMPLE TWO: WORK WITH US!

1. Set up your formula
- Using `polr()` ,
what arguments do you need?

2. Make sure to generate a summary!

```
f <- cong_overall ~ failure
  + amplify + pid7

model <- polr(

formula = f,

data = robbins_data)

summary(model)
```

EXAMPLE TWO: WORK WITH US!

1. Let's try understanding group differences together using **`marginaleffects()`**

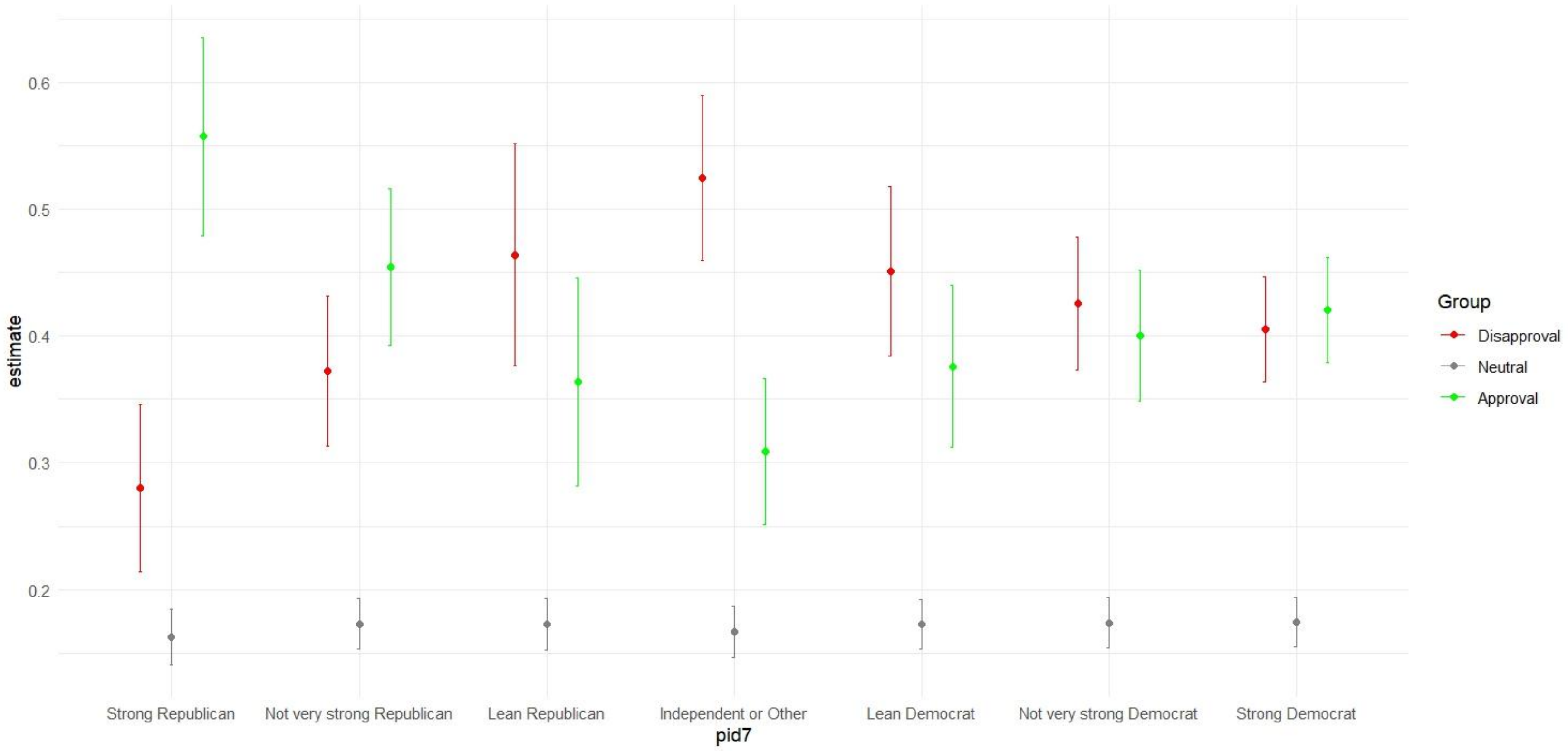
2. Intuitively, what do you think the difference is between predictions and comparisons?

```
pred_approval <- predictions(
  model, by = c("pid7", "group"))
glimpse(pred_approval)

avg_predict <- avg_predictions(
  model, by = c("pid7", "group”))
glimpse(avg_predict)

comp_approval <- comparisons(
  model, by = c("pid7", "group"))
glimpse(comp_approval)

avg_compare <- avg_comparisons(
  model, by = c("pid7", "group"))
```

EXAMPLE TWO: WORK WITH US!

1. And if we do a little fine-tuning on our graph…

   - Remember, we move from low → high on amplify



If we have time, example 3

BRINGING IT BACK TOGETHER

- **Ordered Categorical Outcomes** are familiar friends to political scientists, but it is important that we **handle them correctly**
- When we are **uncertain** about category distance, or we cannot meet the OLS assumptions, we should use an **Ordered Logit Model**
- Ordered Logits tell us the **logged odds of moving from a lower category to a higher category**

R packages like **`MASS()`**, **`marginaleffects()`**, **`and brms()`** can help us better understand what these odds mean

FEEL FREE TO ASK QUESTIONS!

THANK YOU FOR YOUR TIME!

EXAMPLE THREE: BRMS() WITH STAN

`brm()` is the function of interest

- Need both the brms **AND** **`cmdstanr`** packages
- The first three items are fairly similar (formula, data, and family)
- The remaining items are for computer performance and data generation
- Seed ensures our data is replicable!

```
model_logit_brms <- brm(
  formula = state_conf ~pid3 + ideology
+ pres_vote,
  data = survey_data,
  family = cumulative(link = "logit"),
  chains = 4,
  iter = 4000,
  warmup = 2000,
  cores = 2,
  backend = "cmdstanr",
  seed = 137)

conditional_effects(model_logit_brms)
```


Back to end

APPENDIX