Lecture 15: Missing Data

Naijia Liu

March 21 2024

Pset II

- Due next Thusrday 3/28 midnight
- Late submission on Sunday midnight.
- Coding task only.

Multiple imputation

1. A simple imputation, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as "place holders."

Multiple imputation

1. A simple imputation, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as "place holders."
2. The "place holder" mean imputations for one variable ("var") are set back to missing.

Multiple imputation

1. A simple imputation, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as "place holders."
2. The "place holder" mean imputations for one variable ("var") are set back to missing.
3. The observed values from the variable "var" in Step 2 are regressed on the other variables in the imputation model. In other words, "var" is the dependent variable in a regression model and all the other variables are independent variables in the regression model.

Multiple imputation

1. A simple imputation, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as "place holders."
2. The "place holder" mean imputations for one variable ("var") are set back to missing.
3. The observed values from the variable "var" in Step 2 are regressed on the other variables in the imputation model. In other words, "var" is the dependent variable in a regression model and all the other variables are independent variables in the regression model.
4. The missing values for "var" are then replaced with predictions (imputations) from the regression model.

Multiple imputation

1. A simple imputation, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as "place holders."
2. The "place holder" mean imputations for one variable ("var") are set back to missing.
3. The observed values from the variable "var" in Step 2 are regressed on the other variables in the imputation model. In other words, "var" is the dependent variable in a regression model and all the other variables are independent variables in the regression model.
4. The missing values for "var" are then replaced with predictions (imputations) from the regression model.
5. Steps 2–4 are then repeated for each variable that has missing data.

Multiple imputation

1. A simple imputation, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as "place holders."
2. The "place holder" mean imputations for one variable ("var") are set back to missing.
3. The observed values from the variable "var" in Step 2 are regressed on the other variables in the imputation model. In other words, "var" is the dependent variable in a regression model and all the other variables are independent variables in the regression model.
4. The missing values for "var" are then replaced with predictions (imputations) from the regression model.
5. Steps 2–4 are then repeated for each variable that has missing data.
6. Steps 2–4 are repeated for a number of cycles, with the imputations being updated at each cycle.

Missing values and counterfactual values

- Causal inference (observational studies): counterfactual is missing.

  What if the person were not treated?

Missing values and counterfactual values

- Causal inference (observational studies): counterfactual is missing.

  What if the person were not treated?

- Missing data problem:

  What if the person answered the questions?

Missing values and counterfactual values

- Causal inference (observational studies): counterfactual is missing.

  What if the person were not treated?

  - Matching: find similar control group to approximate the outcome for treated group.

Missing values and counterfactual values

- Causal inference (observational studies): counterfactual is missing.

  What if the person were not treated?

  - Matching: find similar control group to approximate the outcome for treated group.

- Missing data problem:

  What if the person answered the questions?

  - Imputation: find similar observed group to approximate the outcome for missing group.

Missing values and counterfactual values

- Causal inference (observational studies): counterfactual is missing.

  What if the person were not treated?

  - Matching: find similar control group to approximate the outcome for treated group.

- Missing data problem:

  What if the person answered the questions?

  - Imputation: find similar observed group to approximate the outcome for missing group.
  - Causal inference is a missing data problem!

    $\rightsquigarrow$ Counterfactual is missing.

  - Missing data covers more than just causal inference.

Attrition

- A repeated randomized experiment to study campaign ads during 2016 election (Coppock et al, 2020)

Attrition

- A repeated randomized experiment to study campaign ads during 2016 election (Coppock et al, 2020)
- Every week, authors invite 1000 respondents to watch partisan campaign ads (Republican, Democrats, or neutral).

Attrition

- A repeated randomized experiment to study campaign ads during 2016 election (Coppock et al, 2020)
- Every week, authors invite 1000 respondents to watch partisan campaign ads (Republican, Democrats, or neutral).
- Survey questions to measure opinion after watching.

Attrition

- A repeated randomized experiment to study campaign ads during 2016 election (Coppock et al, 2020)
- Every week, authors invite 1000 respondents to watch partisan campaign ads (Republican, Democrats, or neutral).
- Survey questions to measure opinion after watching.
- **Attrition in the data**: treatments caused subjects to stop taking the survey

  $\rightsquigarrow$ A republican voter asked to watch Democratic ads.

Attrition

- A repeated randomized experiment to study campaign ads during 2016 election (Coppock et al, 2020)
- Every week, authors invite 1000 respondents to watch partisan campaign ads (Republican, Democrats, or neutral).
- Survey questions to measure opinion after watching.
- **Attrition in the data**: treatments caused subjects to stop taking the survey

  $\rightsquigarrow$ A republican voter asked to watch Democratic ads.
- Whether voters express favorability towards the party / candidate, and vote intention.

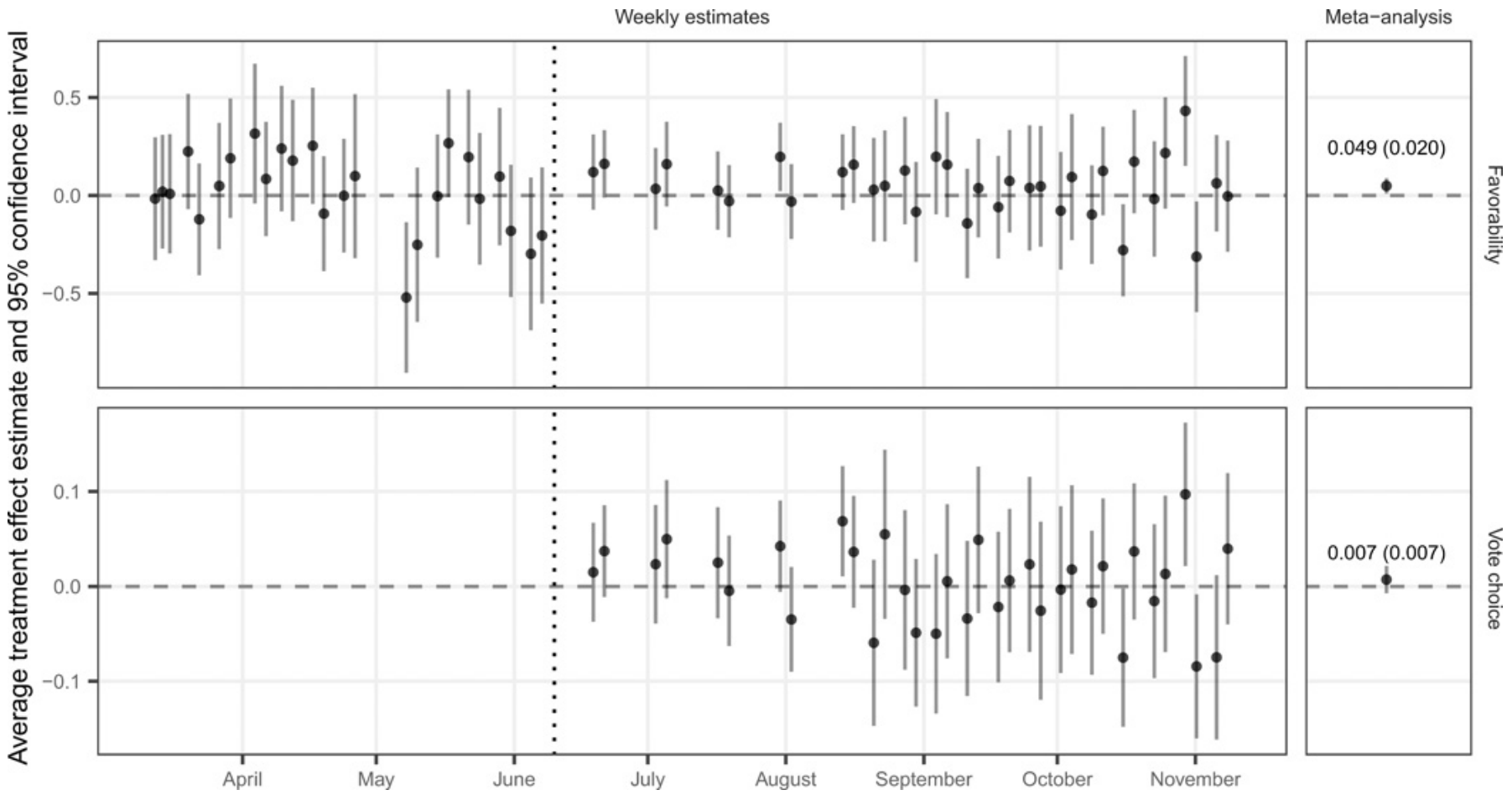
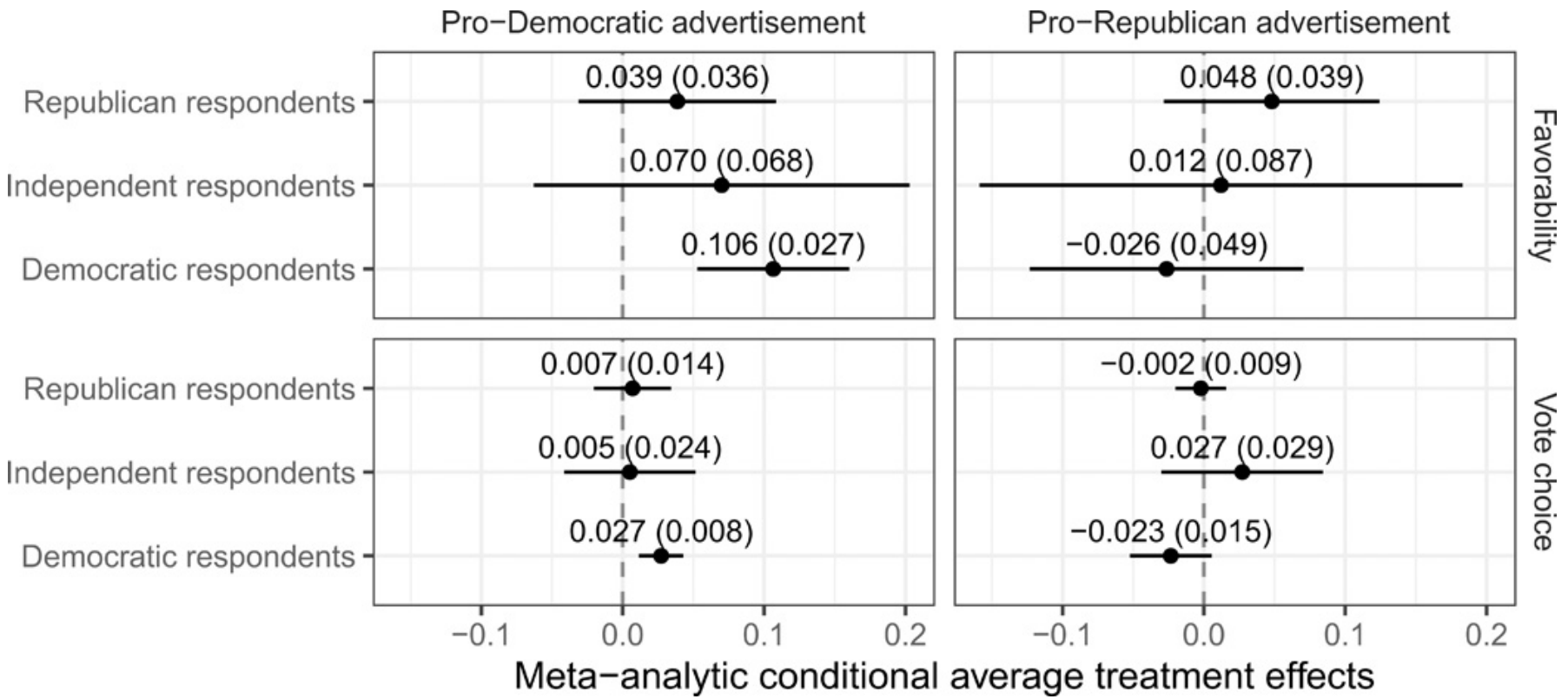

Weekly estimates
Meta−analysis
Average treatment effect estimate and 95% confidence interval
0.5
0.0
−0.5
0.049 (0.020)
Favorability
0.1
0.0
−0.1
0.007 (0.007)
Vote choice
April
May
June
July
August
September
October
November

Figure: Democratic respondents show more significant effects.



Attrition bias

- Attrition: certain types of respondents refused to answer post-treatment survey.

Attrition bias

- Attrition: certain types of respondents refused to answer post-treatment survey.
- One example: Republican respondents got more angrier and left.

  Hence the insignificant results among Republicans.

Attrition bias

- Attrition: certain types of respondents refused to answer post-treatment survey.
- One example: Republican respondents got more angrier and left.

  Hence the insignificant results among Republicans.
- Test potential attrition bias by respondent partisanship:

Attrition bias

- Attrition: certain types of respondents refused to answer post-treatment survey.
- One example: Republican respondents got more angrier and left.

  Hence the insignificant results among Republicans.
- Test potential attrition bias by respondent partisanship:
  - Full set of respondents who began each survey

Attrition bias

- Attrition: certain types of respondents refused to answer post-treatment survey.
- One example: Republican respondents got more angrier and left.

  Hence the insignificant results among Republicans.
- Test potential attrition bias by respondent partisanship:
  - Full set of respondents who began each survey
  - Partial set of respondents who finished each survey