{SCS 23-056} Fall 2023 Summary
IM Agenda

1. Background & Aim

Disinformation dissemination impacts party reputation that matters for reelection chances. Recent societal developments and the advent of social media and networked communication altered its reputational cost and lowered the bar of its spread.

This project use regression to investigate the relationship between the dependent variable "party dissemination of disinformation domestically" and explanatory variables like "political polarization", "electoral system", and etc.

The client develops a reputational cost theory of disinformation dissemination and uses this regression to test it. Specifically, the client wants to test hypotheses like countries with high political polarization and media fractionalization incentivize politicians to use disinformation.

2. Discussion Points

1. For the dependent variable "party dissemination of disinformation domestically", it seems the original data is 5-point Likert scale data.
   --Curious why the client convert it to [-5, 5].
2. The client uses cross-sectional time series data from the Digital Society Project, V-Dem and V-Party from 2000-2021. Some datasets are between the years 2000-2021 and some starts from 2016.
   --better to ask client to present some details of the datasets. She also mentions Unit of Analysis: country-year. Not clear what it is.
3. The client mentions different ways of analyzing the data, for example, panel corrected SEs vs clustered SEs, fixed effects, MICE for dealing with missing data, and etc. She wants to know which one fits best. This might be her main question.
   --fixed effect here means linear mixed model?
4. She mentions that she wants to address the following question:
   "Politicians in democracies should care about reputation for re-election chances. So why risk it by disseminating disinformation?"
   --I feel this question is not related to the project description.

IM Spring 2023

1. Background

Disinformation dissemination impacts party reputation and that matters for reelection chances. Recent societal developments and the advent of social media and networked communication has altered its reputational cost and lowered the bar of its spread.

This project uses regression to investigate the relationship between the dependent variable "party dissemination of disinformation domestically" and explanatory variables such as "political polarization" and "electoral system." The client developed a reputational cost theory of disinformation dissemination and uses regression to test it. Specifically, the client wants to test hypotheses such as do countries with high political polarization and media fractionalization incentivize politicians to use disinformation.

The client is interested in investigating what variables could explain the global variation in disinformation dissemination domestically by political elites including politicians and parties. The client would like to discuss different ways of analyzing the data, for example, panel corrected standard errors, clustered standard errors, fixed effects models, etc.

2. Relevant information presented at meeting

1. The study uses the Digital Society Project dataset, which is an existing dataset based on expert surveys for 179 countries from 2000 to 2021. For each country, 3-5 experts answer over 400 different questions related to its democracy. It is per country panel data from 2000 to 2021 with 8-10 measurements for each country. The client filters the raw dataset to contain 131 countries by setting V-Dem's liberal democracy index larger than 0.42. Some variables in it may not correspond well to the concepts the client is trying to capture. Thus, she also uses other data sources that capture the concepts better.

2. The dependent variable disinformation dissemination is evaluated by aggregating over major parties in each country. Specifically, coders are asked: How often do major political parties and candidates for office use social media to disseminate misleading viewpoints or false information to influence their own population?

3. The client's theorem is about cross-country variation in disinformation dissemination. Independent variables vary across countries and time, for example, polarization likely varies within countries over time, and whether a country has compulsory voting only varies across countries. The cross-country analysis exploits two types of variations.

3. Recommendations for Data and/or Analysis

1. Professor Gu suggested first performing data visualization. Since the same calendar year might mean different things for different countries, calendar year probably does not make sense across countries. Instead of treating the dataset as time series data directly, the client could first take a look at some individual countries by plotting both the dependent variable and independent variables versus time to check if there are similar election cycle patterns among countries with similar political systems. If the election cycle matters, the client could look at the data by election cycles and then check the relationship between the response variable and predictors. After obtaining a general sense based on these plots, the client could integrate things across countries.

2. The client could run different models for different countries if there are enough data for each country. By doing so, interaction terms could be avoided for an easier model explanation. The client could perform quantitative analyses for each country and qualitatively tell stories across countries. Otherwise, countries with similar political systems could be merged into a larger group that share similar election cycle patterns.

3. Each election cycle might be treated as independent observations, ignoring serial correlations.

4. Suppose variables have been adjusted based on the election cycle, the client could run a linear mixed effect model by treating countries as a random effect. Prof. Gu explained mixed effects and random effects to the client.

4. Next Steps

The client will first check the election cycle pattern and the consultant will help her with the data visualization.

Spring 2023 Semester Report

1. Background

Disinformation dissemination impacts party reputation and that matters for reelection chances. Recent societal developments and the advent of social media and networked communication has altered its reputational cost and lowered the bar of its spread.

This project uses regression to investigate the relationship between the dependent variable "party dissemination of disinformation domestically" and explanatory variables such as "political polarization" and "electoral system." The client developed a reputational cost theory of disinformation dissemination and uses regression to test it. Specifically, the client wants to test hypotheses such as do countries with high political polarization and media fractionalization incentivize politicians to use disinformation.

The client is interested in investigating what variables could explain the global variation in disinformation dissemination domestically by political elites including politicians and parties. The client would like to discuss different ways of analyzing the data, for example, panel corrected standard errors, clustered standard errors, fixed effects models, etc.

2. Progress During Current Semester

The client shared her plots with the consultant to check whether the election cycle pattern existed. The information of election years was still missing in her plots and therefore it was difficult to decide the election cycle pattern. The client just figured out a proper data source on election years to merge them in her plots. She also got some feedback from related conferences and decided to use both country/election cycle as well as country/year as units of analysis. She planned to get back to SCS after finishing her prospectus.

Follow-Up Q&A – Round 1

1. 1. On selecting specific time snapshots:

Question 1.1:
Will focusing on specific years reduce sample size and statistical power?

Response:

Suppose that each country had exactly the same data each year (i.e., perfect correlation). **Then having data only for one year is as informative as having it over 10 years.**

Even if one were to include a random effect does not necessarily mean a large increase in the "effective" sample size. If the correlation is high, then the resulting sample size may be closer to the number of countries.

Question 1.2:
[Consultant to SCS Director:]
When looking at the client's data, I looked at how many `country` data points were available for the year 2012. The count was n = 190, which seems as though it would be sufficient for the client's linear model – *but are there any other considerations to bear in mind?*

Response:

I like to often build up to more complex models – so starting on a yearly basis and seeing how the predictors work across years gives an idea of how one might combine the data.

I agree that 190 seems reasonable. They'll only have issues if there are a lot of interaction effects.

2. 2. On `country` as a random effect:

Question 2.1: Can you explain the reasoning for the recommendation to set `country` as a random effect?

Response:

This would only be necessary if the client wants to add `time` into the model as a factor (rather than focusing on specific years).

If there does appear to be a different relationship, that is, there is a need to add interaction terms between `time` and other factors $\implies$ then we would likely have to bring in `country` as a random effect.

However, this will not be necessary and can be disregarded if we proceed with the prior recommendation to focus on time snapshots

3. 3. On another possible approach for managing `time` as a factor:

Question 3.1: If we focus simply on before / after Trump's time in office, how might we address the concern that this will not capture the gradual increased use of social media and disinformation over time?

Response:

One good way of investigating this is to **plot the data for different countries over this time period** $\implies$ to show that this steady increase does indeed increase. If so, then we can determine the necessary adjustments thereafter.

Question 3.2: What if we were to use only the past 5 years or so to increase the quality of the data? (since experts were asked in 2018 to start rating and retrospectively back-casting until 2000; as such, the client believes the 2017 data will be hopefully valid)

Response:

This strategy sounds fine. Note that the plot would help give an idea of a linear trend or something more complex.

Question 3.3: On whether using the past 5 years will provide the client with more data points (rather than focusing on single years), and whether this will allow for more variability in the main explanatory variables (that is, `Political Polarization` and `Media Fractionalization`) ?

Response:

Note that this assumes the predictors behave similarly over years. As such, I would fit each year separately, then combine using year as a nominal factor, interacting with each predictor to assess if it is reasonable to assume predictors behave similarly across years.

Follow-Up Q&A – Round 2

1. Client Question:

I have a question regarding bootstrapping. Am trying to figure out how to increase my sample size for a survey experiment I will be conducting and am wondering the following:

is it possible to bootstrap survey experiments if those involve variables that modify the relationship between the treatment and dependent variable?

Chat gpt tells me the following:

"Yes, it is possible to bootstrap survey experiments, even when the experiment involves variables that modify the relationship between the treatment and dependent variable. Bootstrapping is a versatile resampling technique that can be applied to a variety of experimental setups, including those with interaction effects or other modifications to the relationships between variables.

Here are steps you might consider when using bootstrapping in the context of survey experiments with modifying variables:

1. **Bootstrapping with Interaction Effects:**

- If your survey experiment involves interaction effects between the treatment variable and modifying variables, you can still use bootstrapping to assess the stability of these effects. When resampling, you should maintain the interaction structure in the sampled datasets.

2. **Modifying Variables during Resampling:**

- When you bootstrap, you can modify variables to simulate different scenarios, including variations in the modifying variables. This allows you to explore the robustness of your findings under different conditions.

3. **Assessing Sensitivity of Interaction Effects:**

- Bootstrapping is particularly useful for assessing the sensitivity of interaction effects. By resampling and modifying variables, you can observe how the interaction effects change and evaluate whether they are consistent across different samples.

4. **Creating Confidence Intervals for Interaction Effects:**

- Bootstrapping allows you to create confidence intervals around interaction effects, providing a measure of the uncertainty associated with these effects. This can be valuable in understanding the precision of your estimates.

5. **Testing Hypotheses involving Interactions:**

- Researchers can use bootstrapped samples to perform hypothesis tests related to the interaction effects, helping to determine the statistical significance of these effects.

6. **Model Validation with Interactions:**

- If your survey experiment involves a statistical model with interaction terms,

```
you can use bootstrapping to validate the model by assessing its performance
across different samples generated through resampling.

When working with interaction effects and modifying variables, it's crucial to
carefully consider the theoretical and substantive implications of the
modifications. Additionally, researchers should be aware of potential sources of
bias and confounding and take steps to address these issues in the original
experimental design.

As always, consulting statistical literature, considering best practices in
survey methodology, and seeking guidance from experts in the field can enhance
the rigor and validity of your analysis."
```

Would love to know what you and your director might advise about this.

2. SCS Recommendation

Just to provide some updates after my discussion with the director:

- He did seem a bit concerned about using the Bootstrapping technique for survey data, as this is not precisely what the technique was designed for.
  - While bootstrapping does allow for simulating various scenarios by altering the modifying variables in the resampled data, one risk this poses for survey data is that it may aggravate existing biases in the existing survey (especially if it is not designed in a manner that minimizes biases and confounders).
  - When sharing your current survey design, he acknowledged that the vignette study is well used, but there would likely need to be some careful thought & investment made on how subjects are distributed among the different groups.
- **As such, it would seem most wise to potentially set up a follow-up meeting with one of our directors to thoughtfully review the survey design and make sure that we are maintaining the interaction structures that you are looking for?**
  - Although we have limited availability for the rest of this semester, I would be more than happy to proactively schedule a meeting as soon as their times are open next semester, if you wish!
  - Please feel free to let me know your thoughts, as my only goal is to support you in finishing your PhD studies. I believe this would be the best way to ensure that your study design is empirically validated and would give you a lot of confidence moving forward.