4.

Note: This problem consists of an **ungraded** problem at the end for those who have learned R in *14.310x Data Analysis for Social Scientists* (the data analysis course in the SDS Micromasters Program), and is a preview to those who will be learning to use software to carry out analysis on data sets.

Background:

Understanding why people vote — and how they can be persuaded to do so — is an important concern among social scientists.

To study the effects of applying varying degrees of pressure on voting behavior, a group of political scientists conducted an experiment prior to a primary election. **The sample for the experiment was 20000 voters**.

Voters were randomly assigned to either the control group, or one of 4 treatment groups (i.e. each voter belongs to exactly 1 group):

(You do NOT need to understand the details of the treatments to do the problems.)

0. Treatment 1, **civic duty**: Each voter in this treatment group receives the baseline message “Do your civic duty–vote!".

0. Treatment 2, **Hawthorne effect**: Each voter in this treatment group receives the baseline message in Treatment 1, and in addition, a message that creates a mild form of social pressure : “You are being studied! Your voting behavior will be examined by means of public records. ".

0. Treatment 3, **Self**: Each voter in this treatment group receives all messages as in Treatment groups 1 and 2, and in addition, a message that who votes is public information, and a list of the recent voting record of each registered voter in the recipient's household.

0. Treatment 4, **Neighbors**: Voter in this treatment group receives all messages as in Treatment groups 1, 2, and 3, and in addition, a list of the voting records of all his/her neighbours.

Setup:

You work with a dataset in which each row corresponds to 1 respondent, and there are 6 relevant columns, as described below:
(For your reference, here is the data set: data_gerber_trunc.csv, though you do not need to work directly with to answer the graded problems below.)

- voting, an indicator variable to indicate whether the respondent voted in the election, i.e. $\mathrm{voting} = 1$ if the respondent voted in the election, and $\mathrm{voting} = 0$ otherwise.

- $\mathrm{civicduty}$, an indicator variable to indicate whether the respondent is assigned to **Treatment Group 1** , the “Civic Duty" group, i.e. $\mathrm{civicduty} = 1$ if the respondent is in Treatment Group 1, $\mathrm{civicduty} = 0$ otherwise.

- $\mathrm{hawthorne}$, an indicator variable to indicate whether the respondent is assigned to **Treatment Group 2** , the “hawthorn effect" group, i.e. $\mathrm{hawthorn} = 1$ if the respondent is in Treatment Group 2, $\mathrm{hawthorn} = 0$ otherwise.

- self, an indicator variable to indicate whether the respondent is assigned to **Treatment Group 3** , the “self" group, i.e. $\mathrm{self} = 1$ if the respondent is in Treatment Group 3, $\mathrm{self} = 0$ otherwise.

- $\mathrm{neighbors}$, an indicator variable to indicate whether the respondent is assigned to **Treatment Group 4** , the “neighbor " group, i.e. $\mathrm{neighbors} = 1$ if the respondent is in Treatment Group 4, $\mathrm{neighbors} = 0$ otherwise.

- $\mathrm{control}$, an indicator variable to indicate whether the respondent is assigned to the control group i.e. $\mathrm{control} = 1$ if the respondent is in the control group, $\mathrm{control} = 0$ otherwise.

Model:

You will regress the variable voting on the four treatment variables $\mathrm{civicduty}, \mathrm{hawthorne}, \mathrm{self}, \mathrm{neighbors}$, without any other control variables. That is, you will perform a linear regression using voting as the response variable, and $\mathrm{civicduty}$, $\mathrm{hawthorne}$, self, $\mathrm{neighbors}$ as the explaining variables, using the linear relation:

$$\mathrm{voting} \quad = \quad \beta_0 + \beta_1 \mathrm{civicduty} + \beta_2 \mathrm{hawthorne} + \beta_3 \mathrm{self} + \beta_4 \mathrm{neighbors} + \epsilon$$

Generating Speech Output

where $\epsilon = \mathcal{N}(0, \sigma^2)$ for some unknown σ^2 by assumption.

Goal:

Our goal is to estimate the variance $\mathsf{Var}\left(\widehat{\beta_4} - \widehat{\beta_3}\right)$ of the difference $\widehat{\beta_4} - \widehat{\beta_3}$ of the two estimators.

(This would be useful, for example, in a test for the null hypothesis $H_0 : \beta_4 - \beta_3 = 0$.)

Covariance Matrix

2.5/2.5 points (graded)
Use the setup described as above.

In this problem, you will find the covariance matrix Σ for the least square estimator $\widehat{\beta}$ where $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$.

0. Recall the assumption in the regression model that $\epsilon \sim \mathcal{N}(0, \sigma^2)$. What kind of distribution does $\widehat{\beta}$ follow?

- (selected) Multivariate Gaussian distribution with mean β
- t-distribution with 4 degrees of freedom
- t-distribution with 19999 degrees of freedom
- χ^2-distribution with 5 degrees of freedom
- χ^2-distribution with 19999 degrees of freedom

Generating Speech Output

0. What is the shape of the covariance matrix?

- Σ is 20000×20000 matrix.
- **Σ is 5×5 matrix.**
- Σ is 5×20000 matrix.
- Σ is 20000×5 matrix.

✔

0. Recall the assumption in the regression model that $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Define the design matrix $\mathbb{X}$ of covariates to be:

$$\mathbb{X} = \begin{pmatrix} 1 & \text{civicduty}_1 & \text{hawthorn}_1 & \text{self}_1 & \text{neighbors}_1 \\ 1 & \text{civicduty}_2 & \text{hawthorn}_2 & \text{self}_2 & \text{neighbors}_2 \\ & & \vdots & & \\ 1 & \text{civicduty}_{20000} & \text{hawthorn}_{20000} & \text{self}_{20000} & \text{neighbors}_{20000} \end{pmatrix}$$

where each row i corresponds to row i in the data set (and hence corresponds to the i^{th} respondent.)

Assuming $\mathbb{X}$ is deterministic in our regression model, what is the covariance matrix Σ of the least square estimator $\hat{\beta}$?

- $\Sigma = \sigma^2 (\mathbb{X}^T \mathbb{X})$
- **$\Sigma = \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1}$**
- $\Sigma = \sigma^2 (\mathbb{X} \mathbb{X}^T)$

Generating Speech Output

- $\Sigma = \sigma^2 (\mathbb{X}\mathbb{X}^T)^{-1}$

✔

0. Which of the following equals to $\mathsf{Var}\left(\widehat{\beta_4} - \widehat{\beta_3}\right)$? In all choices below, $\vec{u}^T = [0 \quad 0 \quad 0 \quad -1 \quad 1]$.

- $\vec{u}^T \Sigma$
- $\Sigma \vec{u}$
- $\vec{u}^T \Sigma \vec{u}$
- $\vec{u} \Sigma \vec{u}^T$

✔

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

(Optional Ungraded) Estimate the Variance

0 points possible (ungraded)

Using the (correct) results from the problem above, estimate the variance $\mathsf{Var}\left(\widehat{\beta_4} - \widehat{\beta_3}\right)$ based on the given data.

(Enter a numerical answer accurate to at least 4 decimal places.)

Estimate of $\mathsf{Var}\left(\widehat{\beta_4} - \widehat{\beta_3}\right)$: 0.0001952786 ✔ **Answer:** 0.000195

Generating Speech Output

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

Error and Bug Reports/Technical Issues

Topic: Final exam:Final Exam / 4.

Hide Discussion

Add a Post

Show all posts ▼ by recent activity ▼

Python code for ungraded question Here is my Python code: There was a problem with indexing in my code. Hence, I deleted that. This is modified working code import numpy as np data = np.genfromtxt(open(...	6
My Answers to this question(3)	8
[Staff]: (Optional Ungraded) Estimate the Variance - motivation Hi Staff, this problem doesn't affect the Exam score and I wouldn't even know neither the correct solution nor even whether I did it right. It is a bit sad.	4
Requirements for the UNGRADED exercise Just to be sure, the UNGRADED exercise needs the actual data and the estimation of the variance of the error term, right ?	2

Generating Speech Output