Department of Economics

410 Arps Hall
1945 North High Street
Columbus, OH 43210-1172

Prof. Paul J. Healy
Email: healy.52@osu.edu

ECON 8877
Fall 2023

Regression Assignment

Due by Monday, December 11$^{\text{th}}$ 2023 at 4:00 PM

First, from the course website, download FakeData.csv. This data is fake, but it supposed to represent some hypothetical experiment you might have run. The columns are as follows:

SubjectID: In each there are 120 total subjects. This is their ID number $i \in \{1, \ldots, 120\}$.

DecisionID: Each subject makes 15 decisions during the experiment (say, contributions to a public good). This is the decision ID number $t \in \{1, \ldots, 15\}$.

Constant: This is just a column of ones.

Treatment: The first 60 subjects are in the control ($T_i = 0$) and the last 60 are in the treatment ($T_i = 1$).

SessionID: This is the ID number of the session that they participated in. Denote it by $S_i \in \{1, \ldots, 20\}$.

SwitchPoint: We measured risk aversion using an MPL and their switch point is the row on which they switched from Option A to Option B on the list. It is a number $W_i \in \{0, \ldots, 10\}$.

Gender: Odd-numbered subjects reported male as their gender ($G_i = 0$), and even-numbered subjects reported female ($G_i = 1$). Those were the only two genders reported.

Complexity: Each of the 15 decisions has a different complexity level. The complexity of decision problem t is simply $X_t = t$, which ranges from 1 (simplest) to 15 (most complex). In other words, the decisions become more and more complex during the 15-question experiment.

There is a missing variable that you're going to create. It's their choice on each decision, called Choice and denoted y_{it}.

1. First, we test OLS with homoskedasticity

   (a) For each row in the data, generate an error ϵ_{it} drawn iid from $N(0, 7.5)$ (where 7.5 is the standard deviation, not variance).

   (b) Next, generate a column of data called "Choice" using the formula

   $$y_{it} = \beta_0 + \beta_1 T_i + \beta_2 S_i + \beta_3 W_i + \beta_4 G_i + \beta_5 X_t + \epsilon_{it}.$$

   Use the following "true" values for β:

   $$\beta = (\beta_0, \ldots, \beta_5) = (1, 0.75, 0, 0.10, 0, -0.10)$$

   (c) Run a regression to get $\hat{\beta}$. Note the p-values as well.

   (d) Repeat the previous three steps 100 times.

(e) For each β_j, what is the median value of $\hat{\beta}_j$ (out of 100)? How close is it to the true value of β_j? This is an estimate of the bias in your regression.

(f) For each β_j, count the fraction of the 100 regressions for which the p-value was below 0.05. This is the estimated power of that test.

2. Re-run that excercise, but with heterskedasticity. Specifically, let the standard deviation of ϵ_{it} equal the Decision ID number. So, for example, everyone subject's choice on decision 8 has noise $\epsilon_{it} \sim N(0, 8)$. Re-generate these errors with that structure, and then re-generate the Choice data using the same linear equation as above.

   (a) This time, run 100 regressions of OLS, then 100 each with HC0, HC1, HC2, and HC3.

   (b) Which methods give unbiased estimates?

   (c) Among methods with unbiased estimates, which has the most power (when $\beta_j \neq 0$).

3. Now let's add session effects. The way to do this is first create a random column $\epsilon_{it} \sim N(0, 7.5)$ of uncorrelated errors. Then, for each session S, generate a *single* random number $u_S \sim N(0, 7.5)$. Finally, generate the Choice variable using the linear equation

$$y_{it} = \beta_0 + \beta_1 T_i + \beta_2 S_i + \beta_3 W_i + \beta_4 G_i + \beta_5 X_t + u_{S_i} + \epsilon_{it}.$$

   Use the same values for β as above.

   (a) Run 100 regressions of OLS, ignoring the session effects. Are your estimates biased? Is your power affected? Are the tests still valid (rejecting 5% of the time when $\beta = 0$)? Pay special attention to $\hat{\beta}_2$, which measures session effects directly.

   (b) Run 100 regressions with clustering at the session level. Does this fix any bias? Does it give better power? Are tests valid?

   (c) Run 100 regressions using random effects. Note that the u_S is highly correlated with one of the X columns, so we're told this is not valid. Analyze bias, power, and validity.

   (d) Run 100 regressions with fixed effects. Analyze bias, power, and validity.

   (e) Our regression equation contains $\beta_2 S_i$, which assumes a linear relationship between the session ID and choices. This seems a bit silly. Instead, one could create dummy variables for each of the sessions. So, replace S_i with 19 dummy variables (leaving session 1 as the omitted session). Run 100 regressions of regular OLS (with no clustering, fixed effects, or random effects) but with this dummy variable specification. Presumably this drastically reduces power since we've added so many variables. Analyze bias, power, and validity.

4. Repeat the previous exercise, but this time with individual effects instead of session effects. In other words, one value of $u_i \sim N(0, 7.5)$ is drawn for each individual i. For this part, drop $\beta_2 S_i$ from the regression.

   (a) Run 100 OLS regressions and analyze bias, power, and validity.

   (b) Run 100 regressions with clustering at the individual level. Analyze bias, power, and validity.

   (c) Run 100 regressions with random effects. Analyze bias, power, and validity. Since there's no correlation with any independent variable, this should now be a valid procedure.

   (d) Run 100 regressions with fixed effects. Analyze bias, power, and validity.