PS405: Linear Models

Problem Set 3

Professor: Nicole Wilson TA: Artur Baranov

Due: Thursday, February 6, 2025

Please submit your write-up, including *any* code that you ran, via Canvas. We suggest using Quarto or `RMarkdown`, but any software that renders R code into a PDF document will work. I have provided an `RMarkdown` template if you prefer to use it (note that the process/format is very similar to Quarto, if you are familiar with that). This should be completed *before* class begins at 9:30 AM. Please save your write-up with easy-to-recognize file names (e.g., `ps1_wilson.pdf` or `ps1_baranov.pdf`). Late submissions will not be accepted without special permission from the course instructors. Where instructed to follow a specific coding procedure in `R`, such as the creation of a new function or a data generating process, please include the code in a code chunk in your write up. For example, you can use `echo = TRUE` in the chunk header in Quarto or `RMarkdown` to ensure your code is visible in the output. You do not need to submit the raw file you use to produce your PDF, but the course instructors may request it on a case-by-case basis.

Please post all questions to the Canvas anonymous discussion forum.

Problem 1

Bring in the data from `exp_data.csv`. This data is from an experiment to determine what factors shape whether legislators respond to emails from citizens.

Variables:

- `responded` is whether or not a politician responded to an email. `0` indicated they did not respond, and `1` indicates that they did.
- `treat_out` is whether the email was framed as being from someone outside of the legislators district.
- `leg_black` is whether the legislator receiving the email is Black
- `medianhhincom` the median household income in a legislator's district.
- `statessquireindex` is a measure of professionalization of state legislatures that ranges from 0 to 100.
- `south` is an indicator of whether the legislator is from the south
- `totalpop` is the total population of the legislator's district

(a) First, we will answer the question, are legislators more or less likely to respond to an email from someone outside of their district?

(i) Run a regression and present your findings in a table. In your table make sure that you:

- Have a meaningful title
- Have meaningful names for your variables
- Indicate either the p-value or the confidence interval

(ii) Then provide a concise but accurate interpretation of the output. Make sure to mention the magnitude of the effect and the statistical significance. Think carefully about what your outcome is and the scale it is on. Think about what a "one-unit increase" substantively means in this case. Pretend you are writing an academic paper and you are describing your finding to your audience.

(b) Now, we will answer a slightly more complicated question. Does the effect of an out-of-district email on whether or not a legislator responds vary by the race (specifically, Black or not) of the politician?

(i) Again, run a regression, putting your results in a clear and well-formatted table.

(ii) Again, provide an interpretation of the output. Follow the same instructions as in (a)(ii). Think carefully about what the coefficients mean.

(c) Next, we want to explore how the median household income of the legislator's district is associated their likelihood of responding. I use the word "associated" here because unlike the out-of-district email experimental treatment we looked at before, median household income is not randomly assigned, so we should be cautious about our interpretation.

(i) Again, run a regression. We will use a polynomial (use `poly()` or `I()`) to model a quadratic relationship (i.e., include x and x^2).

(ii) This time, we are going to plot our data instead of outputting a table. You probably want to make use of the `ggeffects` package. Be sure to include confidence intervals for the marginal effects (`geom_ribbon` may be helpful).

(iii) Briefly interpret the output. You can focus on the general nature of the relationship, rather than specific numbers, since the coefficients can be difficult to interpret when modeling non-linearities.

(iv) Your colleague challenges you and says that you are violating the linearity assumption for OLS by running this regression. How do you respond?

(d) Don't worry, we are almost done with this data. Now we are going to run one model with several variables at once. We are interested in the likelihood that a legislator responds to an email based on:

- whether the legislator is black
- the state Squire Index
- whether the legislator is from the South

- the population of the legislators district

(i) First, to help us better compare the variables in our model, we are going to standardize them. We can leave the binary variables alone, but we will divide the continuous variables by two standard deviations (as you read in Gelman, Hill, and Vehtari, this makes it easier to compare them to our binary variables). It's probably best practice to create new variables but add something to the name to indicate that they are standardized versions. Just make sure you use the standardized ones in your regression.

(ii) Next, run your regression with the four independent variables of interest.

(iii) Then, make a plot of the effects with 90% confidence intervals. You can use `tidy()` from the `broom` package to add to your model information the upper and lower bounds you will need for your plot. Then, you can use `geom_errorbar()` (or `geom_errorbarh()` depending on the orientation of your plot) to add the confidence intervals. Put a dotted line at 0 so your audience can easily see whether the confidence interval crosses it. Like with your tables, make sure your variables are clearly labelled.

(iv) Which of these factors seem to matter the most, all else equal?