PS405: Linear Models

Problem Set 2

Professor: Nicole Wilson TA: Artur Baranov

Due: Thursday, January 30, 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

In this problem, we will investigate analyzing regressions using interaction terms. To do so, we will replicate the analysis in Baldwin's (2013) article "Why Vote with the Chief? Political Connections and Public Goods Provision in Zambia." In this paper, Baldwin argues that citizens support political candidates who are personally connected to the local chief, not because of clientelism, but because politicians with personal connections are more effective at providing local public goods. To test her theory, she runs a survey experiment where she asks respondents about their likelihood of voting for a hypothetical MP candidate. Respondents in the treatment group are provided information about the local chief's opinion about the MP candidate, while respondents in the control group are not provided such information.

Please download `zambia.dta` from the course website. For this problem, we will use the following variables:

- `treatment`: an indicator variable that indicates whether the respondent was told that their chief supported the MP candidate (1) or were not provided information about their chief's preferences (0)
- `voteMP`: an indicator variable that takes the value 1 if the respondent indicate they were likely to vote for an MP and 0 otherwise

- `finprim`: an indicator variable that takes the value of 1 if the respondent stated that they had completed primary school and a value of 0 otherwise
- `age`: age of the respondent in years
- `chiefMPimportant` - An indicator variable that takes the value of 1 if the respondent rated the influence of both their chief and their MP highly and a 0 otherwise
- `chiefchangetreat` - An indicator variable that takes a value of 1 if the respondent believed that the chief would treat them differently based on how they cast their ballot, and a value of 0 otherwise

a) Regress `voteMP` on the treatment variable. Report the estimates in a table, and briefly interpret the effect. The data is saved in `.dta` format (from Stata). You can use the `foreign` package to bring it into R (look at the documentation to figure out which function to use).

b) In Baldwin's article, her main quantity of interest is not the main treatment effect that was estimated in Part a). Instead, Baldwin's main quantity of interest is the interaction between the treatment and respondents' answer to two different survey questions: (1) whether or not respondents think both the chief and MP are important, and (2) whether or not respondents fear retribution if they do not vote for the chief's preferred candidate. If the interaction effect between perceived importance of the chief and the treatment is significant, but the interaction effect between perceived fear and treatment is insignifcant, there is support for Baldwin's theory, but not classical theories of clientelism.

i) We will now introduce interaction terms into our model. Before estimating your models, write out two population regression models. In the first model, interact the treatment variable with the variable measuring whether or not both the chief and the MP are important (`chiefMPimportant`). In the second model, interact the treatment variable with the variable measuring whether or not respondents feared retaliation from the chief for their vote choice (`chiefchangetreat`).

ii) Now, estimate the two models that you wrote out in item 2b(i) and report your results in a well-formatted regression table. Interpret the treatment effect (both the main effect and the effect by sub-group), but be sure to specify which sub-group each coefficient refers to. (Hint: It can help to reference the equation you wrote out in part 2b(i) to see which β terms will be included when the right-hand-side variables take a value of either 0 or 1.)

iii) Using your regression model output, complete the following 2x2 table by reporting the estimates of the conditional expectation of Y by treatment status and belief that both the chief and the MP are important. This table should show the conditional expectation (i.e., likelihood of voting for the hypothetical MP) among respondents in the treatment and control condition, subset by those who do or do not believe that both the chief and MP are important.

	Both Important	Both Not Important
Control Group		
Treatment Group		

iv) Now, use your model with interaction effects to replicate the point estimates of Figure 1 on page 804 of Baldwin's article (only the quadrant for the "Importance of both leaders"), including the point estimates, the connecting line that visualizes the slope associated with moving from treatment to control for each sub-group, and the confidence intervals. It doesn't have to look exactly the same, but should communicate the same information. Please label your plot, and include a legend that identifies sub-group membership. You may find the `ggeffects` package useful here. (Other potential options include the `emmip()` function in the `emmeans` package or the `interaction.plot()` function from the `stats` package).

Problem 2

For this problem, we'll look at some fictitious data on high schools in California. We have the following variables:

1. `students`: The number of students at each high school.
2. `size.small`: A dummy variable that indicates whether the high school has fewer than 500 students.
3. `size.medium`: A dummy variable that indicates whether the high school has between 500 and 700 students.
4. `size.large`: A dummy variable that indicates whether the high school has more than 700 students.
5. `expenditure`: The average expenditure per student at each high school (in thousand USD).
6. `teachers`: The number of teachers at each high school.
7. `math`: A math test score.
8. `read`: A reading test score.

(a) Imagine the data get into the hands of your friend who is interested in estimating the relationship between the size of high schools and the average math score of high school students in California. She proposes the following model:

$$\hat{\text{math}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{size.small}_i + \hat{\beta}_2 \text{size.medium}_i + \hat{\beta}_3 \text{size.large}_i$$

for all (i = 1, ..., 300) schools in the data set. Do you think there's any problem with the way your friend specified her model? Run the model specified above in `R`. How would we explain what is displayed and interpret the results? Make a nicely formatted regression table for your output. This is saved as an `RData` file, which you can bring in using the `load()` function.

(b) You decide that you are more interested in the reading scores. You still include the dummies for school size, but you think it would be better to have small high schools as the reference category. In addition, you think that there is probably a nonlinear relationship between the number of teachers and the reading score (since having more teachers is generally better, but there are likely decreasing returns). Check the data first to see if it seems like there appears to be this assumed nonlinearity. Run two models: one with a linear relationship between number of teachers and reading scores, and one with a second-order polynomial (i.e., include an x^2). You may find the `poly()` function useful. Compare the two models. What are the differences in the output? Which do you think is better?