Midterm project

EC 421: Introduction to Econometrics

Due *before* midnight on Sunday, 21 February 2021

Instructions

INTEGRITY: Groups can either have **one or two members**. Only one person needs to submit your final document. If you are suspected of cheating in any way (for example, copying from someone else), then you will receive a zero and fail this course. We will report you to the dean.

GRADING: Your grade for this project will be based upon the accuracy of your answers *and* how well you explain/illustrate your answers. We value short, accurate answers over long, meandering answers. Edit your answers! Make your figures look good (including titles and labeled axes)!

EMAIL POLICY: Do not ask the GEs, the instructor, or people outside your group for help coding or for help answering these questions. You may only ask **clarifying** questions. **Use Google and the course's materials** (lectures, labs, notes, assignment keys).

DUE: One member of your group must upload your answer on Canvas *before* midnight on Sunday, 21 February 2021. All members of the group must be listed on the submission.

IMPORTANT: As with your homework, you must submit **two files**:

1. your typed responses/answers to the question **with figures and regression results** (in a Word file or something similar).
2. the R script you used to generate your answers.

If you are using RMarkdown, you can submit a single file.

README! The last page has a table that describes each variable in the dataset (`data-project-01.csv`).

HELP! The questions below ask for several figures. If you need help creating the figures, check out these `ggplot2` resources (in addition to the class and lab materials):

- An intro to `ggplot2`
- A tutorial on customizing `ggplot2` figures
- The `ggplot2` website (with a `ggplot2` cheatsheet!)

Questions

01. Load the data (`data-project-01.csv`). Summarize and describe the variables in the dataset. Your answer should include:

- Which countries show up in the data? What are their percentages (share of the sample)?
- How many provinces and "tasters" show up in the data? How many varieties are there?
- How skewed are the distributions of price and points?
- Create at least two figures (graphs) that individually summarize the variables `price` and `points`.
- Create at least three figures (graphs) that demonstrate how the key variables relate to each other and other variables (*i.e.*, `variety`, `country`, `province`, `taster_name`).

Explain your decisions on summarizing the data. What do you learn about potential relationships?

02. Does the distribution of `price` appear to be the same across the countries? What about the distribution of `points` across the countries? Use figures (plots) to justify your answer. Explain your answer.

03. We are going to treat `price` as our outcome variable. Regress the price (of the wine bottle) on an intercept, the bottle's rating (`points`) and its country of origin. Explain what the coefficients mean and comment on their statistical significance.

04. Create a scatter plot with the residuals from **03** on the y axis the rating (`points`) on the x axis.

05. Does the scatter plot from **04** suggest that **heteroskedasticity** may be present? Explain your answer.

06. Does the scatter plot from **04** suggest that there are any issues with **your specification**? Explain.

07. Do you think your regression in **03** could suffer from omitted-variable bias? Explain why or why not, using the requirements for omitted-variable bias as part of your explanation.

08. Now include an interaction between country and rating (`points`). Interpret this interaction and comment on the statistical significance. Does this interaction seem important? Explain.

09. Up to this point, we've told you which regressions to run. And we've stuck with pretty simple regressions (*e.g.*, regress `y` on `x1 + x2`). Now we want you to explore the actual complexity of econometric/statistical analyses. **Estimate three new models.** These models **should not match** your previous models (in **03** and **08**). Be creative!! Across these three new models, you should include (at least once):

- a log-transformed outcome variable (*i.e.*, use `log()`)
- new/additional explanatory variables (you've only used two of the variables in the dataset)
- an interaction

10. How did you choose your specifications in **09**? Explain your decision making.

11. Which of your new models is the "best"—*i.e.*, if you must choose one model, which would you choose? Why? Explain your reasoning.

12. For your "best" model (chosen in **11**): Interpret the coefficients and comment on their statistical significance.

13. Do you *trust* the estimates from your *best model*? Explain why/why not.

14. Create a scatter plot with the actual prices (`price`) on the *y* axis the the predictions (see the `fitted.values` outputted by `lm()` or use the `fitted()` or `predict()` functions) on the *x* axis. How well does your model predict the true price? Explain your answer.

15. Write up a one-paragraph summary of what you've learned about pricing in wine. Base your findings on the figures and regressions in this project.

Variable	Description
price	The price of the bottle of wine (US dollars).
points	The points given in the review (more points means better wine).
variety	The wine's variety (think: type).
country	The country that produced the wine.
province	The province that produced the wine.
taster_name	The name of the person who tasted/reviewed/rated the bottle of wine.