Midterm project

EC 421: Introduction to Econometrics

Due *before* midnight on Sunday, 21 February 2021

Solutions

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`).

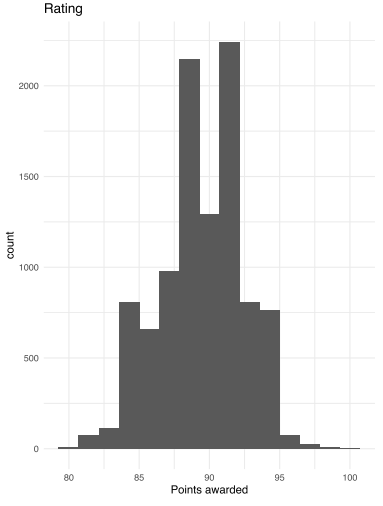
Explain your decisions on summarizing the data. What do you learn about potential relationships?

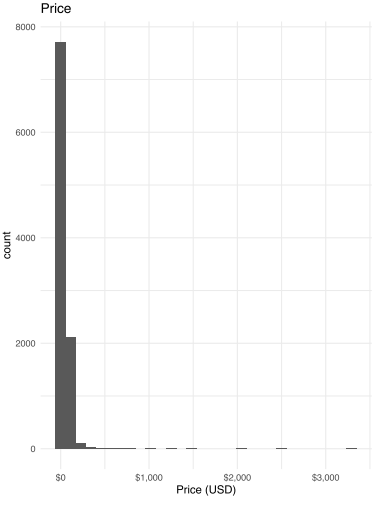
Answer

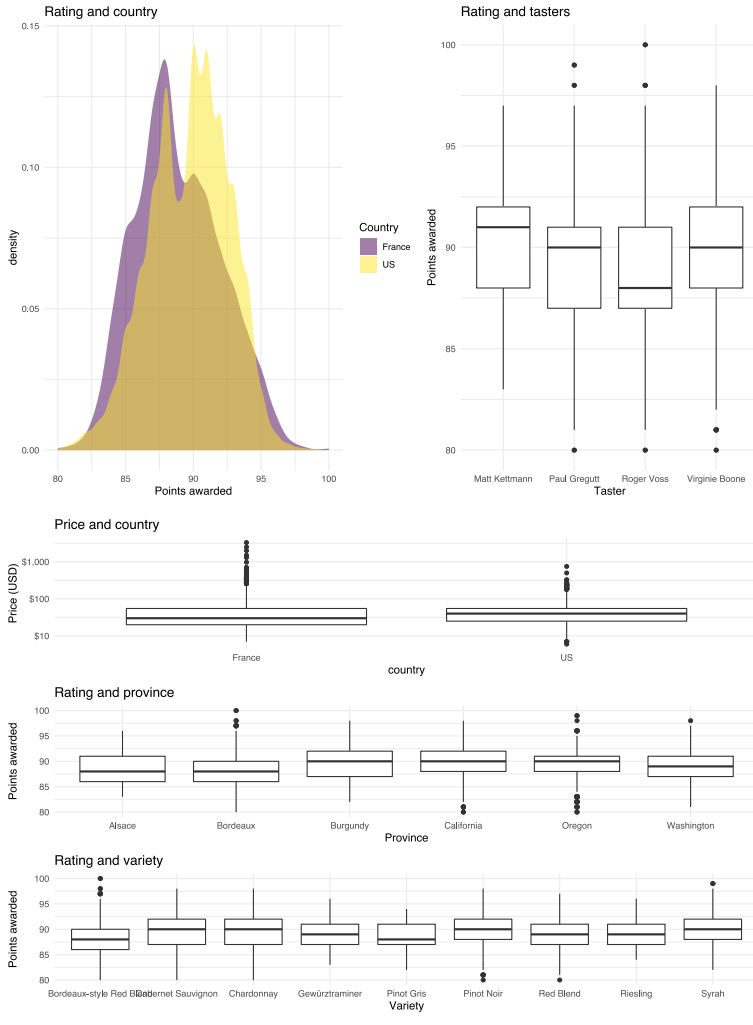
There are 2 **countries**. The US represents 71.6% of the observations, and France represents 28.4% of the observations.

There are 6 unique provinces (3 per country), 4 "tasters", and 9 varieties.

As we see below, the distribution of points is not very skewed (pretty symmetric), but the distribution of price is **very** skewed.







As you can see above, the two countries differ slightly in their ratings, tasters have different distributions of ratings, and there are slight differences across regions and varieties (in rating).

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.

Answer As we saw above, the ratings for US wines tend to be slightly higher. The median price for French wine is slightly lower than that of US wine, but French wine has a much longer tail (higher prices).

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.

Answer

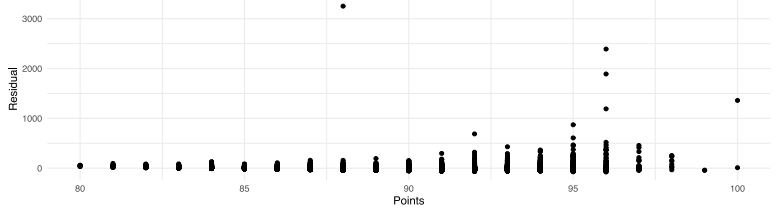
```
#> OLS estimation, Dep. Var.: price
#> Observations: 10,000
#> Standard-errors: Standard
#>              Estimate Std. Error t value  Pr(>|t|)
#> (Intercept)  -623.02    18.5380  -33.61 < 2.2e-16 ***
#> points          7.63     0.2082   36.66 < 2.2e-16 ***
#> countryUS     -17.14     1.3758  -12.46 < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> RMSE: 61.6   Adj. R2: 0.123133
```

Both coefficients are highly statistically significant at the 5% levels.

The coefficient on `points` tells us every additional point is associated with a $7.63 increase in the price (holding all else constant). The coefficient on `countryUS` suggests that a bottle of wine from the US is, on average, $17.14 cheaper than a bottle from France (holding all else constant).

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

Answer



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

Answer The figure above seems to suggest we could have heteroskedasticity: Higher values of points appear to more variance in their residuals.

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

Answer Perhaps: The residuals on the left-hand side of the plot (lower ratings) have means above zero—likely caused by outliers.

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.

Answer Probably: There are likely omitted variables that (1) affect price (many variables affect price) and (2) are correlated with points (again seems like there are many options).

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.

Answer

```
#> OLS estimation, Dep. Var.: price
#> Observations: 10,000
#> Standard-errors: Standard
#>                    Estimate Std. Error t value  Pr(>|t|)
#> (Intercept)      -1208.600    31.7940  -38.01 < 2.2e-16 ***
#> points              14.218     0.3575   39.77 < 2.2e-16 ***
#> countryUS          850.280    38.7550   21.94 < 2.2e-16 ***
#> points:countryUS    -9.729     0.4344  -22.40 < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> RMSE: 60.1   Adj. R2: 0.164945
```

The coefficient on the interaction tells us how each additional point differentially affects the price for US wines, relative to the effect of additional points for French wines. Specifically, it tells us that the increase in price from an additional point is $9.73 less than the effect for French wines.

This effect is highly statistically significant *and* has a large coefficient (it is economically meaningful). Thus, it seems pretty important.

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

Answer Lots of options here. There should be a logged variable, there shoyld be additional explanatory variables (province, taster, variety), and there should be new interactions .

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

Answer Looking for reasonable justification.

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.

Answer Looking for sound reasoning—especially with inference in mind. Not just R squared.

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

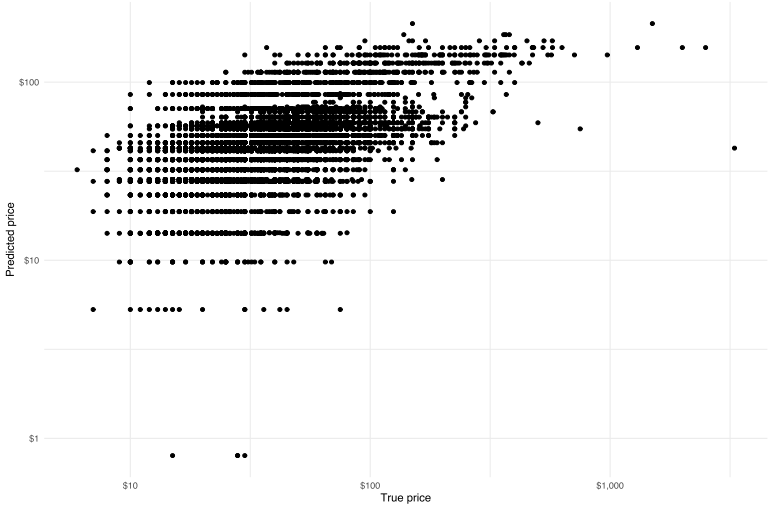
Answer Depends on the best model.

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

Answer Should still worry about omitted-variable bias, but perhaps the model still helps describe the relationship between price and rating for wine. Also could wonder about reverse causality: Does price affect rating?

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.

Answer



There's definitely a positive correlation (a very low). In my figure, there is a lot of room for improvement.

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.

Answer Just want good reasoning.

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.