University of California, Los Angeles, CA
ECON 104, Fall 2023

Project 1

Due: October 20 2023

For this project, you will work any dataset you like, however, it must contain at least 5 different predictors and one response variable (continuous). Your task will be to produce a descriptive model and summary of findings based on your hypothesis by following the steps outlined below. The suggested datasets for this project from the AER package are:

- MurderRates
- HousePrices
- CASchools
- CollegeDistance
- CPS1985/CPS1988

However you may use ANY cross-sectional dataset your group chooses. There are many publicly available datasets through Google DataSet Search, Kaggle, AER - base R datasets, etc. Do not use any datasets used in class, lab or practice sets.

---

1. Briefly discuss the question you are trying to answer with your model.
2. Give a description of your dataset including:
   (a) Citing the dataset
   (b) A summary of what the dataset is about
   (c) Descriptive analysis of your variables. This should include histograms with fitted distributions and correlation matrix, and the five number summary (which can be accompanied by a boxplot). **All** figures must include comments including, but not limited to, the distribution, central tendency and dispersion of the variables.
   (d) Possible violation of the regression assumptions.

3. Estimate a multiple linear regression model that includes main effects only (i.e. no interactions or higher order terms). This is our baseline model.

   (a) Comment on the statistical and economic significance of your individual estimates and provide an interpretation of the estimates obtained. Include any anomalies present if any such as unrealistic magnitudes, unexpected signs, etc.

   (b) Comment on the overall fit of the model and how 1(d) might interfere with this. Comment also on the overall statistical significance of the model.

   FEATURE SELECTION

4. Test the model in (3) for multicollinearity using VIF. Based on this test remove remove the appropriate variables and estimate a new regression model based on these findings. Be sure to justify your reason/criteria for removal.

5. Using AIC or Schwartz Criterion, determine which subset of predictors you will keep and generate a new model. Comment on the performance of this model compared to the one in (3)

6. Using the model in (5) plot the residuals versus its fitted values, $\hat{y}$ and comment on your results.

7. Perform a RESET test on the model in (5) and comment on the results.

8. Using the appropriate method learnt in class, test the model in (4) for heteroskedasticity and comment on the conclusion. If it is present, correct the model before moving on. Based on the results in (c) or (d), this might be helpful in transforming the model in the event that its functional form presents an issue.

9. Using a combination of the results from the previous steps, estimate a model based on your findings which includes interaction terms or higher power terms (if necessary). You may need to use forward or backward selection for this. Comment on the performance of this model. compared to your other models. Make sure to use AIC and Schwartz criterion for model comparison.

10. Provide a short 1 paragraph summary of your overall conclusion, findings, and recommendations not previously stated above.

Please note that submissions will be graded for aesthetics also; this includes but is not limited to code running off the page, including unnecessary messages, errors, and warnings, including extraneous code etc. Whether it is done using R, Python or another language, all code must be included in the output and all documents must be a compiled PDF.