CODING ASSIGNMENT
PREDOCTORAL FELLOWSHIP APPLICATION 2020-21

Instructions. This assignment has the goal of assessing the coding skills of applicants for the position of SIEPR predoctoral fellow for 2020-21. We are looking for work that carefully executes the assignment with clear documentation of how decisions were made. If you can't figure something out, that is OK. Just document your confusion, make a decision of how to proceed, and move on. Be resourceful: look in the documentation files, use google, etc.

Please submit your answers in a single compressed file. Collect all written answers in a single nicely formatted pdf document generated in latex. For codes, we should be able to execute them changing no more than the name of the current directory. Please comment your code and be sure to note things that you were not sure about and the solution you chose. Please submit your results to `name@email.edu` and `name@email.edu` by March 1, 2020, or before.

Data task

We wish to study bidding behavior in auctions for road pavement contracts in California. The data come from the California Department of Transportation (Caltrans) and cover contracts procured between 1999 and 2008. The file `caltrans-bidders1999-2008.dta` is a STATA dataset containing total bids for every bidder who participated in such an auction, together with information about the bidders, the Caltrans employees who worked on the contract, and other relevant auction-specific information. You may use whatever programming software you prefer and are in no way required to use STATA. However, note that the `.dta` file has variable labels that may be useful.

Bids in these auctions are made at a *component* level—for instance, a bidder may ask for $1,000 for each ton of concrete and $200 for each meter of steel that will be used to complete the project. The variable `bidtotal` is the sum of a bidder's unit bid for each component of a contract multiplied by the quantity with which Caltrains expected the component to be used. The winner of each auction is the bidder whose `bidtotal` is lowest. The variable `engestimate` corresponds to Caltrans's engineers' estimate for the total cost of the project across components, i.e. the same as `bidtotal` but replacing unit prices by estimated unit costs.

1. Open the dataset. How many projects are there in the data?

2. We are only interested in contracts that were completed in full at the time our dataset was completed. After conferring with Caltrans, we have determined that projects initiated in 2006 or later are likely to still be ongoing in process or data collection. Furthermore, we found that a number of contracts are "improper" because only one bidder submitted a bid, the contract was awarded to someone who was not the lowest bidder, the contract had missing information, or the auction allowed bidders to also bid on "time to completion". All "improper" contracts are coded with `exclude = 1`. Excluding contracts that are too recent or improper, report the number of observations, mean, standard deviation, min value, and max value at the project level for:

- The `winning_bid`
- Winning bid normalized by engineering estimate `normbid` = `winning_bid/engestimate`
- The winning mark-up (`winning_bid` − `engestimate`)/`engestimate`
- Second-lowest bid `2ndbid`
- Money left on the table `moneyontable` = `winning_bid` − `2ndbid`
- Normalized money left on the table `normmoney` = `moneyontable/engestimate`
- Number of bidders `nbidders`

Note how many observations are dropped by each of the exclusion conditions. How would the summary table be different had we kept them? Apply the exclusion conditions above for the remainder of the work.

3. We want to exploit information about the bidders. Caltrans provided us with each bidder's `distance` to the contract site, as well as the bidder's `backlog`—the total dollar value of contracts won but not yet completed at the time of the auction. Define the following variables and produce a table with summary statistics, stating any assumptions you may make:
   - A measure of bidder `capacity` as its maximum backlog within our dataset
   - A measure of bidder utilization `util` as the ratio of `backlog` to `capacity`
   - Measure the minimum `distance` to the job site `rivaldist` among each bidder's rivals in an auction
   - Define `rivalutil` for the minimium utilization rate among each bidder's rivals
4. We want to distinguish real contenders from fringe bidders. Classify a bidder as a `fringe` bidder if the total of amount of money paid to him by Caltrans across all auctions in the sample constitutes less than 1% of the total amount of money paid for all of the contracts. Note that there are two variables that relate to this: `total_won` and `total_paid`. The difference between them is due to project adjustments made between the auction and whenever the project is completed. How does your `fringe` classification changes if we use `total_won` instead of `total_paid`? How about if we use the number of auctions won?
5. Suppose that you found that 10 fewer bidders would be classified as `fringe` when using `total_won` instead of `total_paid`. Conjecture what might be driving this. Is it a problem for interpreting the relationship between `fringe` status and strategic bidding? Why?
6. To study bidding behavior, start by estimating a linear regression model of normalized total bids on own distance and minimum rival distance. Use distance in 100-mile units. Report the R^2 and standard errors. Should standard errors be clustered? How?
7. Repeat the previous regression, now adding as additional regressors utilization, rival utilization, fringe status and the number of bidders. Interpret the results.
8. Repeat the previous regression, adding firm fixed effects. Why should we expect one of the variables to "drop out" of the regression? Repeat this, now adding project fixed effects instead of firm fixed effects. Interpret the results.

9. Part of the adjustment in payment between the time of the auction and the time of completion is due to an adjustment of the Caltrans engineers' quantity estimates. The variable `bidtotal_actual` is the analog of `bidtotal` but using the final quantity of each component that was used. Repeat the last two regressions using `bidtotal_actual`, normalized as before using the`engestimate` variable. Interpret the results.

Please report answers to parts 2 and 3 in a single table, and parts 6 through 9 in a single table.