ECON 8000/9000 Empirical Energy Economics
PS2: Replication Exercise: Regression Discontinuity
Christy Zhou

Instruction and Notes:

- Deadline: Sunday 1/25 @ 5pm
- Submission: Please compile all results into a single PDF.
- Group: Everyone can work together, but only **up to two students** can submit the assignment together. If so, make sure you include both names at the top when you submit.
- Here I write instructions using Stata. You can use other software if you like.
- This study has a replication package online. I recommend you follow my instructions below, which are slightly less complicated than the online package.

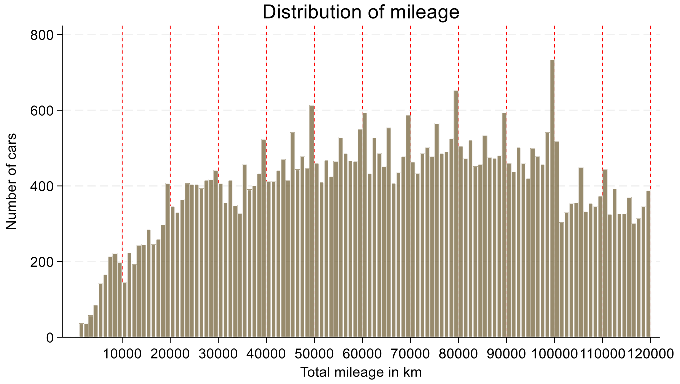
Q1. Replication of Englmaier et al. (Management Science) "Price Discontinuity in Online Market for Used Cars".

In this exercise, you will replicate Figure 1, Figure 7, and Table 3 column 1.

Q1.A Replicate Figure 1

In this exercise, we will demonstrate some key bunching that occurs every 10,000 km of mileage.
Info:

- Dataset: "PSdata_Englmaier_et_al_MgmSci.dta"
- Key vars: price: price
- Throughout all exercises, mileage means the odometer in total. You do not have to convert km to miles.



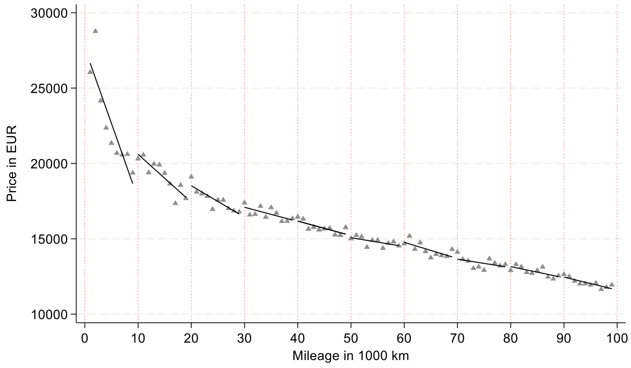
Task

- 1. Plot the above figure using a histogram

Q1.B Replicate An Alternative Version of Figure 7

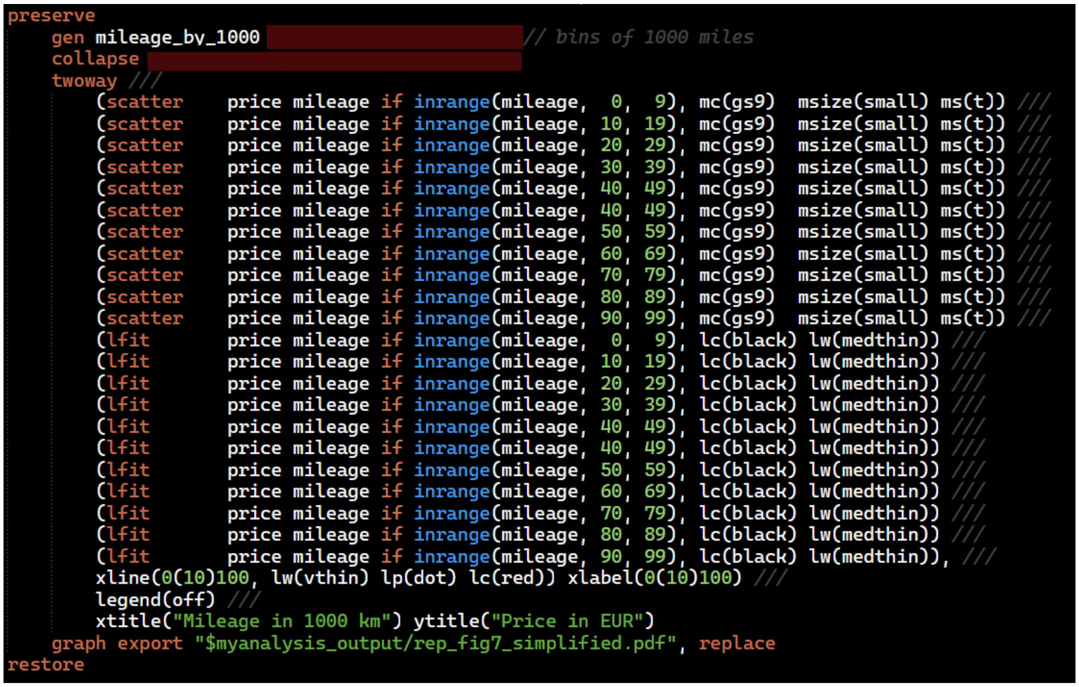
In this exercise, we will offer suggestive evidence of a price jump at each 10,000 km mileage threshold.
Info:

- Dataset: Same
- Key vars: price: price and mileage
- It is more transparent to show how prices jump, rather than how residual prices jump. (However, if you wish to plot residual price, you can follow the replication package and plot Figure 7). I recommend using the price.



Tasks:

- 1. The price is quite dispersed for each mile, meaning the vertical spread of prices for each x. Therefore, it will be a bit too cluttered to do the full scatter plot. It is more readable to see what the average price is at different mileage. To consider different mileage grids, you can create a categorical integer like this: gen mileage_by_1000 = ceil(mileage / 1000). Doing so will eventually allow us to generate an average mileage at 1000km, 2000km, 3000km, … 9000km, 10000km, 11000km, which are more disaggregated than every 10,000 km. Of course, if you want more disaggregated data points, you can define gen mileage_by_500 = ceil(mileage / 500), or however you prefer.
- 2. Compute the average price of each mileage grid calculated by collapsing the data to those grid levels.
- 3. Plot Figure 7 using the aggregated dataset. You can follow my command or write your own.



```
preserve
    gen mileage_by_1000 [redacted]// bins of 1000 miles
    collapse [redacted]
    twoway ///
        (scatter     price mileage if inrange(mileage,  0,  9), mc(gs9)  msize(small) ms(t)) ///
        (scatter     price mileage if inrange(mileage, 10, 19), mc(gs9)  msize(small) ms(t)) ///
        (scatter     price mileage if inrange(mileage, 20, 29), mc(gs9)  msize(small) ms(t)) ///
        (scatter     price mileage if inrange(mileage, 30, 39), mc(gs9)  msize(small) ms(t)) ///
        (scatter     price mileage if inrange(mileage, 40, 49), mc(gs9)  msize(small) ms(t)) ///
        (scatter     price mileage if inrange(mileage, 40, 49), mc(gs9)  msize(small) ms(t)) ///
        (scatter     price mileage if inrange(mileage, 50, 59), mc(gs9)  msize(small) ms(t)) ///
        (scatter     price mileage if inrange(mileage, 60, 69), mc(gs9)  msize(small) ms(t)) ///
        (scatter     price mileage if inrange(mileage, 70, 79), mc(gs9)  msize(small) ms(t)) ///
        (scatter     price mileage if inrange(mileage, 80, 89), mc(gs9)  msize(small) ms(t)) ///
        (scatter     price mileage if inrange(mileage, 90, 99), mc(gs9)  msize(small) ms(t)) ///
        (lfit        price mileage if inrange(mileage,  0,  9), lc(black) lw(medthin)) ///
        (lfit        price mileage if inrange(mileage, 10, 19), lc(black) lw(medthin)) ///
        (lfit        price mileage if inrange(mileage, 20, 29), lc(black) lw(medthin)) ///
        (lfit        price mileage if inrange(mileage, 30, 39), lc(black) lw(medthin)) ///
        (lfit        price mileage if inrange(mileage, 40, 49), lc(black) lw(medthin)) ///
        (lfit        price mileage if inrange(mileage, 40, 49), lc(black) lw(medthin)) ///
        (lfit        price mileage if inrange(mileage, 50, 59), lc(black) lw(medthin)) ///
        (lfit        price mileage if inrange(mileage, 60, 69), lc(black) lw(medthin)) ///
        (lfit        price mileage if inrange(mileage, 70, 79), lc(black) lw(medthin)) ///
        (lfit        price mileage if inrange(mileage, 80, 89), lc(black) lw(medthin)) ///
        (lfit        price mileage if inrange(mileage, 90, 99), lc(black) lw(medthin)), ///
        xline(0(10)100, lw(vthin) lp(dot) lc(red)) xlabel(0(10)100) ///
        legend(off) ///
        xtitle("Mileage in 1000 km") ytitle("Price in EUR")
    graph export "$myanalysis_output/rep_fig7_simplified.pdf", replace
restore
```

- 4. At which threshold do you observe the most pronounced jump in price?

Q1.C Replicate Table 3 Column 1, the Main RD Results

Info:

- Dataset: Same
- Key yvar: price
- Key controls for column 1: quadratic mileage polynomials (mileage mileage2) and seller type dummy (private)

Tasks:

- 1. Construct dummy variables that represent if a vehicle has more than 10,000km, 20,000km, …, and 100,000km. These 10 dummies will be the key variable of interest.
- 2. Regress price on the above 10 dummies (to allow for 10 discontinuities) and include the control variables listed above.
- 3. Observe and examine if the pattern described by the 10 $\hat{\beta}^m$ is more or less consistent with Figure 7.
- 4. Discuss the empirical strategy. What is/are the running variable? How are the 10 β^ms identified?
- 5. Discuss if this is really RDD. Why or why not?