ECON613: Applied Econometrics in Microeconomics

Problem Set # 1

Yueru Li

Due Date: 11 PM, Weds Jan 19

Problem 1. Basic Statistics
Open the corresponding dataset, and report the following statistics:

1. Number of households surveyed in 2007

   There are 10498 households surveyed in year 2007

2. Number of households with marital status ”Couple with kids” in 2005

   Among the households surveyed in 2005, there are 3374 households with marital status ”Couple, with Kids”.

3. Number of individuals surveyed in 2008

   There are 25510 individuals surveyed in year 2008

4. Number of individuals aged between 25 and 35 in 2016.

   Among the individuals surveyed in year 2016, there are 2765 aged between 25 and 35, inclusive of those with age exactly 25 or 35. There are 2237 individuals with age strictly between 25 and 35.

5. Cross-table gender/profession in 2009.

   Please refer to the long table starting next page.

Table 1: Profession Gender Cross Table.

Begin of Table		
Profession	Female	Male
0	11(36.67%)	19(63.33%)
11	30(34.48%)	57(65.52%)
12	8(29.63%)	19(70.37%)
13	29(27.10%)	78(72.90%)
21	63(22.83%)	213(77.17%)
22	65(36.31%)	114(63.69%)
23	8(14.29%)	48(85.71%)
31	68(40.96%)	98(59.04%)
33	85(44.27%)	107(55.73%)
34	184(56.44%)	142(43.56%)
35	50(45.87%)	59(54.13%)
37	179(40.77%)	260(59.23%)
38	78(17.49%)	368(82.51%)
42	258(70.11%)	110(29.89%)
43	437(78.88%)	117(21.12%)
44	1(33.33%)	2(66.67%)
45	153(61.69%)	95(38.31%)
46	410(54.67%)	340(45.33%)
47	82(16.05%)	429(83.95%)
48	22(9.28%)	215(90.72%)
52	782(82.23%)	169(17.77%)
53	27(12.92%)	182(87.08%)
54	584(85.63%)	98(14.37%)
55	353(77.75%)	101(22.25%)
56	694(90.39%)	74(9.61%)
62	64(12.62%)	443(87.38%)
63	35(6.31%)	520(93.69%)
64	29(10.55%)	246(89.45%)
65	19(10.67%)	159(89.33%)
67	147(38.28%)	237(61.72%)
68	120(40.40%)	177(59.60%)
69	40(32.79%)	82(67.21%)

Continuation of Table 10		
Profession	Female	Male
NA	8167(54.03%)	6949(45.97%)
End of Table		

6. Distribution of wages in 2005 and 2019. Report the mean, the standard deviation, the inter-decile ratio D9/D1 and the Gini coefficient.

   I performed the following actions before calculating the summary statistics:

   (a) I focused on the the individuals with employment status ”Employed” and ”unemployed”, which correspond to the segment of the population that are actively in the labor force.

   (b) I remove the NA’s for both employed and unemployed as the proprtion is very small (1.2% for 2005 and 0.1% for 2019).

   (c) I remove the 0 earners for the individuals who are employed because the proportion is small (9.7% for 2005 and 6.5% for 2019) and it does not make much sense for a person to be employed but make 0 in wages.

   (d) I keep the individuals making positive wages with employment status ”unemployment”. This might reflect some non-employment income. Individuals making positive wages account for a significant portion of the unemployed group. (49.7% in 2005 and 61.0% in 2019)

   I report the summary statistics for each year.

   (a) Data in 2005

Variable	Statistics	Value
wage	Min/Max	0/271962.0
wage	Med [IQR]	18283.0 [8260.0;26391.0]
wage	Mean (std)	20194.1 (18593.1)

   The D9/D1 ratio is 13.32 and the Gini coefficient is 0.39.

   (b) Data in 2019

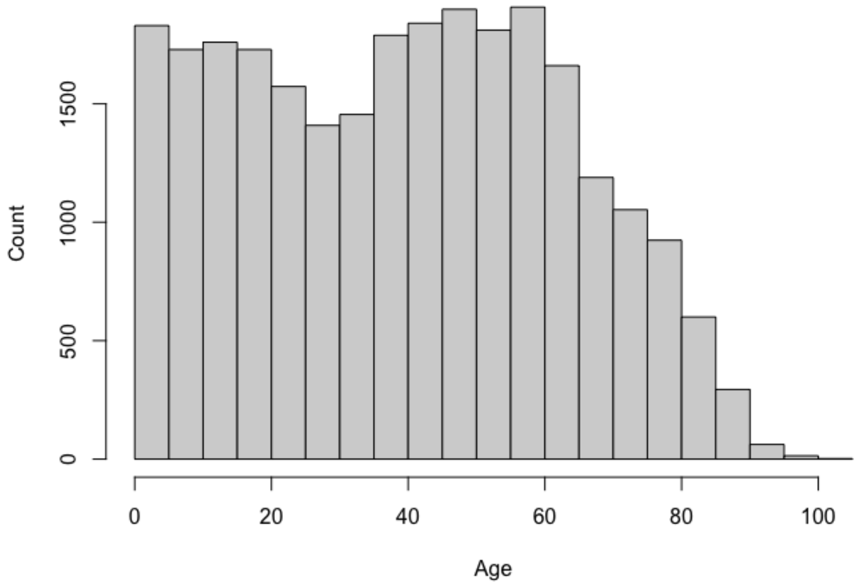

Figure 1: 2010 Sample Age Histogram

Variable	Statistics	Value
wage	Min/Max	0/1068556.0
wage	Med [IQR]	24719.0 [12261.0;35032.0]
wage	Mean (std)	26981.7 (25518.7)

The D9/D1 ratio is 11.27 and the Gini coefficient is 0.38.

7. Distribution of age in 2010. Plot an histogram. Is there any difference between men and women?

   (a) Summary statistics of the distribution of age

Variable	Statistics	Sample	Female	Male
age	Min/Max	0 / 102.0	0 / 102.0	0 / 96.0
age	Med [IQR]	40.0 [19.0;58.0]	42.0 [20.0;59.0]	39.0 [19.0;57.0]
age	Mean (std)	39.9 (23.4)	40.8 (23.6)	38.9 (23.2)

   (b) Histograms: Please refer to figure 1 for the histogram of the entire 2010 sample and to figure 2 for a histogram by gender for the sample in 2010.

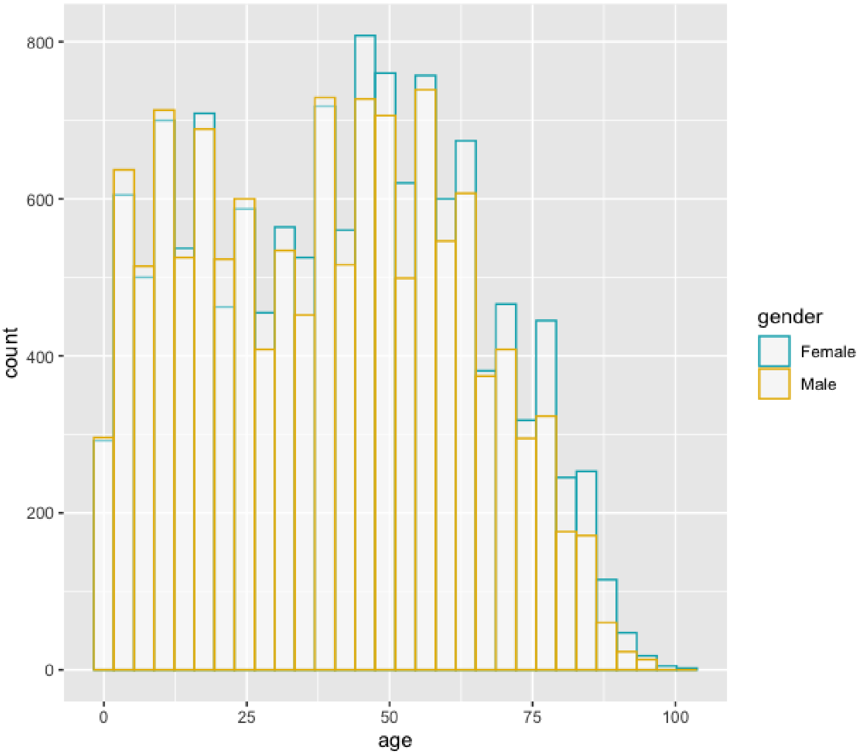

Figure 2: 2010 Sample Age Histogram by gender

(c) Comparison: It seems that the distribution of women's age has a longer and thicker tail on the right (high age section) compared with the male age distribution.

8. Number of individuals in Paris in 2011.

   There are 3514 individuals.

Problem 2. Merge Datasets

1. In the first part of this exercise, we will learn how to merge datasets.

   (a) Read all individual datasets from 2004 to 2019. Append all these datasets.

   Please refer to the *R* code.

   (b) Read all household datasets from 2004 to 2019. Append all these datasets.

   Please refer to the *R* code.

(c) List the variables that are simultaneously present in the individual and household datasets.

The variables that simultaneously present in both datasets are ”idemen” and ”year”.

(d) Merge the appended individual and household datasets.

There seem to be 32 repeated ”idind” in the data for year 2013. Upon close inspection, they seem to be related to children who are double registered in families that went through a divorce that year. I decide to drop the families that are associated with this repeated ”idinds”.

In total, I delete 599 data points from the dataset. Compared to the total size of 413472, this should be a negligible manipulation.

2. In the second part, we use the newly created dataset from the previous to answer the following questions:

(a) Number of households in which there are more than four family members

Here I report the number of households satisfying the requirement for each year.

Table 2: Number of Households with More Than 4 Individuals Each Year

Begin of Table	
Year	Number of Households
2004	745
2005	814
2006	862
2007	874
2008	814
2009	809
2010	820
2011	781
2012	813
2013	748
2014	781
2015	760
2016	749
2017	702

Continuation of Table 10	
Year	Number of Households
2018	647
2019	692
End of Table	

(b) Number of households in which at least one member is unemployed

Number of households in which at least one member is unemployed for each year.

Table 3: Number of Households In Which At Least One Member is Unemployed Each Year

Begin of Table	
Year	Number of Households
2004	950
2005	1039
2006	1030
2007	975
2008	909
2009	1045
2010	1108
2011	1070
2012	1202
2013	1173
2014	1182
2015	1225
2016	1136
2017	1103
2018	991
2019	1086
End of Table	

(c) Number of households in which at least two members are of the same profession

There are many people with blank profession variable, I drop them before computing this number because I think I should only look at individuals who are employed and report a meaningful profession. Similar to the two previous problems, I report the number for each year.

Table 4: Number of Households In Which At Least Two Members Are Of The Same Profession Each Year

Begin of Table	
Year	Number of Households
2004	445
2005	496
2006	480
2007	490
2008	458
2009	452
2010	474
2011	491
2012	517
2013	455
2014	475
2015	467
2016	473
2017	457
2018	454
2019	496
End of Table	

(d) Number of individuals in the panel that are from household Couple with kids I report this number for each year.

Table 5: Number Of Individuals In The Panel That Are From Household ”Couple with kids” Each Year

Begin of Table	
Year	Number of Individuals
2004	11993
2005	13210
2006	13626

Continuation of Table 10

Year	Number of Inviduals
2007	13949
2008	13459
2009	13258
2010	13689
2011	13759
2012	14362
2013	13071
2014	13220
2015	12995
2016	12941
2017	11960
2018	11442
2019	12149

End of Table

(e) Number of individuals in the panel that are from Paris.

I report this number for each year.

Table 6: Number Of Individuals In The Panel That Are From Paris Each Year

Begin of Table

Year	Number of Individuals
2004	3494
2005	3734
2006	3658
2007	3735
2008	3559
2009	3524
2010	3607
2011	3514
2012	3679
2013	2288
2014	2576

Continuation of Table 10	
Year	Number of Inviduals
2015	3033
2016	2946
2017	2836
2018	2797
2019	2924
End of Table	

(f) Find the household with the most number of family members. Report its "idmen".

There are two such families. One is the family with idmen 2207811124040100 in year 2007 and another is the family with idmen 2510263102990100 in year 2010

(g) Number of households present in 2010 and 2011.

i. Number of household that were present in either 2010 or 2011 There are 13401 such households.

ii. Number of household that were present in both 2010 or 2011
There are 8968 such households

iii. Number of household that were present in each year
There are 11034 households in year 2010 and 11335 households in year 2011.

Problem 3. Migration

1. Find out the year each household enters and exit the panel. Report the distribution of the time spent in the survey for each household.

   For the entry and exit for each household, please refer to the R-code. I report the distribution of time spent in the survey for households here:

Variable	Statistics	Years
Years In the Survey	Min/Max	1.0 / 9.0
Years In the Survey	Med [IQR]	6.0 [4.0;8.0]
Years In the Survey	Mean (std)	5.9 (2.4)

2. Based on "datent", identify whether or not a household moved into its current dwelling at the year of survey. Report the first 10 rows of your result and plot the share of individuals in that situation across years.

I report first 10 rows in my testing on whether or not a household moved into it's current dwelling at the year of survey in the following table:

Table 7: First 10 Households in the data and whether or not they moved in the year of Survey based on datent

Begin of Table		
Household idmen	Survey Year	Moved at the Year of Survey
1200010012930100	2004	FALSE
1200010040580100	2004	FALSE
1200010040580100	2005	FALSE
1200010066630100	2004	FALSE
1200010066630100	2005	TRUE
1200010082450100	2004	FALSE
1200010082450100	2005	FALSE
1200010086440100	2004	FALSE
1200010086440100	2005	FALSE
1200010102990100	2004	FALSE
End of Table		

Share of individuals in the situation is plotted in figure 3.

3. Based on "myear" and "move", identify whether or not household migrated at the year of survey. Report the first 10 rows of your result and plot the share of individuals in that situation across years.

   I use the variable myear to fill in the NA's for the variable move, and then use move to decide whether or not a household moved in the year of survey.

   I report the first ten rows of my results for different household-year combinations.

Table 8: First 10 Households in the data and whether or not they moved in the year of Survey based on move and myear

Begin of Table		
Household idmen	Survey Year	Moved at the Year of Survey
1200010012930100	2004	FALSE
1200010040580100	2004	FALSE

Continuation of Table 10

Household idmen	Survey Year	Moved at the Year of Survey
1200010040580100	2005	FALSE
1200010066630100	2004	FALSE
1200010066630100	2005	TRUE
1200010082450100	2004	FALSE
1200010082450100	2005	FALSE
1200010086440100	2004	FALSE
1200010086440100	2005	FALSE
1200010102990100	2004	FALSE

End of Table

We can see that this table is identical to the previous one. For a plot of the share of the individuals in this situation using this method, pleas refer to figure 4

4. Mix the two plots you created above in one graph, clearly label the graph. Do you prefer one method over the other? Justify. Please refer to figure 5 for the two plots mixed together. I prefer the first method (using datent) for the following reasons:

    (a) Reason 1: Less NA's in the data. Notice that the two lines match each other relatively well before year 2015. This is because there are many NA's in the value for move, more specifically, in total, there are 248 NA's in the datent variable but there are 31680 for move after we use myear to fillin the missing values for move. The large amount of NA's concentrated in surveys after year 2015 is likely the cause of the divergence of the two curves after 2015.

    (b) Reason 2: Less manipulation. Using datent saves us the effort of filling in move with myear.

5. For house holds who migrate, find out how many households had at least one family member changed his/her profession or employment status.

    For the limitation of the dataset, while we can identify the households who has migrated the year of the survey, we can only identify a career change or an employment change if that person also appeared in the survey of previous year. It is important that we keep this in mind while we interpret the data. It is also important to note that we can only report data starting in 2005 because of the need of at least two year's record of profession. In the following table, I count the number of families with at least one

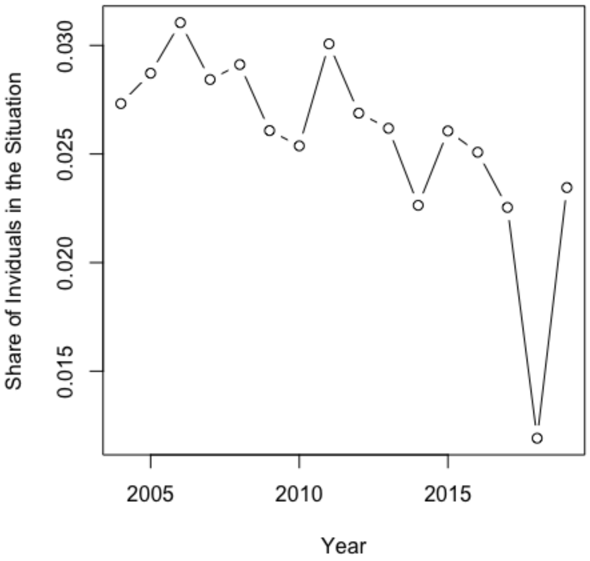

Figure 3: Share of Individuals moved in the Year of Survey

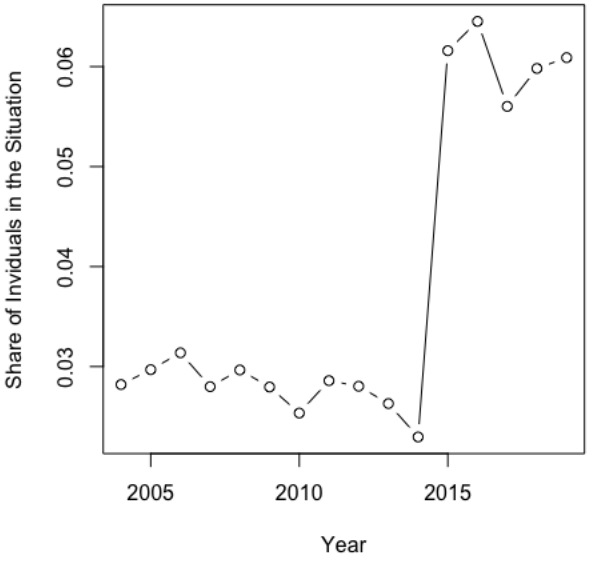

Figure 4: Share of individuals in the situation using myear and move

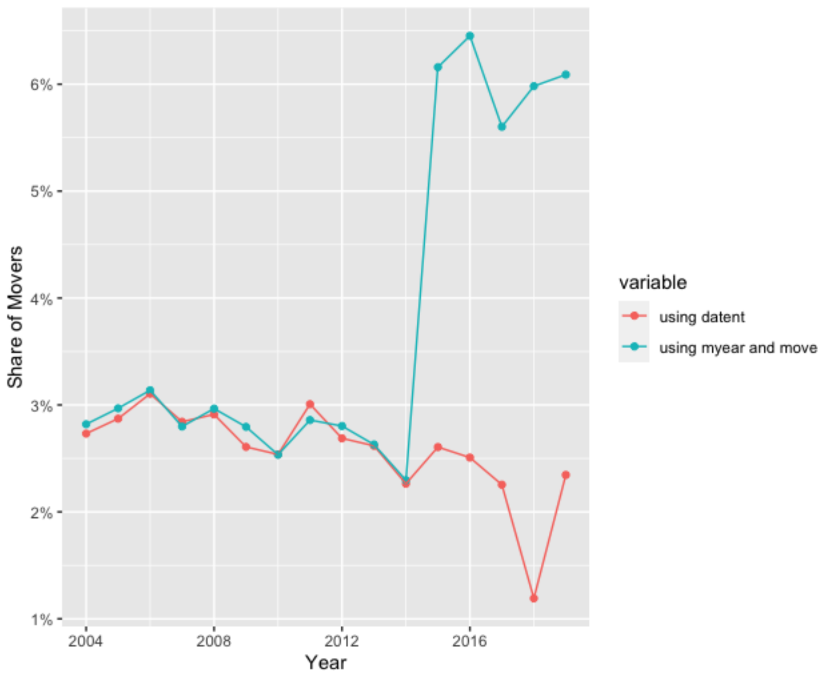

Figure 5: Combining the plot for the previous two parts

member who changed his/her profession or employment status in the year of migration. My criterion for migration is determined using datent as it has less NA values.

Table 9: Number of households that migrate and also have at least one member having a career or employment status change in the year of migration.

Begin of Table			
Year	Households Migrated	Also surveyed in the previous year	satisfying the condition
2005	311	187	63
2006	343	211	76
2007	309	186	64
2008	326	207	86
2009	307	181	86
2010	306	187	57
2011	363	234	84
2012	344	216	92
2013	305	202	77
2014	282	193	62
2015	299	208	73

Continuation of Table 10			
Year	Households Migrated	Also surveyed in the previous year	satisfying the condition
2016	288	191	63
2017	274	173	53
2018	130	92	31
2019	279	164	54
End of Table			

Problem 4. Attrition

Compute the attrition across each year, where attrition is defined as the reduction in the number of indi- viduals staying in the data panel. Report your final result as a table in proportions.

Table 10: Attrition Rate Each Year.

Begin of Table			
Year	Households	Stayer Households	Attrition
2004	22144	NA	$\frac{22144-19148}{22144} = 13.54\%$
2005	24234	19148	$\frac{24234-19384}{24234} = 20.10\%$
2006	24929	19384	$\frac{24929-20472}{24929} = 17.88\%$
2007	25890	20472	$\frac{25890-20017}{25890} = 22.68\%$
2008	25482	20017	$\frac{25482-20237}{25482} = 20.58\%$
2009	25577	20237	$\frac{25577-20870}{25577} = 18.40\%$
2010	26483	20870	$\frac{26483-21348}{26483} = 19.39\%$
2011	27001	21348	$\frac{27001-22405}{27001} = 17.02\%$
2012	28458	22405	$\frac{28458-21192}{28458} = 25.53\%$
2013	26238	21192	$\frac{26238-20470}{26238} = 21.98\%$
2014	26725	20470	$\frac{26725-20862}{26725} = 21.94\%$
2015	26590	20862	$\frac{26590-20808}{26590} = 21.74\%$
2016	26594	20808	$\frac{26594-19939}{26594} = 25.02\%$
2017	25376	19939	$\frac{25376-19181}{25376} = 24.41\%$
2018	24677	19181	$\frac{24677-18679}{24677} = 24.30\%$
2019	26475	18679	NA
End of Table			