Problem Set 2 Solutions

Note: Text that is preceded by a "." is the Stata code used in the analysis. Text enclosed in "*"s explains what each piece of code is doing. Where relevant, I have pasted the actual Stata output.

Part I

Romney votes by Texas County

The election results are available at http://elections.sos.state.tx.us/elchist.exe. After selecting 2012 general election from the dropdown menu, choose county-by-county canvass results and president/vice president. From here, copy and paste into excel or a text editor and save as a CSV file.

```
*Using semicolon as delimiter*
#delimit;

. log using ps2.log

. set more off

. cd "/Users/nlmiller/Desktop/Poli Sci Lab/PS2/"

. insheet using "/Users/nlmiller/Desktop/Poli Sci Lab/PS2/texas_counties.csv"

*converting Romney vote variable from string to numeric, dropping commas*

. destring rep, replace ignore(,)

*converting total vote variable from string to numeric, dropping commas*

. destring votes, replace ignore(,)

*Creating Romney percentage vote variable*

. gen pctromney=rep/votes

*Getting rid of observation that sums up all counties*

. drop if county=="ALL COUNTIES"

. save "/Users/nlmiller/Desktop/Poli Sci Lab/PS2/Romney_Texas.dta"

*Finding mean, standard deviation, skewness, and kurtosis of Romney vote*

. summarize pctromney, detail
```

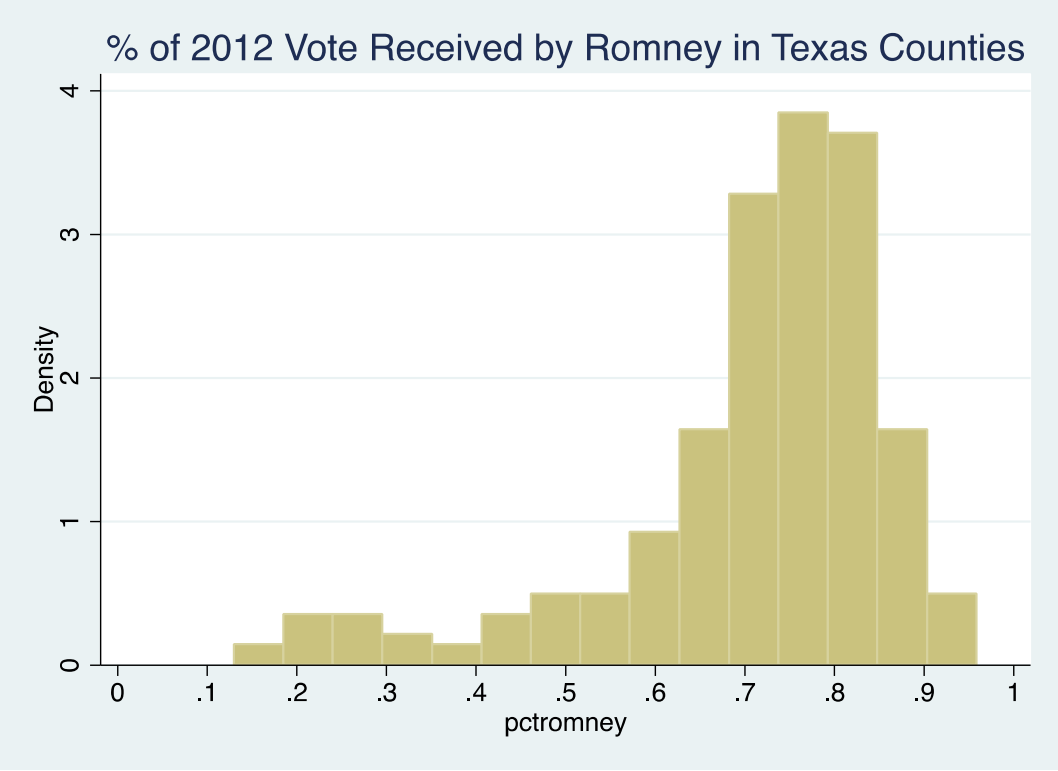
```
                          pctromney
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .2055287       .130186
 5%     .3314734       .1571311
10%     .4930846       .2055287       Obs                 254
25%     .6648976       .2109863       Sum of Wgt.         254

50%     .7458333                      Mean            .708672
                       Largest        Std. Dev.       .1586351
75%     .8119143       .911315
90%     .8568618       .9212598       Variance        .0251651
95%     .8872077       .9291498       Skewness       -1.563087
99%     .9212598       .9586207       Kurtosis        5.365688
```

.

Creating a historgram of Romney vote, customizing x axis

. hist pctromney, title(% of 2012 Vote Received by Romney in Texas Counties) xlabel(0 (0.1) 1)



OECD Growth Rates 2011

The data is available from the OECD statistics website (http://stats.oecd.org/index.aspx?queryid=26646). The data can be exported as a CSV or copy and pasted into excel/a text editor and saved as a CSV (being careful to only include the countries actually part of the OECD, not China, India, etc). Stata will not accept numerical

variable names so the names of the yearly growth rates should be changed (for example, to growth2010, growth2011, etc,)

. clear

. insheet using "/Users/nlmiller/Desktop/Poli Sci Lab/PS2/OECD_2011.csv"

. save "/Users/nlmiller/Desktop/Poli Sci Lab/PS2/OECD_Growth_2011.dta"

Finding mean, standard deviation, skewness, and kurtosis of growth rates

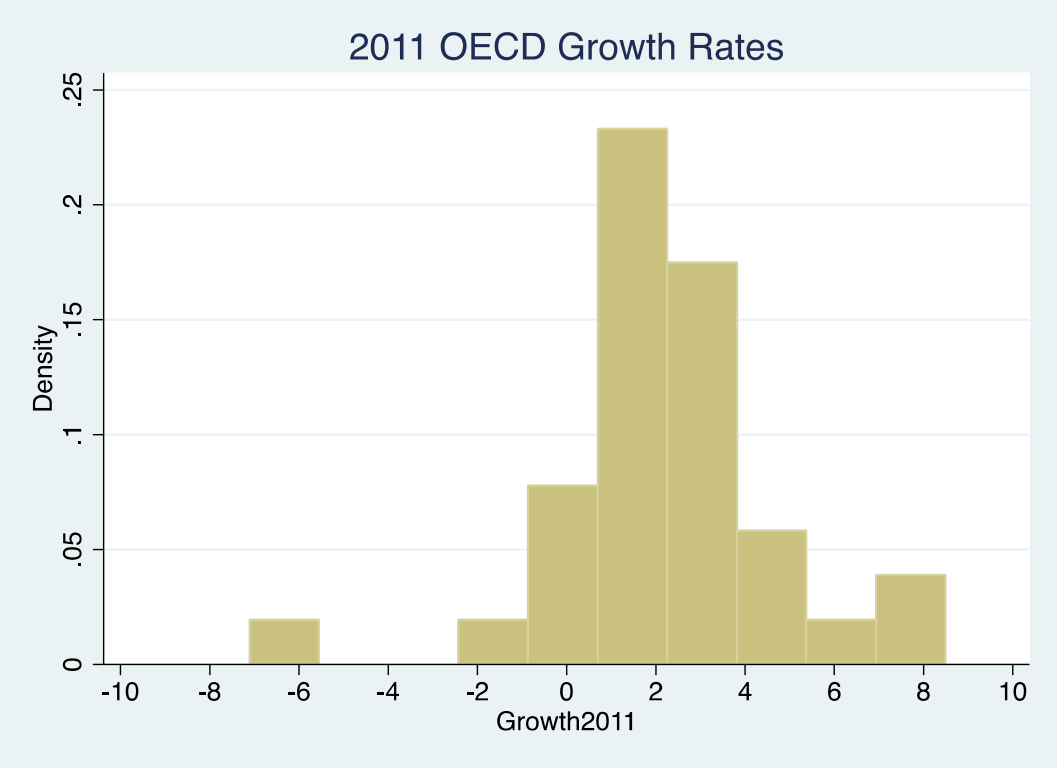
. summarize growth2011, detail

```
                        Growth2011
-------------------------------------------------------------
      Percentiles      Smallest
 1%          -7.1           -7.1
 5%          -1.6           -1.6
10%            .4             -.6       Obs                  33
25%           1.1              .4       Sum of Wgt.          33

50%           1.9                       Mean           2.169697
                      Largest            Std. Dev.      2.719431
75%           3.2            4.6
90%           4.6              6       Variance       7.395303
95%           8.3            8.3       Skewness       -.498337
99%           8.5            8.5       Kurtosis        6.47441
```

Creating a histogram of growth rates, using 10 bins and customizing x axis

. hist growth2011, bin(10) title(2011 OECD Growth Rates) xlabel(-10 (2) 10)



Military expenditure as % of GDP, 2010

The data on military expenditure as percent of GDP is available on the World Bank website: **http://data.worldbank.org/indicator/MS.MIL.XPND.GD.ZS?order=wbapi_data_value_2010+wbapi_data_value&sort=desc**

After downloading the data excel file or copying and pasting into excel/text editor, delete extraneous regional categories/years and save as a CSV file. Rename 2010 variable (to milgdp2010, for example) so that Stata reads it as a variable name

. clear

. insheet using "/Users/nlmiller/Desktop/Poli Sci Lab/PS2/mil_spending_2010.csv"

. save "/Users/nlmiller/Desktop/Poli Sci Lab/PS2/mil_spending_2010.dta"

Finding mean, standard deviation, skewness, and kurtosis of military spending as % of GDP

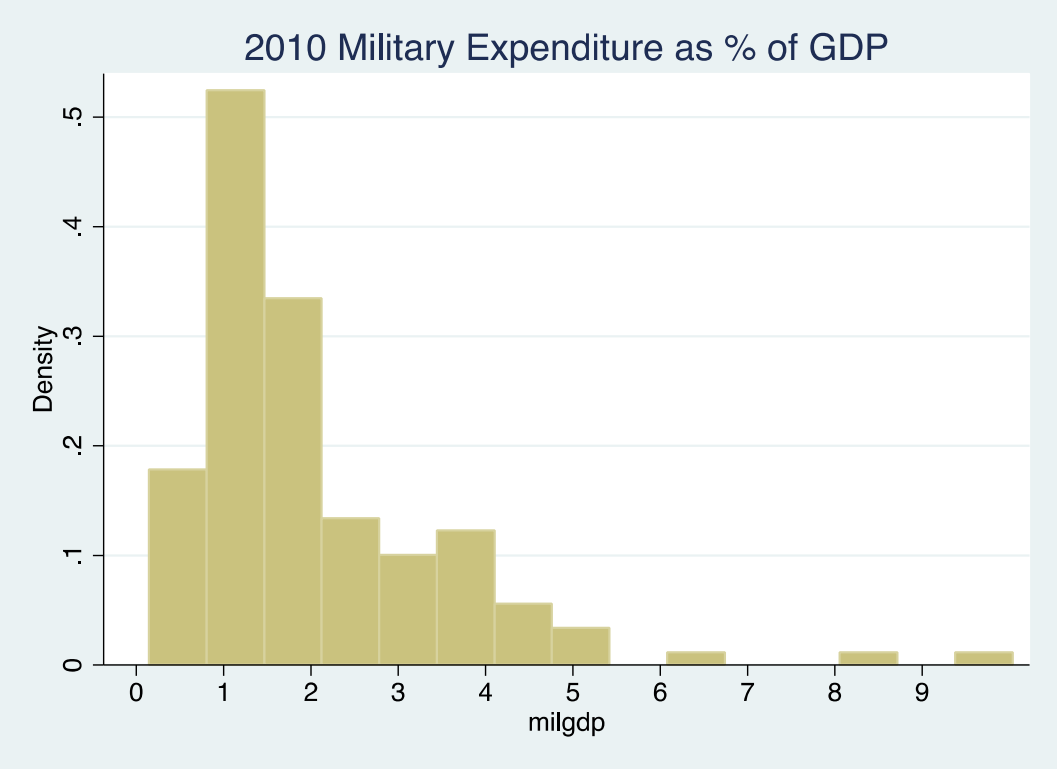
. summarize milgdp, detail

```
                        milgdp2010
-------------------------------------------------------------
      Percentiles      Smallest
 1%      .225932       .1484693
 5%     .4995266        .225932
10%     .7194218       .3157822       Obs                 136
25%     1.100418       .3887702       Sum of Wgt.         136

50%     1.567121                      Mean           2.038118
                       Largest        Std. Dev.      1.494571
75%     2.682297       5.396235
90%     3.904863       6.549769       Variance       2.233743
95%     4.415103       8.461089       Skewness       2.198611
99%     8.461089       10.03668       Kurtosis       10.23931
```

Creating a histogram of growth rates of military spending as a % of GDP, using 15 bins and marking every 1 percentage point on the x axis

. hist milgdp, bin(15) xlabel(0 (1) 9) title(2010 Military Expenditure as % of GDP)



Part II

. clear

. use "/Users/nlmiller/Desktop/Poli Sci Lab/PS2/cces08_common_output.dta"

See from CCES codebook that 1=democrat and 2=republican and generate variable that equals 1 if individual is a democrat and 0 if they are a republican

. gen democrat=1 if cc307==1

. replace democrat=0 if cc307==2

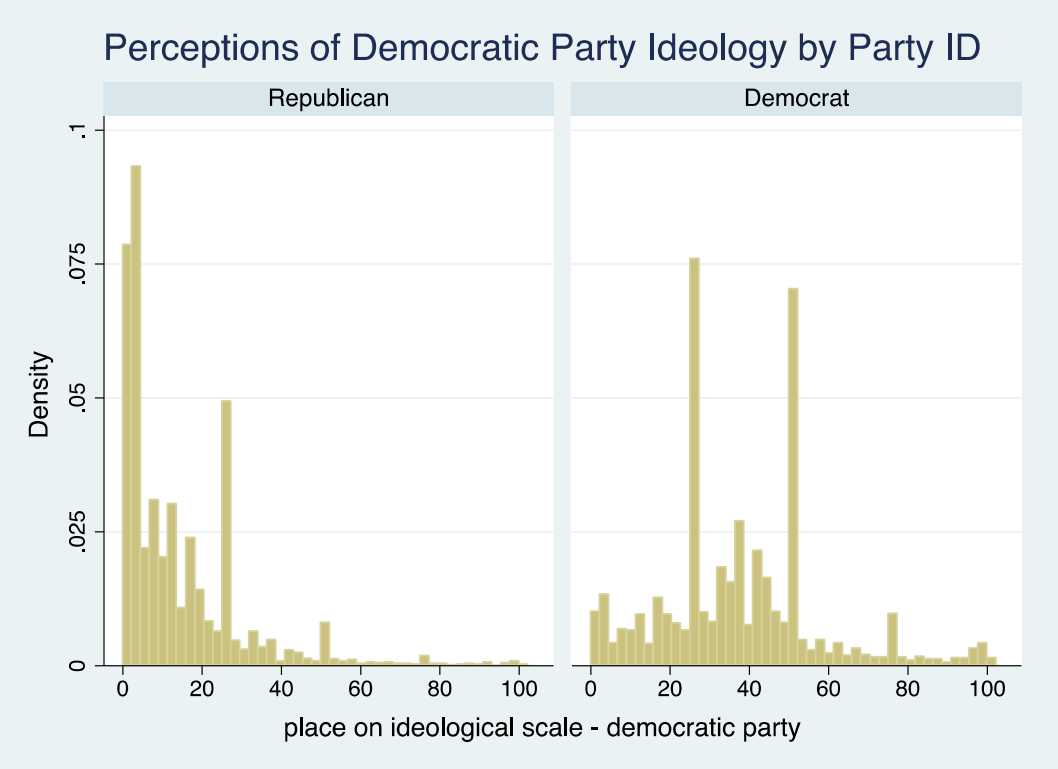
Label the variable to note that 1 denotes Democrat and 0 denotes Republican

. label define democrat 1 "Democrat" 0"Republican

. label values democrat democrat

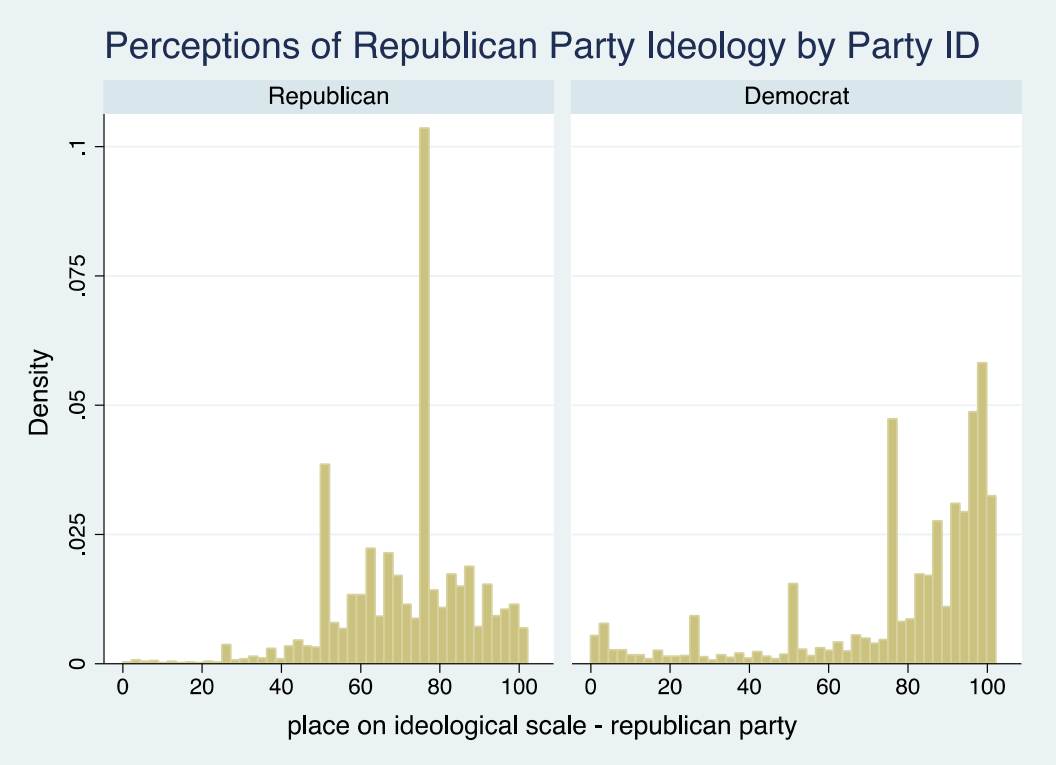
Generate histogram comparing how Republicans and Democrats view the ideology of the Democratic Party, including title, customizing x and y scales and ticks, and removing extraneous legends and notes

. histogram cc317b, ylabel(0(.025).1) xlabel(0(20)105, valuelabel) by(, title(Perceptions of Democratic Party Ideology by Party ID) note("")) by(, legend(off)) by(democrat) width (2)



*Same as above, but comparing how Republicans and Democrats view the ideology of the Republican Party *

. histogram cc317c, ylabel(0(.025).1) xlabel(0(20)105, valuelabel) by(, title(Perceptions of Republican Party Ideology by Party ID) note("")) by(, legend(off)) by(democrat)



Identify univariate statistics on how Democrats view their own party as compared to how they view the Republican Party, using analytical weights

. tabstat cc317b cc317c [aweight=v200], by(democrat) statistics (mean sd skewness kurtosis)

```
Summary statistics: mean, sd, skewness, kurtosis
  by categories of: democrat

democrat |    cc317b    cc317c
---------+--------------------
       0 |  15.30989  71.30035
         |  16.95343  16.92204
         |  2.022479 -.7400853
         |  8.115932  3.972089
---------+--------------------
       1 |  38.31179  76.33653
         |  21.85211  27.01542
         |  .6626436 -1.437213
         |  3.400378  4.052581
---------+--------------------
   Total |  28.61975  74.15994
         |  22.94349  23.33164
         |  .9012175  -1.25966
         |  3.493013    4.2583
-----------------------------
```

The graphs and univariate statistics demonstrate major differences in how the two parties perceive both their own party and the opposing party. Both parties view the opposing party as more ideologically extreme than party members view their own party, although this is more pronounced for Democrats: Democrats give the Republican Party a mean score of 76 while Republicans give themselves 71; Republicans give the Democratic Party a mean score of 15 while Democrats give themselves 38. These figures also suggest that Democrats perceive their party to be more somewhat moderate than Republicans perceive their own party (mean scores of 38 and 71, respectively).

With respect to the spread of the data, the standard deviations reveal that there is more variance in how Democrats perceive the ideological positions of the parties (standard deviations of 21 and 27); Republicans are comparatively more homogenous in their views (both standard deviations around 17). The skewness statistics and histograms show that there is a rightward skew in how both parties perceive Democrats' ideology and a leftward skew in how both parties perceive Republicans' ideology. Thus, there is a greater mass of individuals who view Democrats as liberal and Republicans as conservative but there are longer tails of individuals who view Democrats as relatively conservative and Republicans as relatively liberal. Finally, the kurtosis for all four distributions are quite high (especially with respect to how Republicans view the Democratic Party), suggesting a relatively high concentration of observations at particular values (all more than in a normal distribution, which has a kurtosis of 3).

Part III

```
. clear

. insheet using "/Users/nlmiller/Desktop/Poli Sci Lab/PS2/NMC_v4_0.csv"

. sum milper, detail

*Replacing missing value code (-9) with "." to avoid miscalculated statistics*
. replace milper=. if milper==-9

*Identifying mean, standard deviation, and number of observations for milper in 1900, 1915, 1945, 1980, and 2005*

. sum milper if year==1900

. sum milper if year==1915

. sum milper if year==1945

. sum milper if year==1980

. sum milper if year==2005
```

Year	Mean	Standard Deviation	Minimum	Maximum	N
1900	137.71	264.14	0	1142	42
1915	673.39	1485.13	1	5500	44
1945	1146.98	2862.37	0	12500	45
1980	171.29	527.79	0	4650	156
2005	106.55	260.84	0	2255	186

The changes in mean military size in these years is largely a function of whether a major international war is ongoing, with averages peaking during WWII (1945), and WWI (1915). The mean is also noticeably higher during the Cold War (1980) than in 1900 or 2005, the latter two of which were periods of relative peace globally. The standard deviation follows a similar pattern, with greater average deviations from the mean (both in absolute and relative terms) during periods of world war or international crisis, presumably because there is greater variability in these years in how militarily mobilized nations are (those involved in the war/crisis will have ramped up military size, while those uninvolved will have stayed basically the same, increasing overall variability).

The key factor driving the increase in the number of observations is simply that decolonization post-1945 led to a rapid increase in the number of independent countries (see the n for 1980) and the collapse of the Soviet Union and Yugoslavia led to a further proliferation of countries after 1990 (see the n for 2005).

Part IV

Access licensed driver statistics at http://www.fhwa.dot.gov/policyinformation/statistics/2011/ and population statistics at http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=PEP_2011_PEPAGESEX&prodType=table

. clear

Load data after cleaning spreadsheet to drop extraneous variables and renaming variables as strings: drivers18, drivers19, etc.

. insheet using “/Users/nlmiller/Desktop/Poli Sci Lab/PS2/Drivers_by_state.csv”

Generate stateid variable to facilitate merging with population data and paste geoid2 values from population data

. gen stateid=.

. save "/Users/nlmiller/Desktop/Poli Sci Lab/PS2/Drivers_by_state.dta", replace

. clear

. insheet using "/Users/nlmiller/Desktop/Poli Sci Lab/PS2/PEP_2011_PEPAGESEX/PEP_2011_PEPAGESEX_with_ann.csv"

Rename geoid2 variable to make merging possible
. rename geoid2 stateid

. save "/Users/nlmiller/Desktop/Poli Sci Lab/PS2/population_age_by_state.dta"

. clear

. use "/Users/nlmiller/Desktop/Poli Sci Lab/PS2/Drivers_by_state.dta"

Merge population data using stateid identifier

. merge 1:1 stateid using "/Users/nlmiller/Desktop/Poli Sci Lab/PS2/population_age_by_state.dta"

Drop all population estimates except those from 2010 census that include both sexes

. drop est*

. drop cen42010sex1*

. drop cen42010sex2*

create age 18-19 population and licensed drivers totals
. gen cen42010sex0_age18to19= cen42010sex0_age18to24- cen42010sex0_age20to24

. replace drivers18=drivers18+drivers19

. drop drivers19

. save "/Users/nlmiller/Desktop/Poli Sci Lab/PS2/drivers_and_population_by_age.dta", replace

Collapse dataset to record sums of licensed drivers and population in each age group, nationwide

. collapse (sum) drivers18 drivers20 drivers25 drivers30 drivers35 drivers40 drivers45 drivers50 drivers55 drivers60 drivers65 drivers70 drivers75 drivers80 drivers85 cen42010sex0_age18to19 cen42010sex0_age20to24 cen42010sex0_age25to29 cen42010sex0_age30to34 cen42010sex0_age35to39 cen42010sex0_age40to44 cen42010sex0_age45to49 cen42010sex0_age50to54 cen42010sex0_age55to59 cen42010sex0_age60to64 cen42010sex0_age65to69 cen42010sex0_age70to74 cen42010sex0_age75to79 cen42010sex0_age80to84 cen42010sex0_age85plus

Generate id variable and rename population variables to facilitate reshaping. To reshape long, a variable stem is needed with a numerical suffix to denote the values of the observations

. gen id=1

. rename cen42010sex0_age18to19 pop18
. rename cen42010sex0_age20to24 pop20
. rename cen42010sex0_age25to29 pop25
. rename cen42010sex0_age30to34 pop30
. rename cen42010sex0_age35to39 pop35
. rename cen42010sex0_age40to44 pop40
. rename cen42010sex0_age45to49 pop45
. rename cen42010sex0_age50to54 pop50
. rename cen42010sex0_age55to59 pop55
. rename cen42010sex0_age60to64 pop60
. rename cen42010sex0_age65to69 pop65
. rename cen42010sex0_age70to74 pop70
. rename cen42010sex0_age75to79 pop75
. rename cen42010sex0_age80to84 pop80
. rename cen42010sex0_age85plus pop85

Use reshape command to turn each age group into separate rows in a dataset, recording the total population and number of licensed drivers

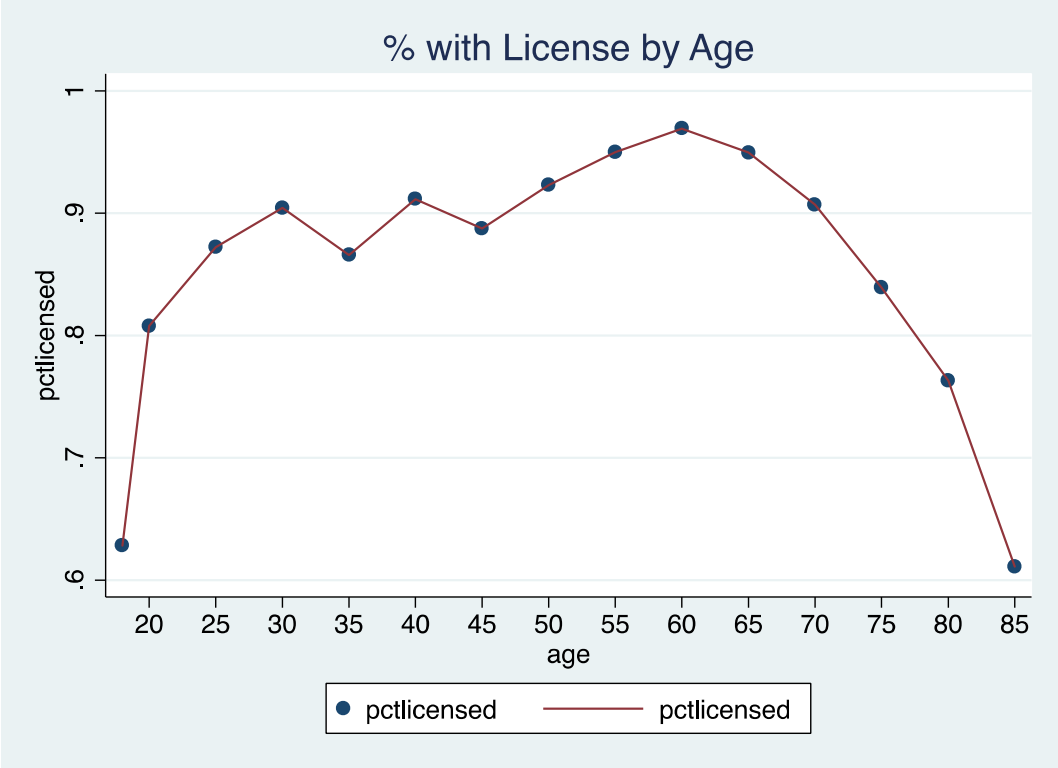
. reshape long drivers pop, i(id) j(age)

Generate a variable that records the percentage of the age group with a license

. gen pctlicensed=drivers/pop

Produce a graph that shows how the rate of license possession changes as a function of age

. graph twoway (scatter pctlicensed age) (line pctlicensed age), xlabel (20 (5) 85) title(% with License by Age)



Save reshaped dataset

. save "/Users/nlmiller/Desktop/Poli Sci Lab/PS2/National_drivers_by_age.dta"

Load state-level dataset

. clear

. use "/Users/nlmiller/Desktop/Poli Sci Lab/PS2/drivers_and_population_by_age.dta"

Drop extraneous census variables

. drop cen42010sex0_age999 cen42010sex0_age0to4 cen42010sex0_age5to9 cen42010sex0_age10to14 cen42010sex0_age15to19 cen42010sex0_age0to17 cen42010sex0_age0to4r cen42010sex0_age5to13 cen42010sex0_age14to17

cen42010sex0_age18to64 cen42010sex0_age18to24 cen42010sex0_age25to44 cen42010sex0_age45to64 cen42010sex0_age65plus cen42010sex0_age85plusr cen42010sex0_age16plus cen42010sex0_age18plus cen42010sex0_age15to44 cen42010sex0_medage

Drop Puerto Rico, which doesn't have drivers data

. drop if stateid==72

Generate a variable that computes the overall rate of licensed drivers in all age groups

. egen totaldrivers=rowtotal(drivers18-drivers85)

. egen totalpop=rowtotal(cen42010sex0_age20to24- cen42010sex0_age18to19)

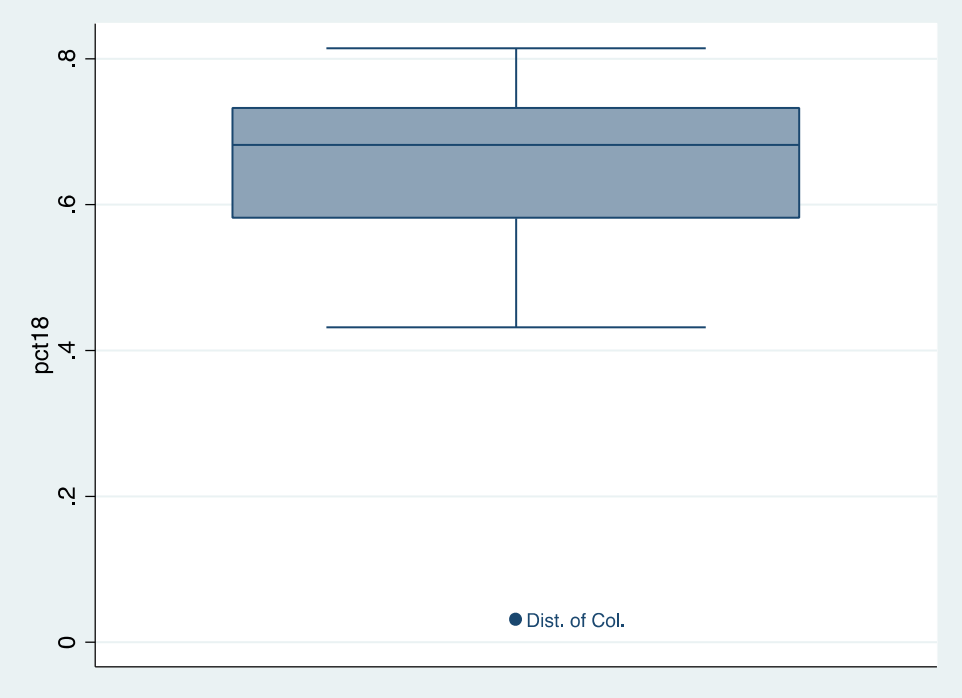
. gen totalrate=totaldrivers/totalpop

Generate box plot to summarize distribution, identifying and labeling outliers

. graph box totalrate, mark(1, mlabel(state))

Generate the same graph, but only from the 18-19 age group

. graph box pct18, mark(1, mlabel(state))



. log close

There are some clear outliers, both for the overall population and the 18-19 age bracket. With respect to the overall population, Indiana is a major outlier (with a license possession rate near 140%), and indeed several states appear to have license possession above 100%, which suggests something strange is going on. This could be because non-residents are allowed to have licenses in certain states, because individuals hold licenses in multiple states simultaneously, or because of poor record keeping. Only if it is due to poor record keeping should it lead us to worry about the general quality of the data. New York is an outlier as well, with a relatively low license possession rate (less than 80%). This could be because such a large portion of New York's population is urban, where driving is less necessary.

In the 18-19 age bracket, the major outlier is the District of Columbia, with a license possession rate around zero. This is presumably due to a law whereby those under the age of 20 are barred from holding a license, and/or because DC is a city, and thus teenagers have less need or opportunity to get a license.