Problem Set 3 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

. clear

. delimit;

. set more off

. log using ps3.log

. use "/Users/nlmiller/Desktop/Poli Sci Lab/PS3/cces08_common_output.dta"

Examine variables and coding schemes

. tab cc317b, m

. tab cc317c, m

Regress view of the Democratic Party on ideological self-placement, using analytical weights

. reg cc317b cc317a [aweight=v200]

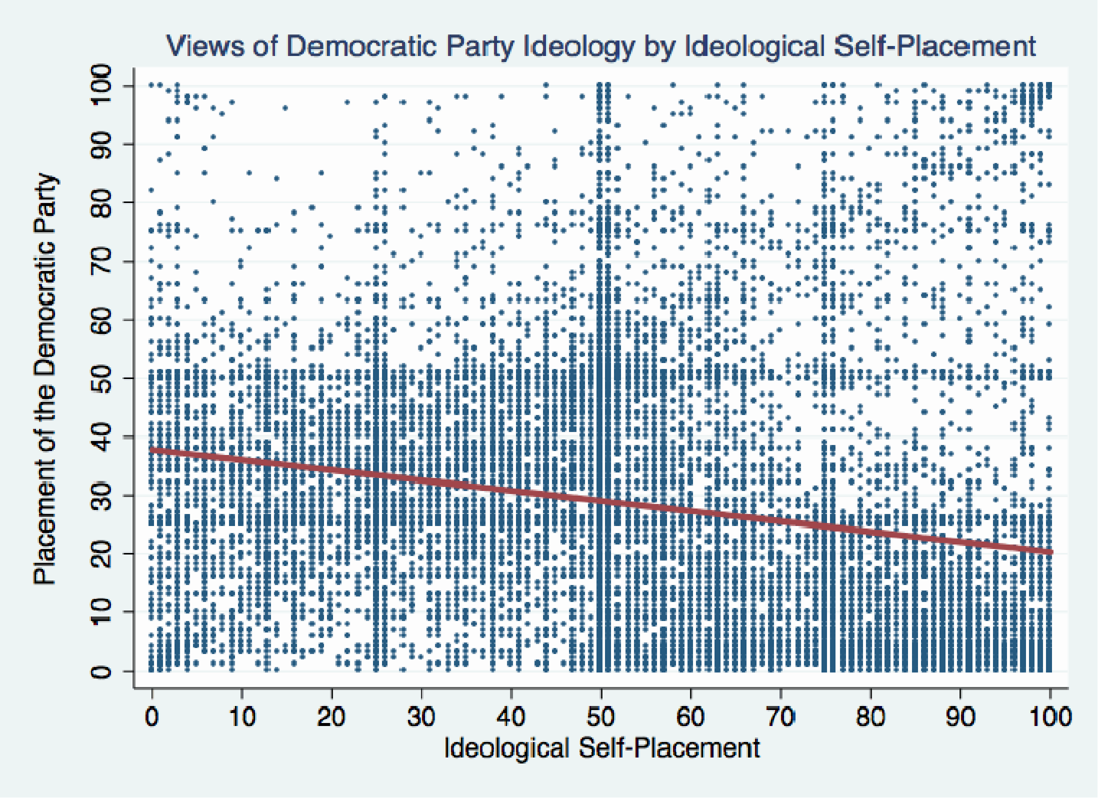
```
      Source |       SS       df       MS              Number of obs =   29493
-------------+------------------------------           F(  1, 29491) = 1382.46
       Model |  642887.802     1  642887.802           Prob > F      =  0.0000
    Residual |  13714213.2 29491  465.030455           R-squared     =  0.0448
-------------+------------------------------           Adj R-squared =  0.0447
       Total |    14357101 29492  486.813405           Root MSE      =  21.565

------------------------------------------------------------------------------
      cc317b |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      cc317a |  -.1752031   .0047121   -37.18   0.000    -.1844391   -.1659672
       _cons |   37.82903   .2834699   133.45   0.000     37.27342    38.38464
------------------------------------------------------------------------------
```

The slope coefficient suggests that a one point increase in one's ideological position (one point more conservative) is associated with a .175 point decrease in one's view of the Democratic party. The confidence interval implies that we can conclude with 95% confidence that the true population parameter lies between -.1844 and -.165. More precisely, if we took repeated samples using the same procedure, the confidence interval would contain the true parameter value 95% of the time. Finally the standard error of the regression tells us that our in-sample predictions are off by 21.56 points on average.

Graphing the relationship, including the best-fit line (adjusting for weights), customizing to make datapoints smaller, including reasonable scales and tick marks on the x and y axes, getting rid of the legend, and making line thicker

```
. graph twoway (scatter cc317b cc317a, msize(tiny)) (lfit cc317b cc317a [aweight=v200],
clwidth(thick)), title("Views of Democratic Party Ideology by Ideological Self-Placement",
size(medium)) ylabel(0 (10) 100) xlabel (0 (10) 100) xtitle("Ideological Self-Placement")
ytitle("Placement of the Democratic Party" " ") legend(off)
```



The slope coefficient tells us, broadly speaking, that the more conservative as individual is, the more liberal they perceive the Democratic Party. This effect, however, does not appear to be very large: a ten point increase in ideological self-placement (10 points more conservative) is only associated with a 1.75 point decrease in the view of the Democratic Party (1.75 points more liberal).

Based on our model, the average individual (in this sample, self-placement of about 55) would give the Democratic Party a score of 28.2.

$Y = \beta_0 + \beta_1 x_1 + \varepsilon_i$

$Y= 37.82 + (-.175 \text{ x } 55) = 28.2$

Part II

Examine variables and coding schemes

```
. tabulate v246, missing

. tabulate v246, missing nol

. tabulate cc307a, missing

. tabulate cc307a, missing nol
```

Recoding family income to meaningful values using the midpoint of the range and an arbitrary value for the highest category. Missing values are recoded as dots

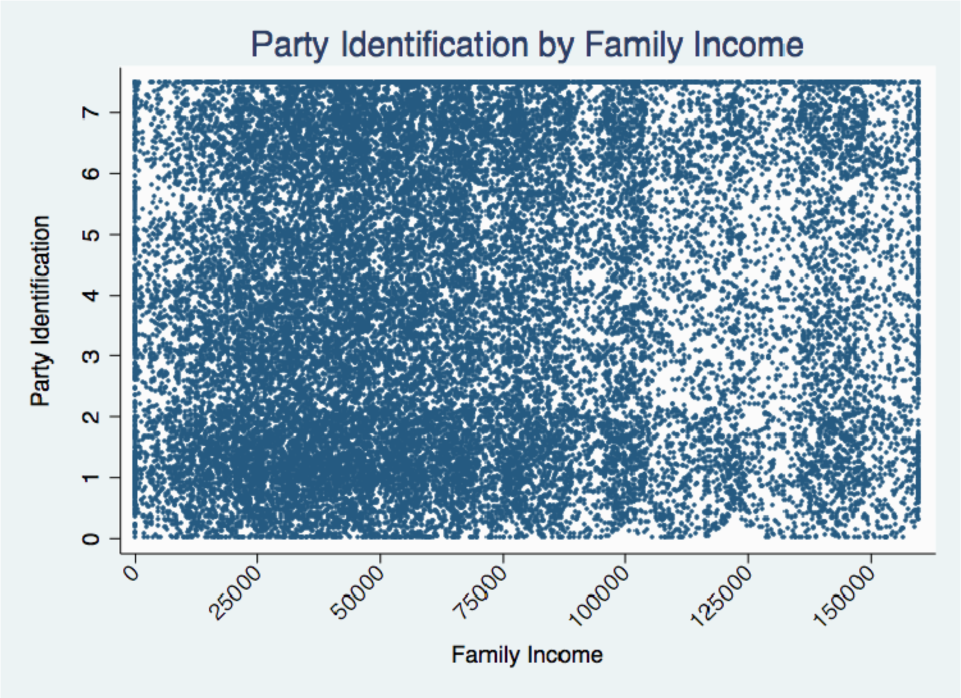
```
. recode v246 (1=5000) (2=12500) (3=17500) (4=22500) (5=27500) (6=35000) (7=45000)
(8=55000) (9=65000) (10=75000) (11=90000)(12=110000) (13=135000) (14=150000) (15=.)
```

Recoding missing values to dots in party ID

```
. recode cc307a (8=.)
```

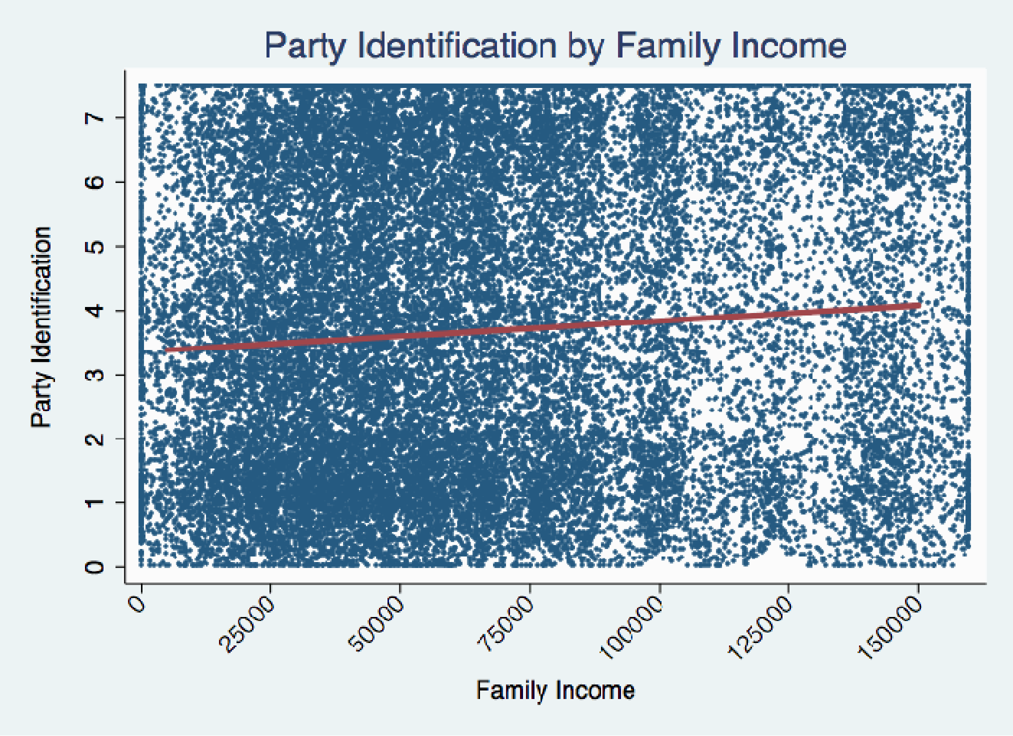
Graphing party ID as a function of family income, using the jitter command to make density of data visible, customizing axes and turning off legend

```
. graph twoway (scatter cc307a v246, jitter(20) msize(tiny)), title("Party Identification by Family
Income") ytitle("Party Identification" " ") xtitle("Family Income") xlabel(0(25000)160000,
angle(45)) ylabel(0(1)7.5, nogrid) legend(off)
```



Adding in regression line with analytical weights

```
. graph twoway (scatter cc307a v246, jitter(20) msize(tiny)) (lfit cc307a v246 [aweight=v200],
clwidth(thick)), title("Party Identification by Family Income") ytitle("Party Identification" " ")
xtitle("Family Income") xlabel(0(25000)160000, angle(45)) ylabel(0(1)7.5, nogrid) legend(off)
```



Recoding family income to be between 0 and 1

. gen v246_01=(v246-5000)/(150000-5000)

Regressing party ID on recoded family income

. reg cc307a v246_01 [aweight=v200]

```
      Source |       SS       df       MS              Number of obs =   29665
-------------+------------------------------           F(  1, 29663) =  216.17
       Model |  1045.44673     1  1045.44673           Prob > F      =  0.0000
    Residual |  143456.156 29663   4.8361985           R-squared     =  0.0072
-------------+------------------------------           Adj R-squared =  0.0072
       Total |  144501.603 29664  4.87127841           Root MSE      =  2.1991

------------------------------------------------------------------------------
      cc307a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     v246_01 |    .693826   .0471902    14.70   0.000     .6013311    .7863208
       _cons |   3.385286    .021906   154.54   0.000     3.342349    3.428223
------------------------------------------------------------------------------
```

The coefficient on the income variable suggests that moving from the minimum to the maximum in family income is associated with .69 point increase in party identification (moving toward stronger Republican). The intercept tells us that those with the minimum family income have a party identification score of 3.38 (between lean Democrat and Independent).

Collapsing dataset by state to get state-level averages of family income, party ID, and analytical weights

. collapse (mean) v246 cc307a v200, by(v206)

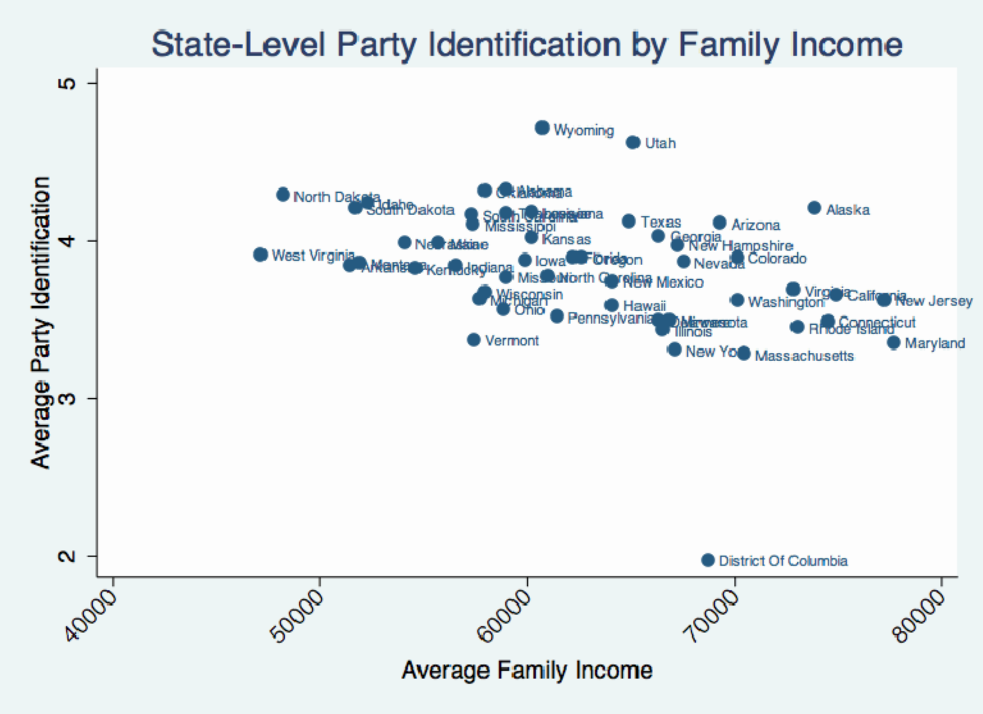
Convert state variable to string in order to facilitate capitalizing first letter with ‘proper’ command

```
. decode v206, generate(state)
```

```
. replace state = proper(state)
```

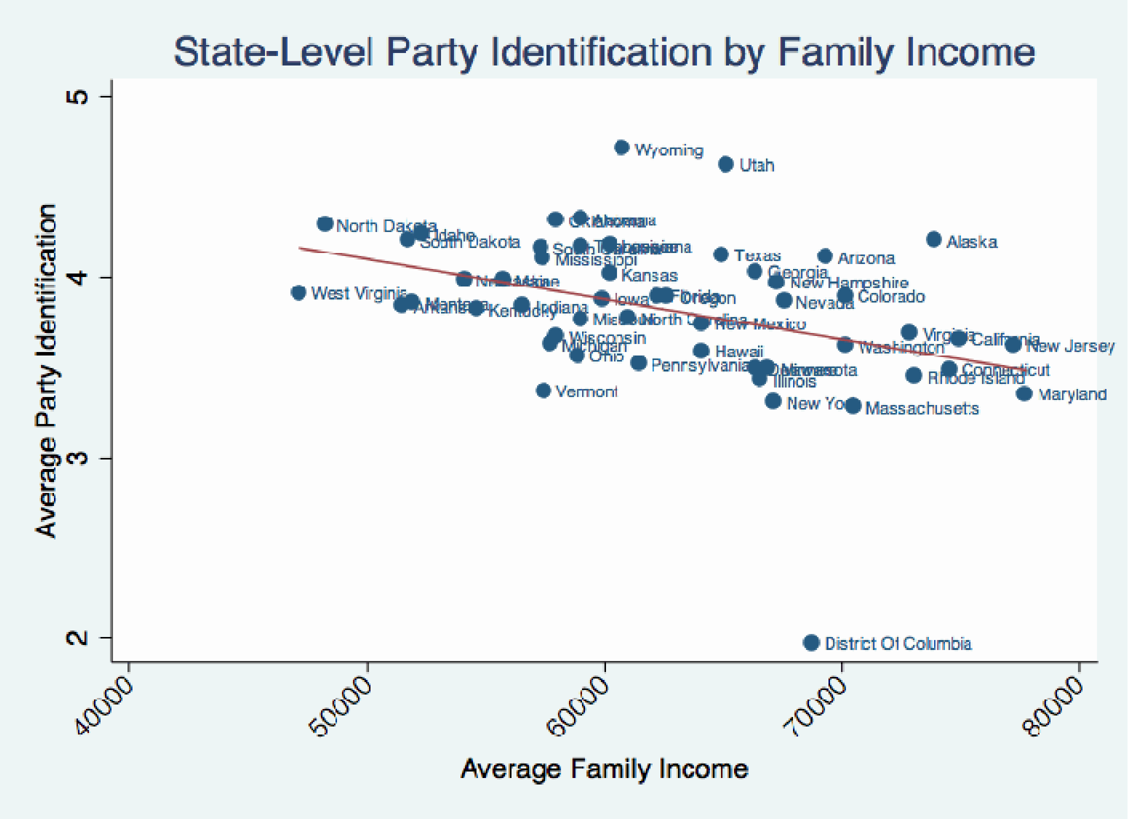
Generating a scatter plot of the state-level relationship between average family income and average party ID, including state marker labels and customizing axes to maximize the visbility of the variation in the data

```
. scatter cc307a v246, mlabel(state) mlabsize(vsmall) title(State-Level Party Identification by Family Income) ytitle(Average Party Identification) xtitle(Average Family Income) xlabel(40000(10000)80000, angle(45)) ylabel(2(1)5, nogrid) legend(off)
```



Adding in regression line with analytical weights

```
. graph twoway (scatter cc307a v246, mlabel(state) mlabsize(vsmall)) (lfit cc307a v246 [aweight=v200]), title(State-Level Party Identification by Family Income) ytitle(Average Party Identification) xtitle(Average Family Income) xlabel(40000(10000)80000, angle(45)) ylabel(2(1)5, nogrid) legend(off)
```



Recoding average family income variable to be between 0 and 1, using egen command store minimum and maximum of the variable

```
. egen v246_min=min(v246)

. egen v246_max=max(v246)

. gen v246_01=(v246-v246_min)/(v246_max-v246_min)
```

Regressing average party ID on average family income at the state-level, including analytical weights

```
. reg cc307 v246_01 [aweight=v200]


      Source |       SS       df       MS              Number of obs =      51
-------------+------------------------------           F(  1,    49) =    8.67
       Model |  1.41455823     1  1.41455823           Prob > F      =  0.0049
    Residual |  7.99141329    49  .163090067           R-squared     =  0.1504
-------------+------------------------------           Adj R-squared =  0.1331
       Total |  9.40597152    50   .18811943           Root MSE      =  .40384

------------------------------------------------------------------------------
      cc307a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
------------+----------------------------------------------------------------
    v246_01 |  -.6799197   .2308665    -2.95   0.005    -1.143863    -.215976
      _cons |   4.163811   .1289379    32.29   0.000       3.9047    4.422921
-----------------------------------------------------------------------------
```

The coefficient on average family income here suggests that moving from the state with the minimum income to the state with the maximum income is associated with a .67 point decrease in party ID score (i.e., becoming more Democratic). The intercept means that when average family income is at its minimum, the expected party ID score is 4.16 (between indepen dent and Lean Republican).

The relationship between family income and party ID differs markedly depending on whether we are analyzing states or individuals. At the individual level, higher income is associated with a preference for the Republican Party; at the state-level, higher average income is associated with a preference for the Democratic Party. One possible explanation for this is that wealth is correlated with urbanization at the state level, and that more urban states tend to vote more Democratic, whereas rural states tend to vote more conservatively (for a variety of cultural reasons unrelated to wealth).

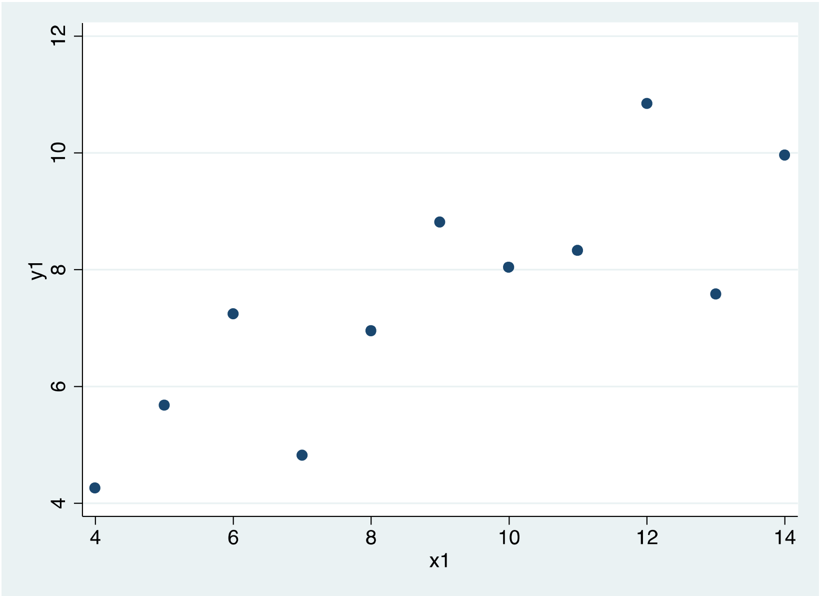
Part III

. clear

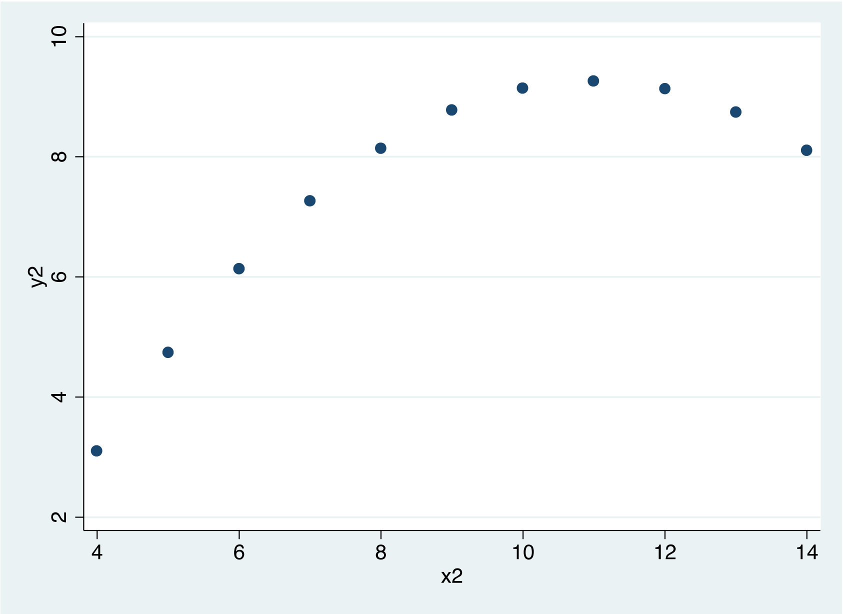
. use "/Users/nlmiller/Desktop/Poli Sci Lab/PS3/quartet.dta", clear

Visualizing the data with scatter plots

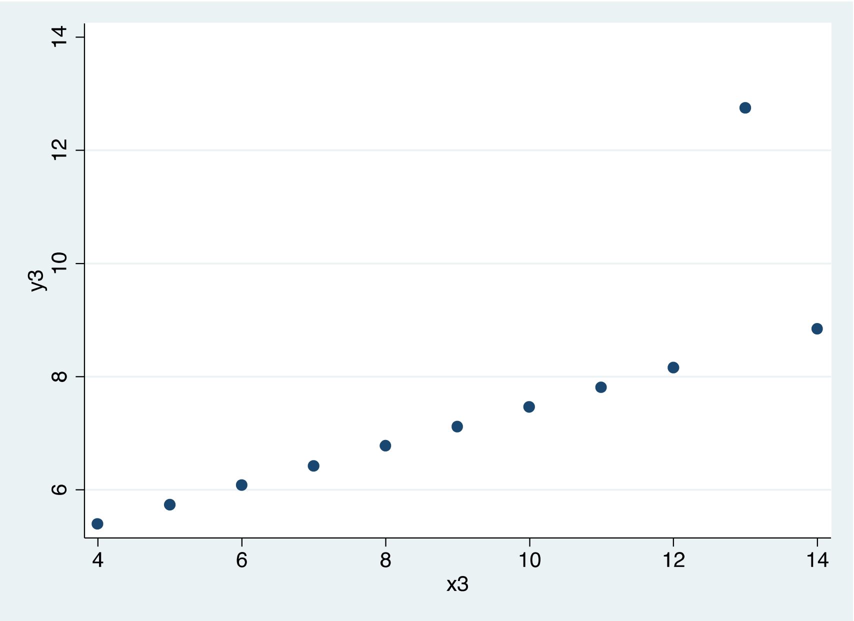
. scatter y1 x1



```
. scatter y2 x2
```

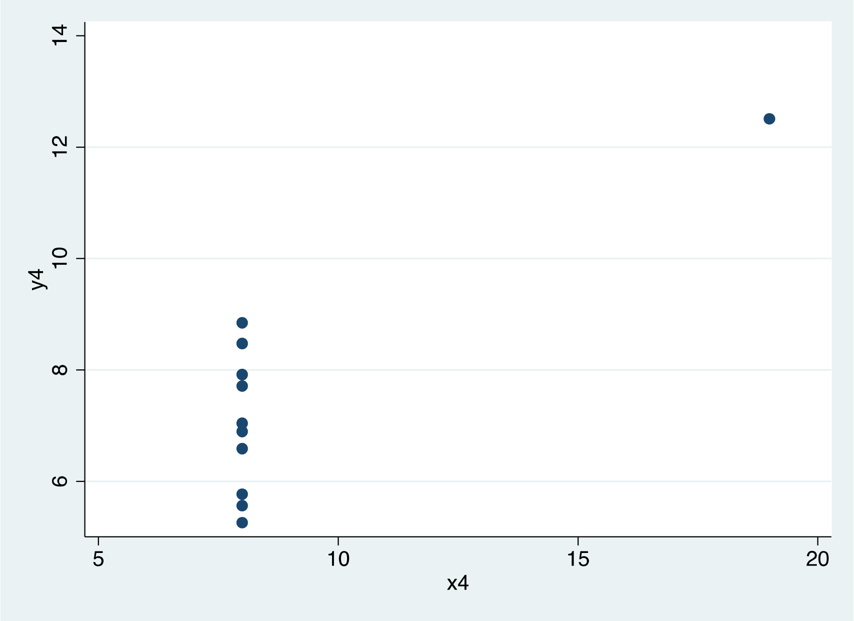


```
. scatter y3 x3
```



.

. scatter y4 x4



Regressing the Ys on the respective Xs

. reg y1 x1

. reg y2 x2

. reg y3 x3

. reg y4 x4

	Y1 on X1	**Y2 on X2**	**Y3 on X3**	**Y4 on X4**
β_1	0.500	0.500	0.499	0.500
β_1 CI	.233, .767	.233, .767	.233, .767	.233, .766
β_o	3.000	3.000	3.002	3.001
SER	1.237	1.237	1.236	1.236

The coefficients on the X variables suggest that a one unit increase in X is associated with a 0.5 unit increase in Y across all four datasets. The intercepts imply that the expected value of Y is 3 when X is 0 in all four datasets. The standard error of the regression shows that in-sample predictions are off the mark by 1.237 on average for all four models.

These estimates are believable in the sense that they identify the line that best fits the data, i.e., the line that minimizes the sum of the squared residuals. However, as the scatter plots above illustrate, the models do a poor job capturing the true relationships between the Xs and Ys in cases where the underlying relationships aren't linear (Y2 and X2, Y4 and X4), and/or when there is very little variation in X that can be used to explain the variation in Y (Y4 and X4).

We should conclude from this exercise that it is important to visually examine the data before fitting a simple linear model. Many relationships aren't linear and require transformation of the variables in order to be accurately captured. In other cases (i.e. Y4 and X4), we probably shouldn't be fitting a model at all since there is so little variation in our explanatory variable.

Part IV

```
. clear

. use "/Users/nlmiller/Desktop/Poli Sci Lab/PS3/cces08_common_output.dta"

*Recoding family income variable as above*

. recode v246 (1=5000) (2=12500) (3=17500) (4=22500) (5=27500) (6=35000) (7=45000)
(8=55000) (9=65000) (10=75000) (11=90000)(12=110000) (13=135000) (14=150000) (15=.)

*Recoding missing values to dots in party ID*

. recode cc307a (8=.)

*Generating logged family income variable*

. gen logincome=log(v246)

*Regressing party ID on logged family income*

. reg cc307a logincome [aweight=v200]
```

```
      Source |       SS       df       MS              Number of obs =   30548
-------------+------------------------------           F(  1, 30546) =   53.38
       Model |  291.573734     1  291.573734           Prob > F      =  0.0000
    Residual |  166840.756 30546  5.46195101           R-squared     =  0.0017
-------------+------------------------------           Adj R-squared =  0.0017
       Total |  167132.329 30547  5.47131729           Root MSE      =  2.3371

------------------------------------------------------------------------------
      cc307a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
   logincome |   .1211822   .0165859     7.31   0.000     .0886732    .1536912
       _cons |   2.545954   .1782457    14.28   0.000     2.196585    2.895323
------------------------------------------------------------------------------
```

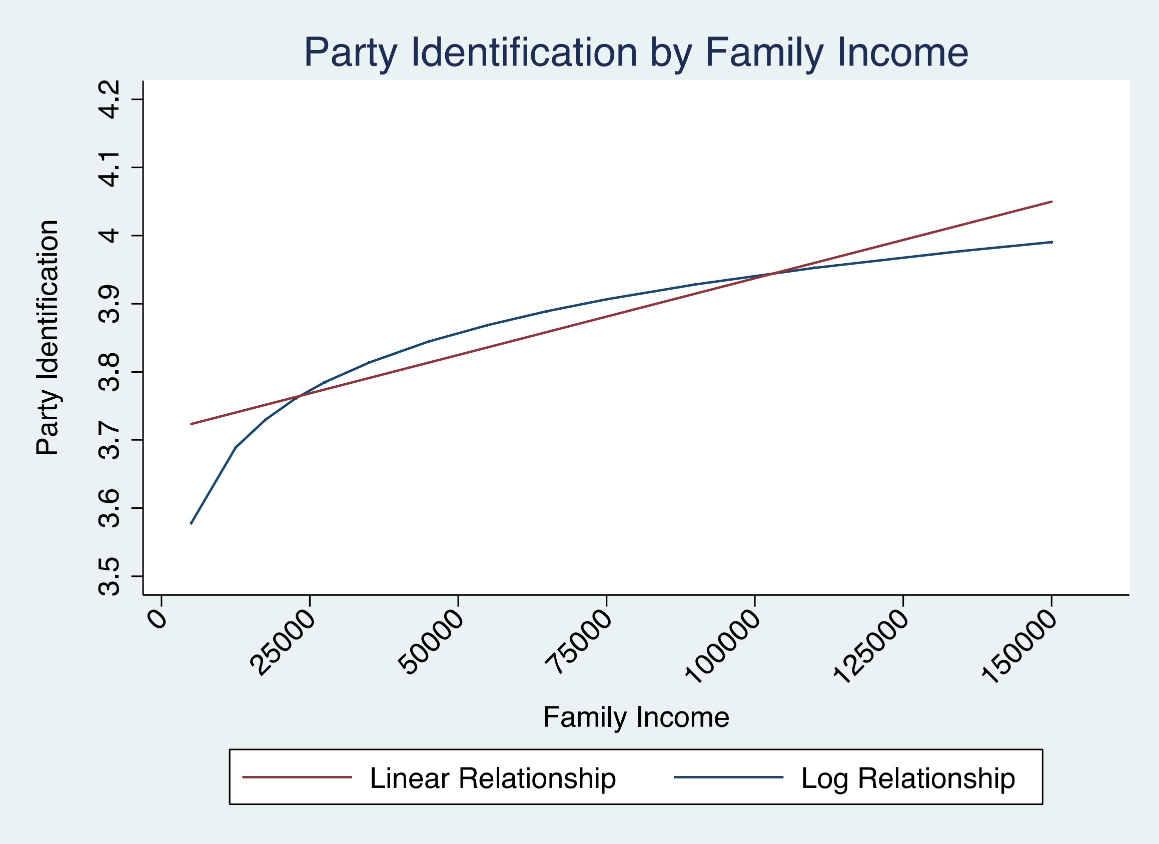
The coefficient on logincome tells us that a one percentage point increase in family income is associated with a .12/100 (.0012) point increase in Party ID score (i.e. moving toward more Republican).

* Generating a variable that records the fitted values for the logged income model*

. predict log_fitted

Graphing the linear and logged relationships between income and party ID, using the fitted values from the regression and the line command for the logged relationship

. graph twoway (line log_fitted v246, sort) (lfit cc307a v246 [aweight=v200]), title(Party Identification by Family Income) ytitle("Party Identification" " ") xtitle(Family Income) xlabel(0(25000)160000, angle(45)) ylabel(3.5 (0.1) 4.2, nogrid) name(income, replace) legend(order (2 "Linear Relationship" 1 "Log Relationship"))



The log transformation on the independent variable changes the interpretation from a one dollar change in income to a one percentage point change in income. While the two models are both linear in the parameters (the beta coefficients), the log-transformed model is no longer linear in the variables, allowing us to fit a relationship where similar absolute changes in X have bigger or smaller effects on Y depending on the level of X, as the graph above illustrates.

Part V

```
. clear

. use "/Users/nlmiller/Desktop/Poli Sci Lab/PS3/NMC.dta"
```

Recoding missing values as dots

```
. replace milex=. if milex==-9

. replace irst=. if irst==-9
```

Creating military expenditure and iron and steel per capita variables. Note that all variables are measured in thousands

```
. gen milex_pc=milex/tpop if year==2007

. gen irst_pc=irst/tpop if year==2007
```

Pulling up the variance-covariance matrix for the two variables of interest

```
. corr milex_pc irst_pc, covariance
```

```
             | milex_pc  irst_pc
-------------+------------------
    milex_pc |   143088
     irst_pc |  42.6647  .263279
```

$$\beta_1 = \frac{cov\ (X,Y)}{var\ (X)}$$

$$\beta_{irst\_pc} = \frac{42.6647}{.263} = 162.23$$

The coefficient suggests that a one-ton increase in iron and steel production per capita is associated with a $162.23 increase in military expenditures per capita.

```
. log close
```