17.871: Solutions for Problem Set 3

- Red Text denotes stata code
- Green text denotes stata output, e.g. tables
- */*Italics*/* denotes comments on code/output

Part I

(1)

```
use cces12_common_subset.dta,clear

gen ideo5_new = .  /*recode to get rid of 'not sure'*/
replace ideo5_new = ideo5 if ideo5<6
gen CC308c_new = .
replace CC308c_new = CC308c if CC308c<5
reg CC308c_new ideo5_new [aw=V103]  /*Regression with weights*/
```

```
CC308c_new    Coef.     Std. Err.    t      P>t      [95% Conf.  Interval]

ideo5_new   .1309917   .0055644   23.54   0.000    .120085     .1418983
_cons        2.36804   .0184406   128.41   0.000   2.331896    2.404185
```

(2)

*/*Moving one point along the 5-point ideology scale (from "very liberal" to "liberal", for example) is associated with an increase of 0.13 in respondents' disapproval of the Supreme Court, where disapproval is measured on a 4-point scale*/*

(3)

/[we didn't cover this yet in class, so it is extra credit and doesn't count toward the pset grade]. Correct answer: there is a 95% probability that the 95% confidence interval contains the true population coefficient; in other words, if we did this again and again on many different samples, 95% of the confidence intervals we calculate would contain the true coefficient. So in this case, there is a 95% probability that the interval (0.12, 0.14) contains the true coefficient. [Note that it is **NOT** correct to say that there is a 95% probability that the true coefficient is between 0.12 and 0.14!]*/*

(4)

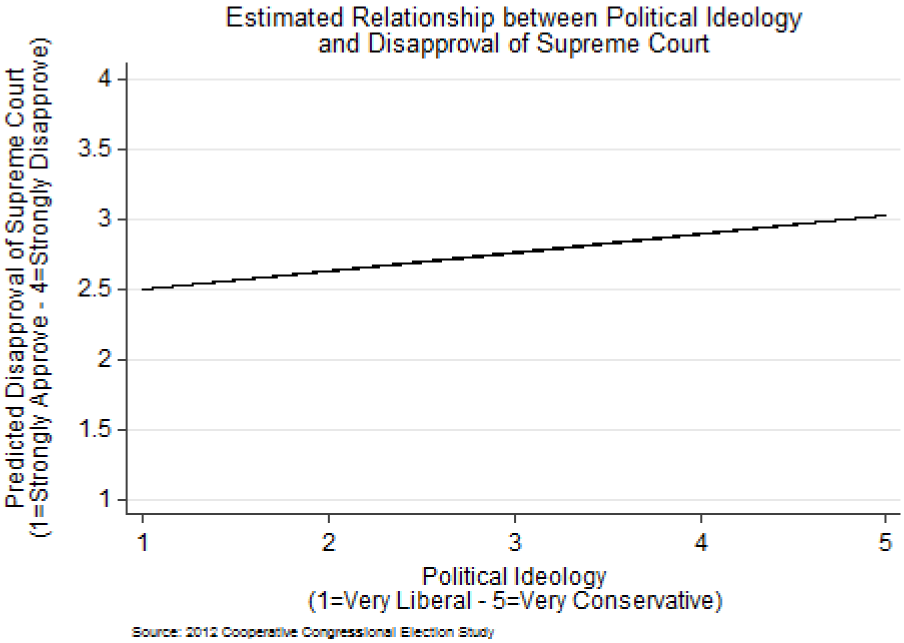
*/*The SER is 0.862. It measures how far the predicted values deviate from the actual sample values. It means that the predicted values on average deviate from the true values by about 0.862.*/*

(5)

```
predict fitted_vals  /*Get the fitted values from our regression*/
```

*/*Graph of regression line – we don’t need the original data because it would be completely uninformative. It would form a ‘lattice’; i.e. there would be at least one observation for every possible combination of ideology and supreme court approval.*/*

```
set scheme tufte
graph twoway line fitted_vals ideo5_new, ///
xtitle("Political Partisanship" "(1=Strong Democrat - 7=Strong Republican)", margin(0 0 0 2)) ///
xlabel(1(1)5) ///
yscale(range(1 4)) ylabel(1(0.5)4) ///
ytitle("Predicted Disapproval of Supreme Court" "(1=Strongly Approve - 4=Strongly Disapprove)",
margin(0 2 0 0)) ///
title("Estimated Relationship between Political Partisanship" "and Disapproval of Supreme
Court",size(medium)) ///
note("Source: 2012 Cooperative Congressional Election Study",size(vsmall))
```



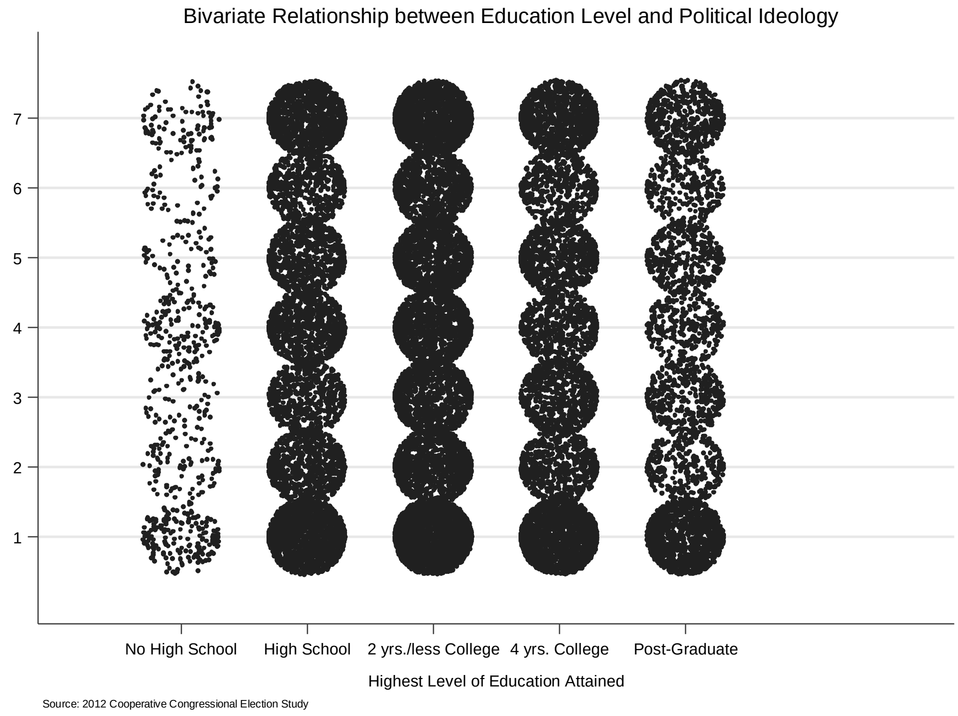
(6)

*/*The slope coefficient suggests that more Conservative voters are somewhat more likely than Liberals to disapprove of the Supreme Court. This could be because 2012 saw the Supreme Court rule in favor of the Affordable Care Act, which was controversial amongst Conservative voters. But more generally, the Supreme Court has been somewhat unpopular with Conservatives ever since the Roe vs. Wade decision, and the pattern see here may be a lingering legacy of that. The coefficient of 0.089 implies that moving from being very liberal to very conservative is associated with an increase of 5*0.131=0.66 points along the 4-point scale (about two-thirds of a standard deviation). This is arguably quite a modest effect, as the quite flat regression line above implies.*/*

Part 2

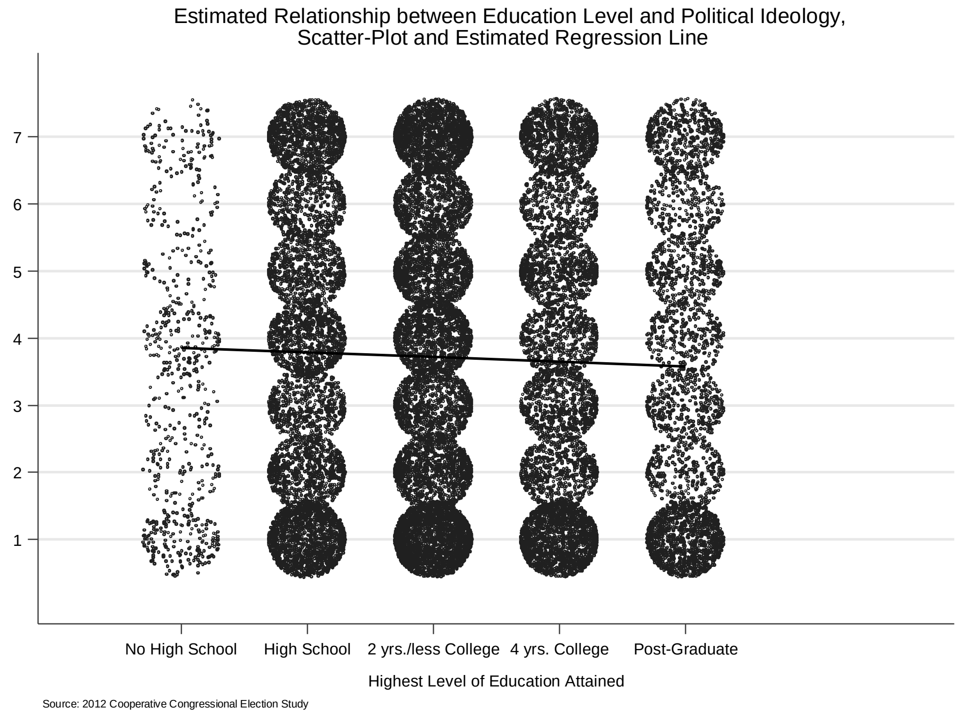
(1)

```
gen pid7_new=.
replace pid7_new=pid7 if pid7<8
gen educ_new = . /*Let’s throw together those with ‘some college’ and ‘2 years college*/
replace educ_new = 1 if educ==1 /*no high school*/
replace educ_new = 2 if educ==2 /*completed high school*/
replace educ_new = 3 if educ==3|educ==4  /*2 years or less college*/
replace educ_new = 4 if educ==5 /*4-year college grad*/
replace educ_new = 5 if educ==6 /*postgrad*/
/*Label the new variable for the graph*/
label define educ_cats 1 "No High School" 2 "High School" 3 "2 yrs. or less College" 4 "4 yrs. College" 5 "Post-Graduate"
label values educ_new educ_cats
/*Scatter-plot with jitter. Use the msize() option to lower the size of points*/
twoway (scatter pid7_new educ_new,jitter(10) msize(tiny)), ///
xscale(range(0 7)) xlabel(1 2 3 4 5, valuelabel labsize(tiny)) ///
yscale(range(0 8)) ylabel(1(1)7, labsize(vsmall)) ///
ytitle("Political Partisanship" "(1=Strong Democrat - 7=Strong Republican)", margin(0 2 0 0) size(vsmall)) ///
xtitle("Highest Level of Education Attained", margin(0 0 0 2) size(vsmall)) ///
title("Estimated Relationship between Education Level and Political Partisanship",size(small)) ///
note("Source: 2012 Cooperative Congressional Election Study",size(tiny))
```



(2)

```
graph twoway (scatter pid7_new educ_new,jitter(10) msize(vtiny)) (lfit pid7_new educ_new, lpattern(solid) lwidth(medium)), ///
xscale(range(0 7)) xlabel(1 2 3 4 5, valuelabel labsize(tiny)) ///
yscale(range(0 8)) ylabel(1(1)7, labsize(vsmall)) ///
ytitle("Political Partisanship" "(1=Strong Democrat - 7=Strong Republican)", margin(0 2 0 0) size(vsmall)) ///
xtitle("Highest Level of Education Attained", margin(0 0 0 2) size(vsmall)) ///
title("Estimated Relationship between Education Level and Political Partisanship," "Scatter-Plot and Estimated Regression Line",size(small)) ///
note("Source: 2012 Cooperative Congressional Election Study",size(tiny)) ///
legend(off)
```


Estimated Relationship between Education Level and Political Ideology,
Scatter-Plot and Estimated Regression Line

Source: 2012 Cooperative Congressional Election Study

(3)

```
reg pid7_new educ_new
```

```
pid7_new    Coef.       Std. Err.    t     P>t        [95% Conf.Interval]

educ_new    -.0677673  .0133764    -5.07  0.000      -.0939858   -.0415488
_cons       3.921039   .0449281    87.27  0.000      3.832978    4.009101
```

*/*The coefficient on education suggests that a rise in education of 1 – from not graduating high school to graduating high school, for example – is associated with a -0.068 decline in political partisanship, where partisanship is measured on a 7-point scale, and lower values indicate that the respondent identifies more with the Democrats.*/*

(4)

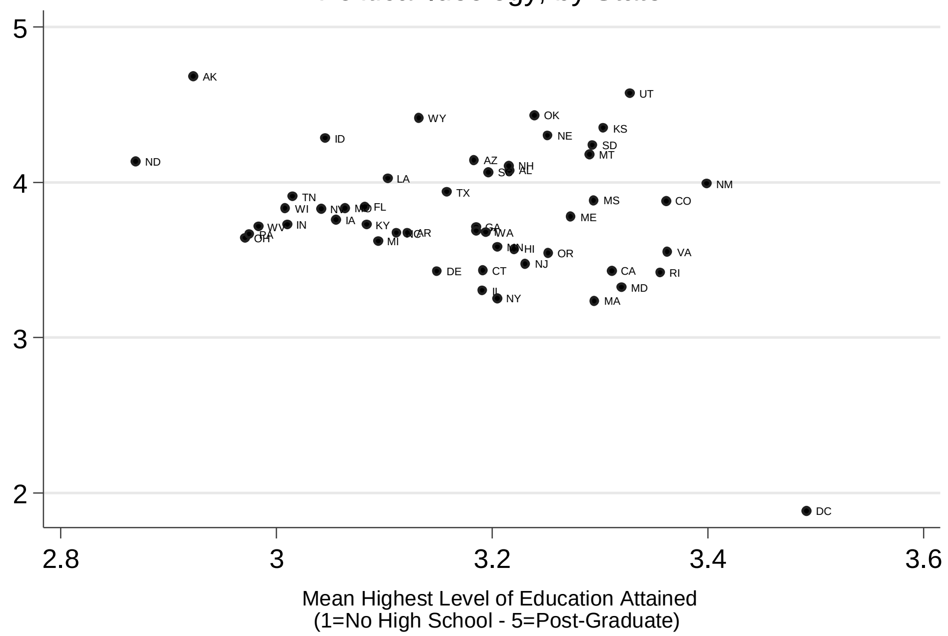
*/*Create dataset of means*/*

```
collapse (mean) mean_educ=educ_new mean_pid7=pid7_new, by(inputstate)
replace mean_educ=round(mean_educ, 0.1)
format mean_educ %12.2g  /*This changes the display format to two decimal places, making for better labelling on the chart
```

```
/*Generate a variable for abbreviated state names. _n indexes the row, so the first line just creates a variable from 1 to 51*/
gen statename = _n
label define snames 1 "AL" 2 "AK" 3 "AZ" 4 "AR" 5 "CA" 6 "CO" 7 "CT" 8 "DE" 9 "DC" 10 "FL" 11 "GA" 12 "HI" 13 "ID" 14 "IL" ///
15 "IN" 16 "IA" 17 "KS" 18 "KY" 19 "LA" 20 "ME" 21 "MD" 22 "MA" 23 "MI" 24 "MN" 25 "MS" 26 "MO" 27 "MT" 28 "NE" 29 "NV" ///
30 "NH" 31 "NJ" 32 "NM" 33 "NY" 34 "NC" 35 "ND" 36 "OH" 37 "OK" 38 "OR" 39 "PA" 40 "RI" 41 "SC" 42 "SD" 43 "TN" 44 "TX" ///
45 "UT" 46 "VT" 47 "VA" 48 "WA" 49 "WV" 50 "WI" 51 "WY"
label values statename snames

/*Chart – notice we use options beginning with "m" to control the look of the points and associated labels*/
twoway (scatter mean_pid7 mean_educ,mlabel(statename) mlabsize(tiny) msymbol(O) mfcolor(black) msize(small)), ///
ytitle("Mean Political Partisanship" "(1=Strong Democrat - 7=Strong Republican)", margin(0 2 0 0) size(small)) ///
xtitle("Mean Highest Level of Education Attained" "(1=No High School - 5=Post-Graduate)", margin(0 0 0 2) size(small)) ///
title("Relationship between Mean Education Level and Mean" "Political Partisanship, by State",size(medium)) ///
note("Source: 2012 Cooperative Congressional Election Study",size(tiny))
```

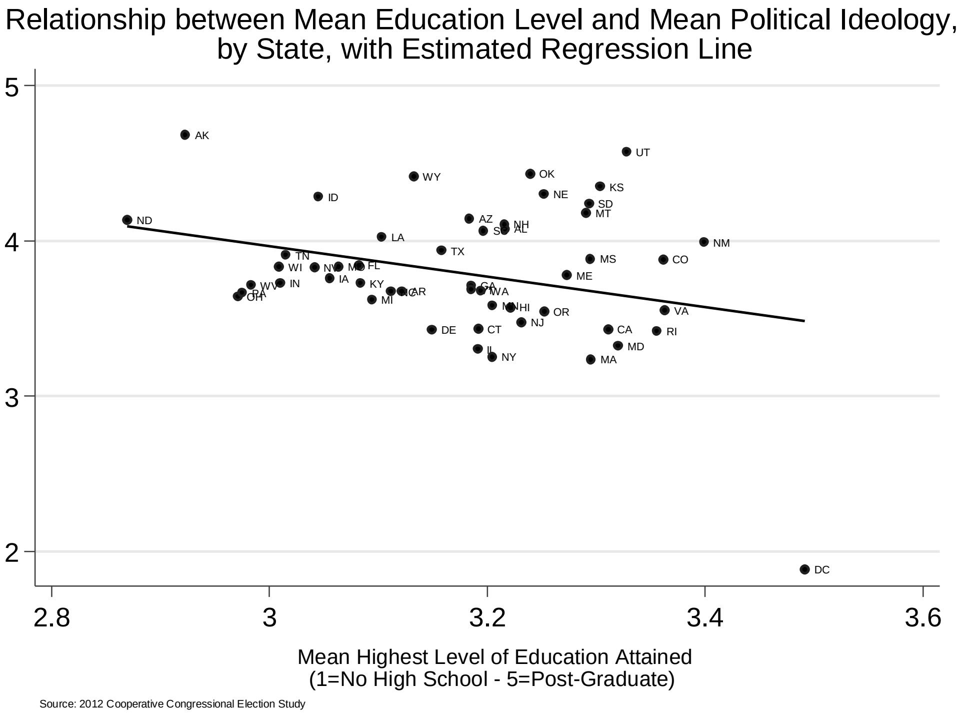
Relationship between Mean Education Level and Mean Political Ideology, by State



Source: 2012 Cooperative Congressional Election Study

(5)

```
graph twoway (scatter mean_pid7 mean_educ,mlabel(statename) mlabsize(tiny) msymbol(O) mfcolor(black) msize(small)) ///
(lfit mean_pid7 mean_educ, lpattern(solid) lwidth(medium)), ///
ytitle("Mean Political Partisanship" "(1=Strong Democrat - 7=Strong Republican)", margin(0 2 0 0) size(small)) ///
xtitle("Mean Highest Level of Education Attained" "(1=No High School - 5=Post-Graduate)", margin(0 0 0 2) size(small)) ///
title("Relationship between Mean Education Level and Mean Political Partisanship," "by State, with Estimated Regression Line",size(medium)) ///
note("Source: 2012 Cooperative Congressional Election Study",size(tiny)) ///
legend(off)
```


Relationship between Mean Education Level and Mean Political Ideology, by State, with Estimated Regression Line

Source: 2012 Cooperative Congressional Election Study

(6)

reg mean_pid7 mean_educ

```
mean_pid7   Coef.       Std. Err.    t       P>t         [95% Conf. Interval]

mean_educ   -.9790724   .4528302    -2.16   0.036       -1.889069   -.0690761
_cons        6.903584   1.440034     4.79   0.000        4.009726    9.797441
```

*/*The results suggest that a rise in mean education of 1 – from not graduating high school to graduating high school, for example – is associated with a -0.97 decline in political partisanship, where partisanship is measured on a 7-point scale, and lower values indicate that the respondent identifies more with the Democrats.*/*

(7)

*/*We might have expected a strong negative relationship between education and partisanship: assuming that education is a good proxy for earnings, we would probably expect poorer voters to vote Democrat, and richer voters to vote Republican. However, the relationship between education and partisanship at the individual level is actually very weak: even going from the lowest to the highest level of education is not associated with much change in partisanship.*

But at the state level, the relationship is very strong: better educated states are, on average, much more likely to be Democratic. Overall, the pattern suggests that geography is a very strong predictor of partisanship – much more so than individuals' education. The weak individual-level relationship is probably the result of voters' moral values competing with their economic interests. Highly-educated people are often socially liberal, and therefore support the Democrats, and some poorly-educated people are socially conservative. At the state level, it seems that better-educated states are more ideologically liberal, perhaps for long-standing cultural reasons, and/or because people tend to migrate to states that fit with their ideology. */

Part 3

(1)

```
use quartet.dta,clear
reg y1 x1
reg y2 x2
reg y3 x3
reg y4 x4
```

	Regression 1	**Regression 2**	**Regression 3**	**Regression 4**
Slope	0.5	0.5	0.5	0.5
Constant	3	3	3	3
Confidence Interval	(0.23, 0 .77)	(0.23, 0 .77)	(0.23, 0 .77)	(0.23, 0 .77)
S.E.R.	1.24	1.24	1.24	1.24

/*Each regression produces identical results!*/

(2)

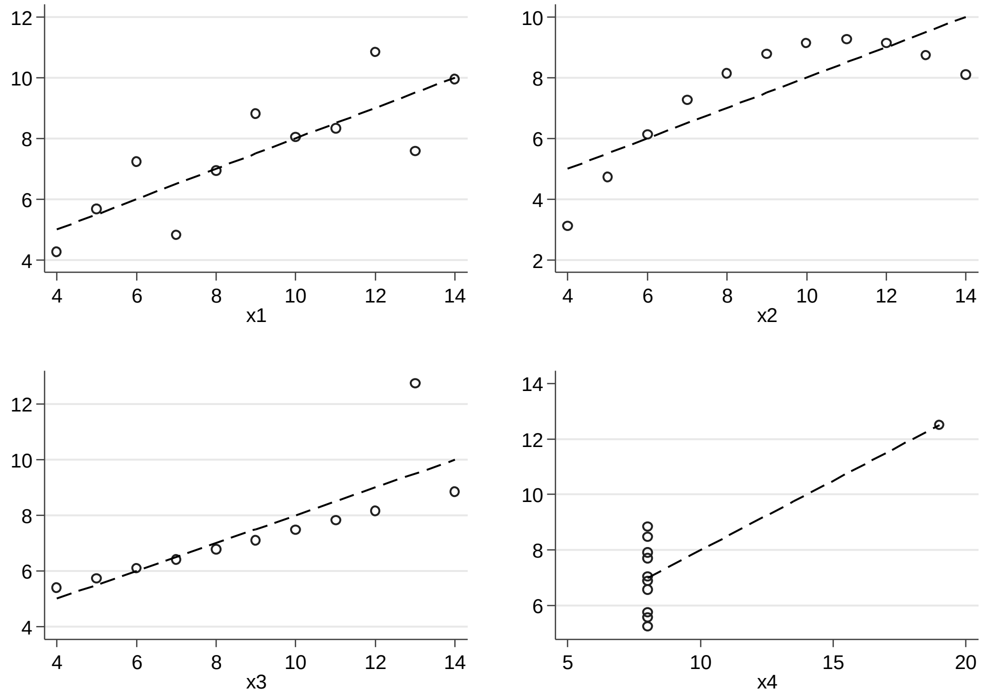
/*The coefficients imply that a one-unit change in x is associated with a rise of 0.5 in y. The SERs suggest that, on average, the predicted values differ from the true values by about 1.24*/

(3)

/*It's best to plot the data and the associated regression lines to assess plausibility*/

```
graph twoway (scatter y1 x1) (lfit y1 x1), legend(off) saving(first,replace)
graph twoway (scatter y2 x2) (lfit y2 x2), legend(off) saving(second,replace)
graph twoway (scatter y3 x3) (lfit y3 x3), legend(off) saving(third,replace)
graph twoway (scatter y4 x4) (lfit y4 x4), legend(off) saving(fourth,replace)

graph combine first.gph second.gph third.gph fourth.gph, rows(2) cols(2)
```



*/*The first and third regressions seem like reasonable fits to the data, although the third is quite influenced by a single positive outlier. If you were writing a paper with this dataset, you would want to investigate this datapoint more closely (to check whether it is miscoded, for example).*

The second and fourth regression lines, however, do not seem very believable. The second dataset clearly shows a curvilinear relationship between x2 and y2, but we have fitted a straight line. While we are still capturing some of the true relationship, predictions are particularly poor for low and high values of x2. In this case, it would be better to change the functional form to use x2 squared. The fourth dataset is the least believable. The regression line is entirely determined by one outlier; other than that, there is no variation at all in x4. It is questionable whether it is even worthwhile running this regression in the first place./*

(4)

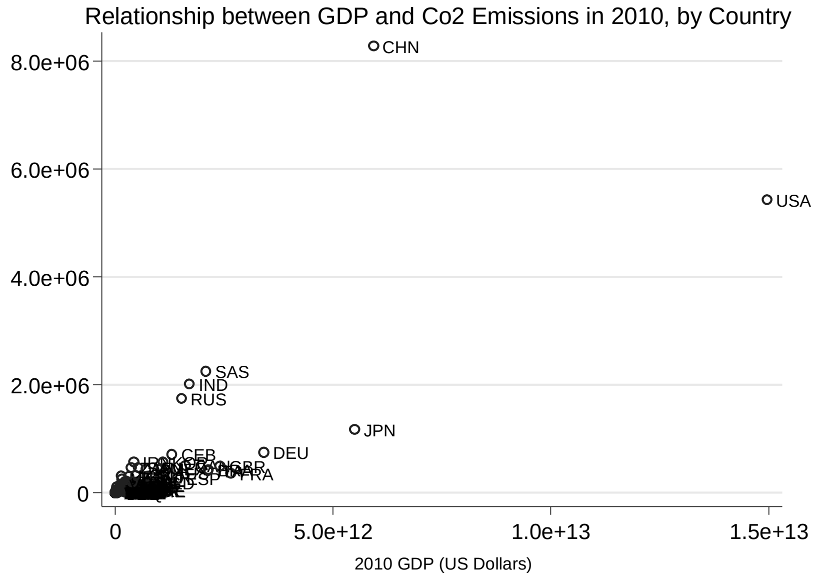
*/*The conclusion we should draw is that fancy statistical techniques are no substitute for simply plotting your data and looking at relationships visually. It's always a good idea to explore your data before moving to running regressions. In some situations, like the fourth dataset, regression may be entirely inappropriate.*/*

Part 4

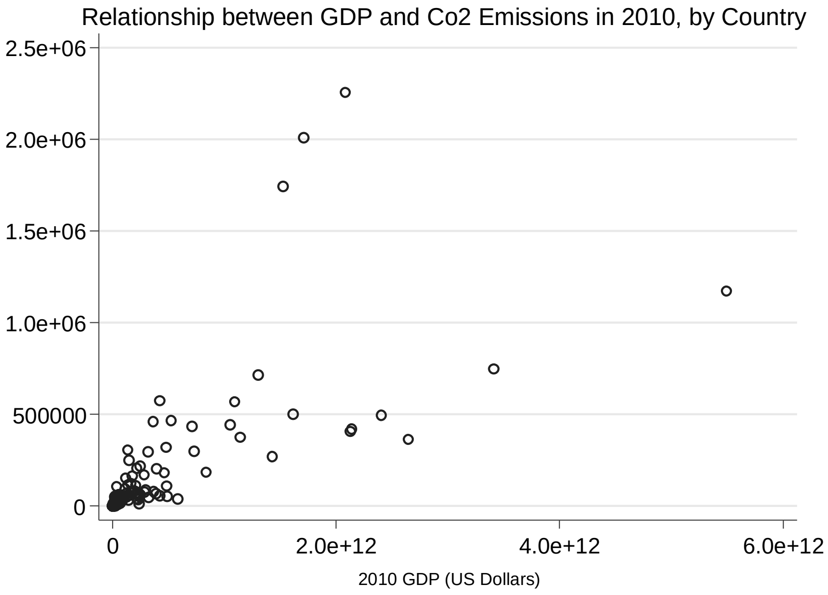
(1)

*/*Let’s plot the data first*/*

```
twoway (scatter co2_2010 gdp2010, mlabel(country_code)), ///
ytitle("2010 CO2 emissions (kilotons", margin(0 2 0 0) size(small)) ///
xtitle("2010 GDP (US Dollars)", margin(0 0 0 2) size(small)) ///
title("Relationship between GDP and Co2 Emissions in 2010, by Country",size(medium))
```



*/*We can immediately see two significant outliers: the US and China, which make it hard to discern the overall relationship – although it looks somewhat positive. So let’s plot it without them:*/*



*/*Now, it looks as though there is a weakly positive relationship between GDP and emissions, with variability increasing with both GDP and emissions*/*

(2)

```
gen gdp_bill = gdp2010/1000000000  /*re-code into billions*/
reg co2_2010 gdp_bill
predict co2_hat, xb  /*fitted vals - we'll use these later*/
```

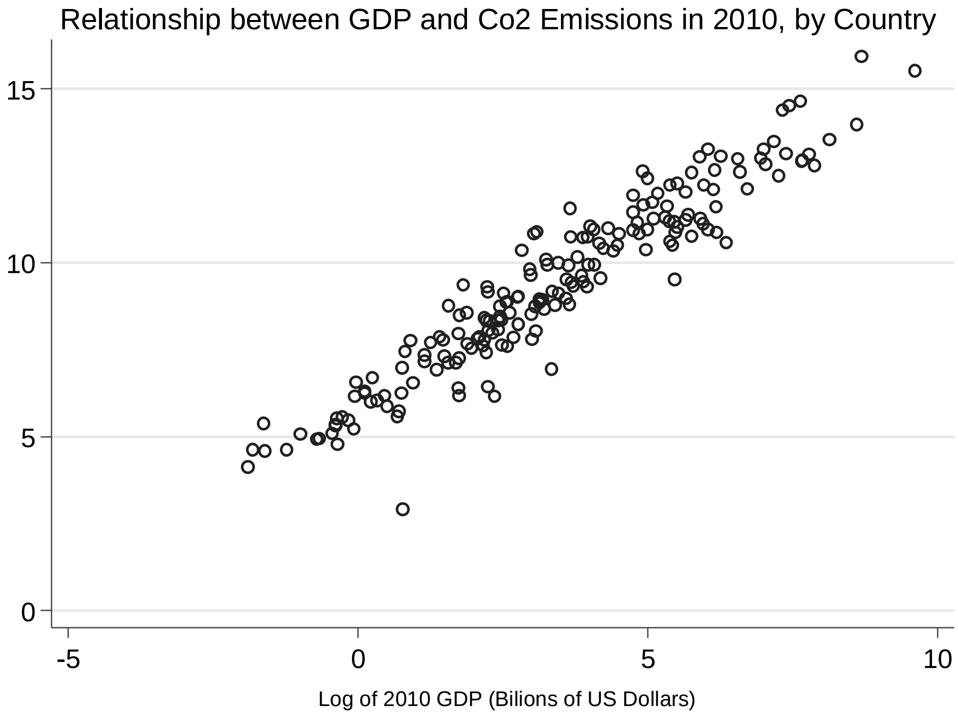
```
----------------------------------------------------------------------
    co2_2010 | Coef.   Std. Err.    t     P>|t|     [95% Conf. Interval]
-------------+--------------------------------------------------------
    gdp_bill |457.3367 26.48057  17.27   0.000     405.0977    509.5757
       _cons |16555.95 35932.55  0.46    0.646    -54329.31    87441.21
----------------------------------------------------------------------
```

(3)

/ The coefficient on gdp suggests that a rise of GDP of 1 billion dollars is associated with an increase in emission of 457 kilotons. The constant implies that a country with GDP of 0 is estimated to produce 16,555 kt of emissions.*/*

(4)

```
gen l_co2  = log(co2_2010)
gen l_gdp = log(gdp_bill)
/*Let's plot it to see the effect of logging…*/
scatter l_co2 l_gdp, ///
ytitle("Log of 2010 CO2 emissions (kilotons", margin(0 2 0 0) size(small)) ///
xtitle("Log of 2010 GDP (Bilions of US Dollars)", margin(0 0 0 2) size(small)) ///title("Relationship
between GDP and Co2 Emissions in 2010, by Country",size(medium))
/*You can see that the relationship now looks much clearer. Outliers are greatly reduced */
```



```
reg l_co2 l_gdp
```

```
l_co2  Coef.     Std. Err.   t    P>t   [95% Conf. Interval]

l_gdp 1.01712   .0243813  41.72  0.000  .9690226  1.065218
_cons 5.888671  .1011981  58.19  0.000  5.689035  6.088308
```

*/*The results suggest that a 1% rise in GDP is associated with a 1.02% rise in carbon emissions. In other words, there is a strong linear relationship between the logged variables*/*

(5)

```
predict l_co2_hat, xb  /*fitted values for the logged regression*/
```

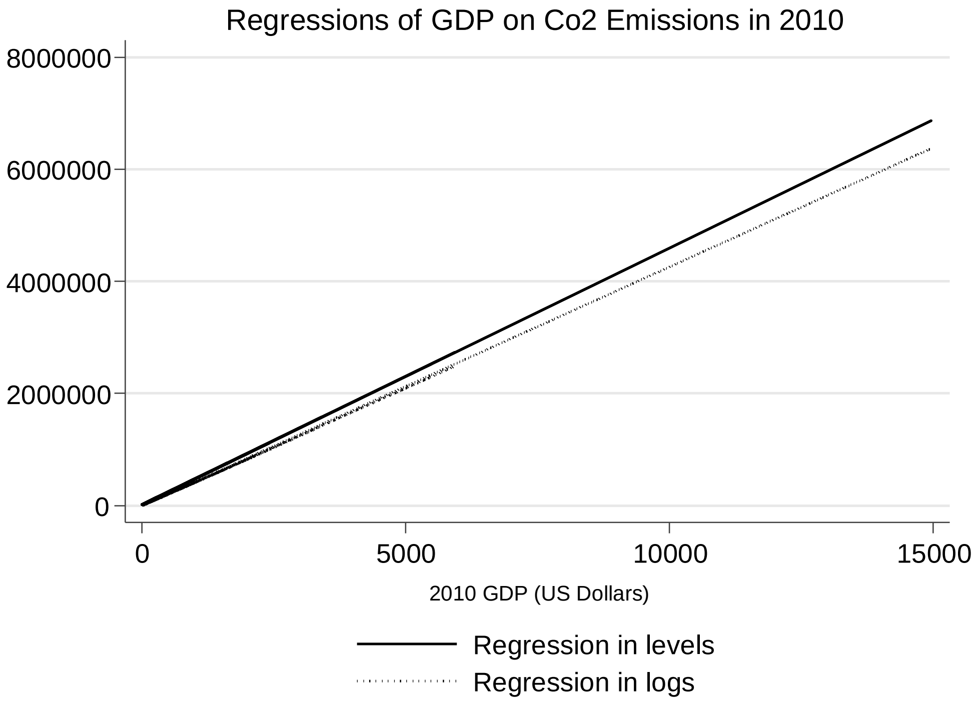
*/*Now we want to convert the logged predictions into levels. Stata's log() function uses the natural log, so we just do e^predicted_values to go back to levels*/*

```
gen l_co2_hat_tr =  exp(2.7182818)
```

*/*Plot both regression lines on the same chart:*/*

```
label variable co2_hat   "Regression in levels"
label variable l_co2_hat_tr   "Regression in logs"
graph twoway (line co2_hat gdp_bill) (line l_co2_hat_tr gdp_bill, lpattern(dot)), ///
```

```
ytitle("2010 CO2 emissions (kilotons", margin(0 2 0 0) size(small)) ///
xtitle("2010 GDP (US Dollars)", margin(0 0 0 2) size(small)) ///
title("Regressions of GDP on Co2 Emissions in 2010",size(medium))
```



*/*We find that the regression in logs leads to slightly lower predicted values. This is because logging reduces variability in the data (i.e., it 'shrinks' outliers); in the logged case, China is no longer exerting such a strong influence in pulling up the regression line.*/*

Part 5

(1)

```
use gdp_co2_pc_2010.dta,clear
gen gdp_th = gdppc_2010/1000
corr gdppc_2010 co2pc_2010,c
```

```
. corr gdp_th co2pc_2010,c
(obs=188)

             |   gdp_th co2~2010
-------------+------------------
      gdp_th |  341.092
  co2pc_2010 |  70.3877  40.8905
```

/ We now that the slope is equal to cov(x,y)/var(x) = 70.387/341.092 = 0.21*/*

(2)

/ The slope implies that an increase in per capita GDP of $1000 is associated with an increase of 0.21 units of per capita Co2 emissions.*/*