Econometrics for Public Policy II
Spring 2025

Problem Set 1
Due before Class 3, September 13

There are two datasets for this problem set: one for 1950 and another for 2010. (You should see links in the previous sentence for data downloading.) If you are using R, use the `haven` package and the `read_dta` function to read these data. If you need another type of dataset, please let me know and I'm happy to provide.

Each dataset has one observation per US county in that year (1950 or 2010). Data come from the Decennial Cenus (1950, some 2010) and the American Community Survey (2010, which is really 2008-2012 5-year average). All variables are labeled. The census tabulates data from the individual collection at a variety of levels of geography; here we use county-level data.

The variables statefips/countyfips uniquely identify observations in each dataset. You can find definitions for the statefips and countyfips variables at this helpful webpage from the University of Missouri. and many other websites.

Use Stata's `describe` command to see the definitions of the remaining variables, or reference the definitions I give below.

Please turn in three documents, **all in pdf form**: (i) a set of written answers to these problems, (ii) a do file (or program from the statistical software of your choice), and and (iii) the output from the program of your choice. For the third, it is sufficient to copy the output into a word doc and save as a pdf. The program file should have comments that indicate the commands associated with each question.

You are both welcome and encouraged to work on this problem set with your classmates. The problem set you turn in should be your own work – both the code and the written output. If we notice exactly duplicative work, we will give zero credit to both assignments.

1. Summary statistics

   a. Make a panel dataset from 1950 and 2010, meaning a dataset that has one observation per county and year. In this dataset, most counties will have two observations, one for 1950 and one for 2010. Stata's `append` command stacks one dataset on top of another.

   b. By year, find the average of

      - population (`cv1`)
      - log of population (create yourself from `cv1`)

- share white (`s1`)
- share black (`s2`)
- share women age 25+ with education of some college or more (`s3`)
- share men age 25+ with education of some college or more (`s4`)

Use one command to find all these averages. In Stata, you can use `collapse` combined with `, by(year)`.

After using `collapse`, you can use `outsheet` to output the resulting dataset as a txt or csv. Using this output file, it should not be difficult to create a labeled table.

Answer: See Table 1 at end.

c. Find averages of the same variables by year and state for California, Mississippi and New Jersey.

Note that California's state code is 6, and that it has a leading zero – so write it `06`. Again, use Stata's `collapse` combined with `, by(year state)` to generate these outputs. Also again, you can use Stata's `outsheet` command to output the data you've created to a .csv or .txt file.

Answer: See Table 2 at end.

2. Matching Data

a. How many counties are in both the 1950 and 2010 datasets?

In the previous question you created a panel dataset. To answer this question, you may prefer to make a "wide" dataset with one observation per county. It may be helpful to make a indicator variable (0/1) for having an observation in a given year in the 1950 and 2010 datasets, and merge the datasets.

In the merged dataset, you can use Stata's `tab` command to see a cross-tab of two indicator variables. For example, if your variables are called `y1950` and `y2010`, you can tell Stata to report `tab y1950 y2010`. Correctly interpreting this table will tell you the answer to the first three parts of this question.

Answer: 3,090

See notes in program and log file about how to find this.

b. How many counties are in the 1950 dataset, but not the 2010 dataset?

Answer: 12

c. How many counties are in the 2010 dataset, but not the 1950 dataset?

Answer: 19

d. Investigate two counties that are in the 2010 dataset, but not the 1950 dataset. Why is this?

Answer: Here are two examples – you answer can include any valid examples. My two examples did not exist in 1950.

- Menominee County, Wisconson (55/078) was created in 1959 (see Wikipedia)
- La Paz County, Arizona (04/012) was established in 1983 (again, see Wikipedia)

3. Regressions

a. Return to the panel dataset from question 1.

b. Regress log of population on the four share variables from question 1 and a fixed effect for year = 2010.

For this and the next question, it is sufficient to paste the results from the log; for future problem sets you will need to make a regression table, but you do not need one here.

In Stata, there are multiple says to create indicator variables and use them in a regression. Here are two equivalent methods:

- ```
  gen y2010 == 0
  replace y2010 = 1 if year == 2010
  regress y x y2010
  ```
- `xi:  regress y x i.y2010`
- You can test for yourself whether these yield equivalent results

Answer: Results are in the log file.

c. Interpret the coefficient on the year indicator variable

Answer: The coefficient on the year indicator I estimated is -0.605, which means that the average county has 0.605 log points lower population in 2010 than in 1950. This negative coefficient may strike you as surprising, but remember that the US has had a big shift to urban areas – so while the biggest counties got bigger, most counties lost population.

Usually when the dependent variable is in logs, we can interpret the coefficient as a percentage change. We can do this because the coefficient tells us that there is a β log point change in the dependent variable for a one-unit change in X. To convert this log
  ```

point change into a regular old change, we do e^{β}, since $\beta = \ln(\Delta Y)$. Exponentiating both sides gives $e^{\beta} = e^{\ln(\Delta Y)}$, or $e^{\beta} = \Delta Y$, where Δ denotes the change.

For small β, $e^{\beta} \sim 1 + \beta$, so we can interpret, for example, $\beta = 0.03$ as a 3% change. However, our change is pretty big! So let's do the math: $e^{-0.606} = 0.546$, or a 55% decline, rather than the 60% decline the naive interpretation of the coefficient would suggest.

For this question, I was hoping that you would interpret the coefficient as an average decline in population of about 60%.

d. Repeat the previous regression with state fixed effects

Answer: Results are in the log file.

e. Interpret one of the share coefficients from the second regression

Answer: First note that the shares and percentages are equivalent. A share of 0.01 is 1 percent. A share of 1 is 100 percent. A one unit change in the share is a change from 0 to 1, which is a change from 0 percentage points to 100 percentage points.

In the regression, the coefficients are -2.06 (share white), -1.02 (share AA), -3.8 (share women at least college) and 9.00 (share men at least college). Using the first one, a one hundred percentage point increase in the share of the white population (a one-unit change) is associated with a 206 percent decrease in a county's population. (Because the dependent variable is in logs, we can interpret the coefficient as a percentage point change.)

However, no counties experience a one hundred percentage point decline in white population share.

4. Long and Short Regressions and Omitted Variable Bias

- From the lecture, we learned the omitted variable bias formula. Now you're going to calculate a specific example.
- We limit our analysis just to 2010.
- We are interested in the impact of the share of college educated men on the employment to population ratio and on the extent of omitted variable bias if we exclude the share of women who are college educated.
  - Let E_i, defined as `cv59 / cv1`, denote the employment to population ratio in county i
  - Let M_i be the share of men age 25 or above who are college educated in county i
  - Let W_i be the share of women age 25 or above who are college educated in county i

- Let's suppose that we have a "true" long equation

$$E_i = \beta_0 + \beta_l M_i + \gamma W_i + \epsilon_{l,i} \tag{1}$$

- However, we sometimes want to be lazy and estimate a "short" regression:

$$E_i = \beta_0 + \beta_s M_i + \epsilon_{s,i} \tag{2}$$

- How bad is the short regression? The omitted variable bias formula tells us that

$$\beta_s - \beta_l = \pi * \gamma \tag{3}$$

  where γ is the coefficient on W from the long regression and π is the coefficient on M from this regression that estimates the strength of the correlation between M and W:

$$W_i = \alpha + \pi M_i + \epsilon_{c,i} \tag{4}$$

a. Estimate equations 1, 2 and 4 above.

b. Use your estimated coefficients to show that the omitted variable bias formula (equation 3) holds. To do so, write the estimates for β_s and β_l, and show that their difference is equal to the product of your estimates of π and γ.
Answer: Results are in the log file.

How to Turn This In

Write a Piazza email, attaching the three items I describe above in pdf form (code, code output, and clearly written questions and answers). **Send these items as a note to "Instructors" and select the folder "hw1_submissions"**
Tables, Code and Output

Table 1: National County Averages by Year

	1950	2010
population	48581	98641
log(population)	9.9	10.3
Share white	0.891	0.843
Share African American	0.101	0.09
Share of women age 25+ with at least some college	0.119	0.515
Share of men age 25+ with at least some college	0.1	0.466

Table 2: State Means by Year

				share			
						at least college, age 25+	
State	year	population	log(pop)	white	African Am.	women	men
CA	1950	182521	10.8	0.951	0.019	0.168	0.157
	2010	642310	12	0.752	0.033	0.613	0.575
MS	1950	26572	10	0.564	0.435	0.086	0.075
	2010	36187	10.1	0.565	0.41	0.477	0.41
NJ	1950	230254	11.9	0.936	0.063	0.106	0.138
	2010	418662	12.7	0.738	0.121	0.575	0.573

```
# delimit;

*************************************************************************

this problem set asks students to do the following things
 - take average by year (stack data)
 - merge data across years (figure out which obs arent consistent across years)
 - collapse by year and state
 - do a simple regression
 - find average change in share educated by state

january 17, 2017
february 14, 2017
january 16, 2018
february 11, 2020 ** update to fix
january 11, 2022
august 30, 2023
august 31, 2023

ps1v04.do

*************************************************************************;

clear all;
pause on;
set more off;

capture log close;
log using ps1.log, replace;

dateo;

***** 0. prepare data for students ***************************************;

*** bring in data ***;

* bring in 1950 data *;
use /home/lfbrooks/pppa6022/2017/problem_sets/stata_basics/data/d1950_20170117;

* append (stack) 2010 data *;
append using
/home/lfbrooks/pppa6022/2017/problem_sets/stata_basics/data/d2010_20170117;

* dataset should be unique by statefips/countyfips/year *;
duplicates report statefips countyfips year, analyze;

*** calculate needed variables ***;

* share white *;
gen s1 = (cv1 - cv3 - cv4) / cv1 ;
```

```
label variable s1 "share white";

* share black *;
gen s2 = cv3 / cv1;
label variable s2 "share black";

* share of women age 25+ college educated *;
gen s3 = (cv25 + cv26)/(cv18 + cv19 + cv20 + cv21 + cv22 + cv23 + cv24 + cv25 +
cv26) if year == 1950;
replace s3 = (cv25 + cv26 + cv27)/(cv18 + cv19 + cv20 + cv21 + cv22 + cv23 + cv24 +
cv25 + cv26 + cv27) if year == 2010;
label variable s3 "share of women age 25+ college educated";

* share of men age 25+ college educated *;
gen s4 = (cv15 + cv16)/(cv8 + cv9 + cv10 + cv11 + cv12 + cv13 + cv14 + cv15 + cv16)
if year == 1950;
replace s4 = (cv15 + cv16 + cv17)/(cv8 + cv9 + cv10 + cv11 + cv12 + cv13 + cv14 +
cv15 + cv16 + cv17) if year == 2010;
label variable s4 "share of men age 25+ college educated";

* keep needed variables *;
keep statefips countyfips year cv1 cv59 s1 s2 s3 s4 d1950 d2010 name;

* preserve so I can save a 1950 and a 2010 *;
preserve;

* save 1950 version *;
keep if year == 1950;
drop d2010;
save /home/lfbrooks/pppa6022/2023/problem_sets/ps1/data/d1950_${date}, replace;

* save 2010 version *;
restore;
keep if year == 2010;
drop name d1950;
save /home/lfbrooks/pppa6022/2023/problem_sets/ps1/data/d2010_${date}, replace;

***** 1. averages by year, and by state and year ***************************;

* find average share white, average share black,
  average share any college or more by gender
  in both years (national)
  and by state (make output a dataset, not just printed to the screen)*;

*** bring in data ***;

* date of data *;
local date_of_data "20230830";

* bring in 1950 data *;
```

```
use /home/lfbrooks/pppa6022/2023/problem_sets/ps1/data/d1950_`date_of_data';

* append (stack) 2010 data *;
append using
/home/lfbrooks/pppa6022/2023/problem_sets/ps1/data/d2010_`date_of_data';

* dataset should be unique by statefips/countyfips/year *;
duplicates report statefips countyfips year, analyze;

*** calculate needed variables ***;

* log of population *;
gen ln_cv1 = log(cv1);
label variable ln_cv1 "log of population (cv1)";

* save for later use *;
save /groups/brooksgrp/junk/uselater, replace;

*** find national averages by year ***;

preserve;
sort year;
collapse (mean) cv1 ln_cv1 s1 s2 s3 s4, by(year);
outsheet using output/natl_lvl_averages.txt, replace;

list;

*** find averages by state and year for CA, MS and ME ***;

restore;
preserve;
keep if statefips == "06" | statefips == "28" | statefips == "34";
table statefips;
sort statefips year;
collapse (mean) cv1 ln_cv1 s1 s2 s3 s4, by(statefips year);
outsheet using output/state_lvl_averages.txt, replace;


****** 3. county change over time
*******************************************************;

* clear all data *;
drop _all;

* bring in 1950 data *;
use /home/lfbrooks/pppa6022/2023/problem_sets/ps1/data/d1950_`date_of_data';
keep statefips countyfips d1950 cv1 name;
rename cv1 cv1_1950;
rename name name_1950;
sort statefips countyfips;
```

```
* merge in 2010 data *;

merge 1:1 statefips countyfips using
  /home/lfbrooks/pppa6022/2023/problem_sets/ps1/data/d2010_`date_of_data';

* fix markers so i can clearly see merging issues *;
replace d1950 = 0 if d1950 == .;
replace d2010 = 0 if d2010 == .;

* look at cross-tab of which counties appear when *;
tab d1950 d2010;

* list the counties that dont merge *;
list statefips countyfips year d1950 d2010 name_1950 name if d1950 + d2010 != 2;


****** 4. regression
**********************************************************************;

* 3(a) *;
*** bring in data from 1 that I saved ***;
drop _all;
use  /groups/brooksgrp/junk/uselater;

* 3(b) *;
* population as a function of our four variables, with and without state fixed
effects *;
regress ln_cv1 s1 s2 s3 s4;
regress ln_cv1 s1 s2 s3 s4 i.year;

* 3(c) *;
* answer in words *;

* 3(d) *;
xi: regress ln_cv1 s1 s2 s3 s4 i.statefips;
xi: regress ln_cv1 s1 s2 s3 s4 i.statefips i.year;

* 3(e) *;
* answer in words *;

* 3(f) *;
* find the standard deviation in share white *;
summ s1 if year == 2010;
* answer in words *;

***** 4. long and short regressions ****************************;

* look at 2010 employment-to-population ratio as a function of male and female
education **;
```

```
-------------------------------------------------------------------------------
-----------------------------------------
      name:  <unnamed>
       log:  /home/lfbrooks/pppa6022/2023/problem_sets/ps1/ps1.log
  log type:  text
 opened on:  30 Aug 2023, 10:44:48

. dateo;

. ***** 0. prepare data for students ****************************************;
. *** bring in data ***;
. * bring in 1950 data *;
. use /home/lfbrooks/pppa6022/2017/problem_sets/stata_basics/data/d1950_20170117;

. * append (stack) 2010 data *;
. append using
/home/lfbrooks/pppa6022/2017/problem_sets/stata_basics/data/d2010_20170117;
(note: variable cv87 was int, now double to accommodate using data's values)
(note: variable cv11 was long, now double to accommodate using data's values)
(note: variable cv12 was long, now double to accommodate using data's values)
(note: variable cv21 was long, now double to accommodate using data's values)
(note: variable cv22 was long, now double to accommodate using data's values)

. * dataset should be unique by statefips/countyfips/year *;
. duplicates report statefips countyfips year, analyze;

Duplicates in terms of statefips countyfips year

--------------------------------------
   copies | observations       surplus
----------+---------------------------
        1 |         6211             0
--------------------------------------

. *** calculate needed variables ***;
. * share white *;
. gen s1 = (cv1 - cv3 - cv4) / cv1 ;

. label variable s1 "share white";

. * share black *;
. gen s2 = cv3 / cv1;

. label variable s2 "share black";

. * share of women age 25+ college educated *;
. gen s3 = (cv25 + cv26)/(cv18 + cv19 + cv20 + cv21 + cv22 + cv23 + cv24 + cv25 +
cv26) if year == 1950;
(3,109 missing values generated)
```

```
. replace s3 = (cv25 + cv26 + cv27)/(cv18 + cv19 + cv20 + cv21 + cv22 + cv23 + cv24
+ cv25 + cv26 + cv27) if year == 2010;
(3,109 real changes made)

. label variable s3 "share of women age 25+ college educated";

. * share of men age 25+ college educated *;
. gen s4 = (cv15 + cv16)/(cv8 + cv9 + cv10 + cv11 + cv12 + cv13 + cv14 + cv15 +
cv16) if year == 1950;
(3,109 missing values generated)

. replace s4 = (cv15 + cv16 + cv17)/(cv8 + cv9 + cv10 + cv11 + cv12 + cv13 + cv14 +
cv15 + cv16 + cv17) if year == 2010;
(3,109 real changes made)

. label variable s4 "share of men age 25+ college educated";

. * keep needed variables *;
. keep statefips countyfips year cv1 cv59 s1 s2 s3 s4 d1950 d2010 name;

. * preserve so I can save a 1950 and a 2010 *;
. preserve;

. * save 1950 version *;
. keep if year == 1950;
(3,109 observations deleted)

. drop d2010;

. save /home/lfbrooks/pppa6022/2023/problem_sets/ps1/data/d1950_${date}, replace;
file /home/lfbrooks/pppa6022/2023/problem_sets/ps1/data/d1950_20230830.dta saved

. * save 2010 version *;
. restore;

. keep if year == 2010;
(3,102 observations deleted)

. drop name d1950;

. save /home/lfbrooks/pppa6022/2023/problem_sets/ps1/data/d2010_${date}, replace;
file /home/lfbrooks/pppa6022/2023/problem_sets/ps1/data/d2010_20230830.dta saved

. ***** 1. averages by year, and by state and year **************************;
. * find average share white, average share black,
>   average share any college or more by gender
>   in both years (national)
>   and by state (make output a dataset, not just printed to the screen)*;
. *** bring in data ***;
. * date of data *;
```

```
. local date_of_data "20230830";

. * bring in 1950 data *;
. use /home/lfbrooks/pppa6022/2023/problem_sets/ps1/data/d1950_`date_of_data';

. * append (stack) 2010 data *;
. append using
/home/lfbrooks/pppa6022/2023/problem_sets/ps1/data/d2010_`date_of_data';

. * dataset should be unique by statefips/countyfips/year *;
. duplicates report statefips countyfips year, analyze;

Duplicates in terms of statefips countyfips year

--------------------------------------
   copies | observations       surplus
----------+---------------------------
        1 |         6211             0
--------------------------------------

. *** calculate needed variables ***;
. * log of population *;
. gen ln_cv1 = log(cv1);

. label variable ln_cv1 "log of population (cv1)";

. * save for later use *;
. save /groups/brooksgrp/junk/uselater, replace;
file /groups/brooksgrp/junk/uselater.dta saved

. *** find national averages by year ***;
. preserve;

. sort year;

. collapse (mean) cv1 ln_cv1 s1 s2 s3 s4, by(year);

. outsheet using output/natl_lvl_averages.txt, replace;

. list;

     +-------------------------------------------------------------------------------+
     | year           cv1     ln_cv1         s1         s2         s3         s4 |
     |-------------------------------------------------------------------------------|
  1. | 1950   48580.70954   9.894452   .8910063   .1009891   .1189899   .1004977 |
  2. | 2010   98641.04407   10.27845    .842963   .0904683   .5153075   .4664086 |
     +-------------------------------------------------------------------------------+

. *** find averages by state and year for CA, MS and ME ***;
. restore;
```

```
. preserve;

. keep if statefips == "06" | statefips == "28" | statefips == "34";
(5,889 observations deleted)

. table statefips;

----------------------
state     |
fips code |      Freq.
----------+-----------
       06 |        116
       28 |        164
       34 |         42
----------------------

. sort statefips year;

. collapse (mean) cv1 ln_cv1 s1 s2 s3 s4, by(statefips year);

. outsheet using output/state_lvl_averages.txt, replace;

. ****** 3. county change over time
*********************************************************;
. * clear all data *;
. drop _all;

. * bring in 1950 data *;
. use /home/lfbrooks/pppa6022/2023/problem_sets/ps1/data/d1950_`date_of_data';

. keep statefips countyfips d1950 cv1 name;

. rename cv1 cv1_1950;

. rename name name_1950;

. sort statefips countyfips;

. * merge in 2010 data *;
. merge 1:1 statefips countyfips using
>   /home/lfbrooks/pppa6022/2023/problem_sets/ps1/data/d2010_`date_of_data';

    Result                           # of obs.
    -----------------------------------------
    not matched                             31
        from master                         12  (_merge==1)
        from using                          19  (_merge==2)

    matched                              3,090  (_merge==3)
```

```
    ----------------------------------------

. * fix markers so i can clearly see merging issues *;
. replace d1950 = 0 if d1950 == .;
(19 real changes made)

. replace d2010 = 0 if d2010 == .;
(12 real changes made)

. * look at cross-tab of which counties appear when *;
. tab d1950 d2010;

 1 if year |   1 if year is 2010
   is 1950 |         0          1 |     Total
-----------+----------------------+----------
         0 |         0         19 |        19
         1 |        12      3,090 |     3,102
-----------+----------------------+----------
     Total |        12      3,109 |     3,121

. * list the counties that dont merge *;
. list statefips countyfips year d1950 d2010 name_1950 name if d1950 + d2010 != 2;

      +-------------------------------------------------------------------------+
      | statef~s   county~s   year   d1950   d2010   nam~1950   nam~1950 |
      |-------------------------------------------------------------------------|
1619. |       30        113      .       1       0                       |
1725. |       32        025      .       1       0                       |
2327. |       46        001      .       1       0                       |
2392. |       46        131      .       1       0                       |
2814. |       51        055      .       1       0                       |
      |-------------------------------------------------------------------------|
2848. |       51        123      .       1       0                       |
2851. |       51        129      .       1       0                       |
2862. |       51        151      .       1       0                       |
2881. |       51        189      .       1       0                       |
2891. |       51        560      .       1       0                       |
      |-------------------------------------------------------------------------|
2908. |       51        785      .       1       0                       |
3102. |       56        047      .       1       0                       |
3103. |       04        012   2010       0       1                       |
3104. |       08        014   2010       0       1                       |
3105. |       32        510   2010       0       1                       |
      |-------------------------------------------------------------------------|
3106. |       35        006   2010       0       1                       |
3107. |       51        515   2010       0       1                       |
3108. |       51        550   2010       0       1                       |
3109. |       51        580   2010       0       1                       |
3110. |       51        595   2010       0       1                       |
      |-------------------------------------------------------------------------|
```

```
3111. |       51        600   2010       0       1                       |
3112. |       51        620   2010       0       1                       |
3113. |       51        640   2010       0       1                       |
3114. |       51        678   2010       0       1                       |
3115. |       51        683   2010       0       1                       |
      |-----------------------------------------------------------------|
3116. |       51        685   2010       0       1                       |
3117. |       51        720   2010       0       1                       |
3118. |       51        735   2010       0       1                       |
3119. |       51        775   2010       0       1                       |
3120. |       51        810   2010       0       1                       |
      |-----------------------------------------------------------------|
3121. |       55        078   2010       0       1                       |
      +-----------------------------------------------------------------+

. ****** 4. regression
**********************************************************************;
. * 3(a) *;
. *** bring in data from 1 that I saved ***;
. drop _all;

. use  /groups/brooksgrp/junk/uselater;

. * 3(b) *;
. * population as a function of our four variables, with and without state fixed
effects *;
. regress ln_cv1 s1 s2 s3 s4;

      Source |       SS           df       MS      Number of obs   =     6,211
-------------+----------------------------------   F(4, 6206)      =    464.29
       Model |  2505.48738         4  626.371845   Prob > F        =    0.0000
    Residual |  8372.51911     6,206  1.34910073   R-squared       =    0.2303
-------------+----------------------------------   Adj R-squared   =    0.2298
       Total |  10878.0065     6,210  1.75169187   Root MSE        =    1.1615

------------------------------------------------------------------------------
      ln_cv1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          s1 |  -1.116101   .2210633    -5.05   0.000    -1.549461   -.6827401
          s2 |   .2454641   .2359306     1.04   0.298    -.2170415    .7079697
          s3 |  -10.61731   .3288104   -32.29   0.000     -11.2619   -9.972731
          s4 |    12.6286   .3421147    36.91   0.000     11.95793    13.29926
       _cons |    10.8182   .2242445    48.24   0.000     10.37861     11.2578
------------------------------------------------------------------------------

. regress ln_cv1 s1 s2 s3 s4 i.year;

      Source |       SS           df       MS      Number of obs   =     6,211
-------------+----------------------------------   F(5, 6205)      =    386.01
       Model |  2580.83455         5   516.16691   Prob > F        =    0.0000
```

```
    Residual |  8297.17193     6,205  1.33717517   R-squared       =    0.2373
-------------+----------------------------------   Adj R-squared   =    0.2366
       Total |  10878.0065     6,210  1.75169187   Root MSE        =    1.1564

------------------------------------------------------------------------------
      ln_cv1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          s1 |  -1.172123   .2202106    -5.32   0.000    -1.603812    -.740434
          s2 |   .3030146   .2350106     1.29   0.197    -.1576875    .7637167
          s3 |  -8.913226   .3983665   -22.37   0.000    -9.694163    -8.13229
          s4 |   12.21275   .3450752    35.39   0.000     11.53628    12.88921
             |
        year |
       2010  |   -.605431   .0806539    -7.51   0.000    -.7635406   -.4473214
             |
       _cons |   10.74145   .2234852    48.06   0.000     10.30334    11.17956
------------------------------------------------------------------------------

. * 3(c) *;
. * answer in words *;
. * 3(d) *;
. xi: regress ln_cv1 s1 s2 s3 s4 i.statefips;
i.statefips       _Istatefips_1-49    (_Istatefips_1 for statefips==01 omitted)

      Source |       SS           df       MS      Number of obs   =     6,211
-------------+----------------------------------   F(52, 6158)     =     93.84
       Model |  4809.09983        52  92.4826891   Prob > F        =    0.0000
    Residual |  6068.90666     6,158  .985532097   R-squared       =    0.4421
-------------+----------------------------------   Adj R-squared   =    0.4374
       Total |  10878.0065     6,210  1.75169187   Root MSE        =    .99274

------------------------------------------------------------------------------
       ln_cv1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
           s1 |  -1.612403   .2085785    -7.73   0.000     -2.02129   -1.203516
           s2 |  -.4506948   .2433753    -1.85   0.064    -.9277954    .0264058
           s3 |  -8.476128    .297724   -28.47   0.000    -9.059772   -7.892485
           s4 |   10.38369   .3094779    33.55   0.000      9.77701    10.99038
 _Istatefips_2 |   .2395945   .2096143     1.14   0.253    -.1713227    .6505117
 _Istatefips_3 |   -.341987   .1191919    -2.87   0.004    -.5756447   -.1083293
 _Istatefips_4 |   .7549761    .132003     5.72   0.000     .4962042    1.013748
 _Istatefips_5 |   -.956634   .1285559    -7.44   0.000    -1.208648   -.7046196
 _Istatefips_6 |   1.661874   .2647189     6.28   0.000     1.142933    2.180816
 _Istatefips_7 |   1.078478   .4146538     2.60   0.009      .265612    1.891345
 _Istatefips_8 |   1.766339   .7078582     2.50   0.013       .37869    3.153989
 _Istatefips_9 |   .1072419    .122085     0.88   0.380    -.1320872    .3465711
_Istatefips_10 |  -.6087009   .1022703    -5.95   0.000    -.8091864   -.4082154
_Istatefips_11 |  -1.111877   .1413167    -7.87   0.000    -1.388907   -.8348467
_Istatefips_12 |     -.0232   .1153071    -0.20   0.841    -.2492422    .2028422
_Istatefips_13 |   .0033265   .1181734     0.03   0.978    -.2283347    .2349876
```

```
_Istatefips_14 |  -.2464388   .1172589    -2.10   0.036    -.4763071   -.0165705
_Istatefips_15 |   -1.01042   .1156064    -8.74   0.000    -1.237049   -.7837915
_Istatefips_16 |  -.1639021   .1121195    -1.46   0.144    -.3836954    .0558913
_Istatefips_17 |  -.0388832   .1227779    -0.32   0.751    -.2795708    .2018044
_Istatefips_18 |   .6963108   .1989496     3.50   0.000     .3063001    1.086322
_Istatefips_19 |    .557126   .1677198     3.32   0.001     .2283366    .8859153
_Istatefips_20 |   1.509039   .2090897     7.22   0.000     1.099151    1.918928
_Istatefips_21 |   .1461311   .1203122     1.21   0.225    -.0897228     .381985
_Istatefips_22 |  -.1091943   .1202632    -0.91   0.364    -.3449521    .1265636
_Istatefips_23 |  -.5141518   .1164799    -4.41   0.000    -.7424931   -.2858106
_Istatefips_24 |  -.2831452   .1131821    -2.50   0.012    -.5050216   -.0612689
_Istatefips_25 |  -1.311575   .1330621    -9.86   0.000    -1.572423   -1.050727
_Istatefips_26 |  -1.289386    .118749   -10.86   0.000    -1.522175   -1.056596
_Istatefips_27 |  -1.412422    .194399    -7.27   0.000    -1.793512   -1.031332
_Istatefips_28 |   .7105157   .2409119     2.95   0.003     .2382441    1.182787
_Istatefips_29 |    1.35665     .17826     7.61   0.000     1.007198    1.706102
_Istatefips_30 |   -.727946   .1563599    -4.66   0.000    -1.034466   -.4214259
_Istatefips_31 |   1.089503   .1280204     8.51   0.000     .8385379    1.340467
_Istatefips_32 |    .230987   .1114315     2.07   0.038     .0125423    .4494317
_Istatefips_33 |  -1.268355   .1348613    -9.40   0.000     -1.53273    -1.00398
_Istatefips_34 |   .6013818   .1186587     5.07   0.000     .3687693    .8339943
_Istatefips_35 |  -.5275815   .1231592    -4.28   0.000    -.7690165   -.2861465
_Istatefips_36 |  -.2262173   .1498855    -1.51   0.131    -.5200452    .0676105
_Istatefips_37 |   .9065174   .1259987     7.19   0.000     .6595159    1.153519
_Istatefips_38 |   .7997704   .3274598     2.44   0.015     .1578348    1.441706
_Istatefips_39 |   .1328684   .1349881     0.98   0.325    -.1317554    .3974921
_Istatefips_40 |  -1.307799   .1293623   -10.11   0.000    -1.561394   -1.054204
_Istatefips_41 |  -.0312807   .1153096    -0.27   0.786    -.2573277    .1947664
_Istatefips_42 |   -.744918   .1003851    -7.42   0.000    -.9417079    -.548128
_Istatefips_43 |  -1.121234   .1606454    -6.98   0.000    -1.436155   -.8063124
_Istatefips_44 |   .1941165   .2100361     0.92   0.355    -.2176277    .6058607
_Istatefips_45 |  -.4409798   .1061472    -4.15   0.000    -.6490654   -.2328942
_Istatefips_46 |  -.0176285   .1464438    -0.12   0.904    -.3047095    .2694525
_Istatefips_47 |   .0094313   .1323346     0.07   0.943    -.2499907    .2688534
_Istatefips_48 |   .2541545   .1249985     2.03   0.042     .0091137    .4991953
_Istatefips_49 |   -.763836   .1728659    -4.42   0.000    -1.102714   -.4249584
         _cons |   11.53488   .2333342    49.44   0.000     11.07746     11.9923
------------------------------------------------------------------------------

. xi: regress ln_cv1 s1 s2 s3 s4 i.statefips i.year;
i.statefips       _Istatefips_1-49    (_Istatefips_1 for statefips==01 omitted)
i.year            _Iyear_1950-2010    (naturally coded; _Iyear_1950 omitted)

      Source |       SS           df       MS      Number of obs   =     6,211
-------------+----------------------------------   F(53, 6157)     =    103.84
       Model |  5134.19898        53  96.8716789   Prob > F        =    0.0000
    Residual |   5743.8075     6,157  .932890613   R-squared       =    0.4720
-------------+----------------------------------   Adj R-squared   =    0.4674
       Total |  10878.0065     6,210  1.75169187   Root MSE        =    .96586
```

```
#-------------------------------------------------------
#
# econometrics ii: problem set 1 of 3
#
# january 14, 2025
#
# ps01v01.R
#
#-------------------------------------------------------

# ---- A. clear and load packages -----------------------

# clear everything
rm(list = ls())

# load packages
library(data.table)
library(haven)

# data paths
data.1950 <-
"H:/pppa6022/2023_fall/problem_sets/ps1/data/d1950_20230830.dta"
data.2010 <-
"H:/pppa6022/2023_fall/problem_sets/ps1/data/d2010_20230830.dta"

# ---- B. load data -------------------------------------

# 1950
d1950 <- read_dta(data.1950)
setDT(d1950)

# 2010
d2010 <- read_dta(data.2010)
setDT(d2010)

# -------------------------------------------------------
# ----- Question 1 ----------------------------------------
# -------------------------------------------------------

# ---- Question 1(a) ------------------------------------

# make a panel dataset
dyears <- data.table::rbindlist(list(d1950,d2010), fill = TRUE)
setDT(dyears)

# ---- Question 1(b) ------------------------------------

dyears[,
       log.cv1 := log(cv1)]

# by year, find average of cv1, log(cv1), s1, s2, s3, s4
dyears.sum <- dyears[,
                     lapply(.SD, function(x){mean(x, na.rm = TRUE)}),
                     .SDcols = c("cv1","log.cv1","s1","s2","s3","s4"),
```

```
                    by = "year"]

print(dyears.sum)

# ---- Question 1(c) ----------------------------------

# find average by state and year
# ca is 06, MS is 28 and NJ is 34
dyears.st.sum <- dyears[,
                        lapply(.SD, function(x){mean(x, na.rm = TRUE)}),
                        .SDcols = c("cv1","log.cv1","s1","s2","s3","s4"),
                        by = c("statefips","year")]

print(dyears.st.sum)

# ------------------------------------------------------
# ----- Question 2 -------------------------------------
# ------------------------------------------------------

# how many counties are
# -- in 1950, not 2010
# -- not 1950, in 2010
# -- appear in both years

# make small datasets and merge to answer this question
d1950p <- d1950[,c("statefips","countyfips","d1950")]
d2010p <- d2010[,c("statefips","countyfips","d2010")]

# merge the two small datasets
mds <- merge(x = d1950p,
             y = d2010p,
             by = c("statefips","countyfips"),
             all = TRUE)
setDT(mds)

# re-code missings to zeros
mds[,
    `:=`
    (d1950 = ifelse(is.na(d1950) == TRUE, yes = 0, no = d1950),
     d2010 = ifelse(is.na(d2010) == TRUE, yes = 0, no = d2010))]

# find types
mds.sum <- mds[,
               .(counties = .N),
               by = c("d1950","d2010")]
print(mds.sum)


# ------------------------------------------------------
# ----- Question 3 -------------------------------------
# ------------------------------------------------------

# make d2010 = 0 if not defined
dyears[,
```

```
       d2010 := ifelse(is.na(d2010) == TRUE, 0, d2010)]

# ---- Question 3(b) -----------------------------------

r1 <- lm(log.cv1 ~ d2010 + s1 + s2 + s3 + s4,
         data = dyears)

print(summary(r1))

# ---- Question 3(d) -----------------------------------

r2 <- lm(log.cv1 ~ d2010 + s1 + s2 + s3 + s4 + statefips,
         data = dyears)

print(summary(r2))

# ------------------------------------------------------
# ----- Question 4 -------------------------------------
# ------------------------------------------------------

# create employment to population ratio
d2010[,
      E := cv59 / cv1]

# ---- Question 4(a) -----------------------------------

# women are s3, men are s4

# Eq 1:
e1 <- lm(E ~ s4 + s3,
         data = d2010)
print(summary(e1))

# Eq 2:
e2 <- lm(E ~ s4,
         data = d2010)
print(summary(e2))

# Eq 4:
e4 <- lm(s3 ~ s4,
         data = d2010)
print(summary(e4))

# ---- Question 4(b) -----------------------------------

# copying coefficients by hand into here
# beta_s - beta_l = pi * gamma

beta_s <- 0.317586
beta_l <- 0.146891
pi <- 0.772901
gamma <- 0.220850

print("beta_s - beta_l")
```

```
beta_s - beta_l

print("pi * gamma")
pi * gamma
```

```
R version 4.4.1 (2024-06-14 ucrt) -- "Race for Your Life"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> source("H:/pppa6022/2025/problem_sets/ps1/ps1v01.R", echo=TRUE)

> #-------------------------------------------------------------
> #
> # econometrics ii: problem set 1 of 3
> #
> # january 14, 2025
> #
> # ps01v01.R
>  .... [TRUNCATED]

> # load packages
> library(data.table)
data.table 1.16.0 using 7 threads (see ?getDTthreads).  Latest news: r-datata
ble.com

> library(haven)

> # data paths
> data.1950 <- "H:/pppa6022/2023_fall/problem_sets/ps1/data/d1950_20230830.dt
a"

> data.2010 <- "H:/pppa6022/2023_fall/problem_sets/ps1/data/d2010_20230830.dt
a"

> # ---- B. load data -------------------------------------
>
> # 1950
> d1950 <- read_dta(data.1950)

> setDT(d1950)

> # 2010
> d2010 <- read_dta(data.2010)

> setDT(d2010)

> # -----------------------------------------------------
> # ----- Question 1 ------------------------------------
> # ------------------------------ .... [TRUNCATED]

> setDT(dyears)

> # ---- Question 1(b) ----------------------------------
>
> dyears[,
```

```
+        log.cv1 := log(cv1)]

> # by year, find average of cv1, log(cv1), s1, s2, s3, s4
> dyears.sum <- dyears[,
+                      lapply(.SD, function(x){mean(x, na.rm = TRU .... [TRUN
CATED]

> print(dyears.sum)
    year      cv1   log.cv1        s1         s2        s3        s4
   <num>    <num>     <num>     <num>      <num>     <num>     <num>
1:  1950 48580.71  9.894452 0.8910063 0.10098907 0.1189899 0.1004977
2:  2010 98641.04 10.278454 0.8429631 0.09046832 0.5153075 0.4664086

> # ---- Question 1(c) ----------------------------------
>
> # find average by state and year
> # ca is 06, MS is 28 and NJ is 34
> dyears.st.sum < .... [TRUNCATED]

> print(dyears.st.sum)
    statefips  year        cv1   log.cv1        s1           s2         s3
s4
       <char> <num>      <num>     <num>     <num>        <num>      <num>
<num>
 1:        01  1950  45697.657 10.388312 0.6675145 0.3309898567 0.07683309 0.
06290232
 2:        04  1950  53541.929 10.270315 0.8243797 0.0222193489 0.15055767 0.
14107540
 3:        05  1950  25460.147  9.891663 0.8155300 0.1839498469 0.06808349 0.
06761742
 4:        06  1950 182521.086 10.759696 0.9506171 0.0192448054 0.16819786 0.
15651813
 5:        08  1950  21033.159  8.957428 0.9917534 0.0032263476 0.16725428 0.
13364969
 6:        09  1950 250910.000 11.972605 0.9820467 0.0171338674 0.12544054 0.
14173709
 7:        10  1950 106028.333 11.287437 0.8375276 0.1593514358 0.10425909 0.
12254167
 8:        11  1950 802178.000 13.595086 0.6455737 0.3500507474 0.22323199 0.
28419062
 9:        12  1950  41362.761  9.791397 0.7488376 0.2489996167 0.10548337 0.
10642360
10:        13  1950  21664.013  9.525608 0.6578370 0.3418942325 0.08467782 0.
06627897
11:        16  1950  13378.114  9.057056 0.9884751 0.0010192552 0.16928118 0.
14752387
12:        17  1950  85413.490 10.263131 0.9790048 0.0206081619 0.10358239 0.
09905886
13:        18  1950  42763.304 10.184294 0.9876170 0.0120904872 0.10131556 0.
10457536
14:        19  1950  26475.485  9.939814 0.9966983 0.0027739611 0.13866256 0.
10411826
15:        20  1950  18145.705  9.317308 0.9846597 0.0139086435 0.14751885 0.
12750077
16:        21  1950  24540.050  9.706174 0.9486827 0.0510783871 0.07796962 0.
06642033
17:        22  1950  41929.938 10.190414 0.6453274 0.3531917098 0.08099276 0.
07243502
18:        23  1950  57110.875 10.693592 0.9971979 0.0009006487 0.12623208 0.
10111506
19:        24  1950  97625.042 10.716330 0.8206011 0.1788297129 0.11483402 0.
12007770
20:        25  1950 335036.714 11.875641 0.9824048 0.0155521049 0.14536744 0.
15752742
```

```
21:        26  1950  76768.265 10.194469 0.9807581 0.0149588731 0.11824390 0.
09712141
22:        27  1950  34281.414  9.899677 0.9909519 0.0007818175 0.14010725 0.
08761293
23:        28  1950  26572.122 10.003431 0.5637197 0.4346092825 0.08567864 0.
07496693
24:        29  1950  34388.287  9.745118 0.9749855 0.0247454993 0.09666413 0.
08317939
25:        30  1950  10368.842  8.721889 0.9675630 0.0013437682 0.19406769 0.
12244907
26:        31  1950  14252.796  8.985516 0.9940893 0.0016884960 0.13205410 0.
09373548
27:        32  1950   9416.647  8.335317 0.9253942 0.0117230165 0.17250521 0.
15115036
28:        33  1950  53324.200 10.690068 0.9983881 0.0012041007 0.14523033 0.
13354900
29:        34  1950 230253.762 11.901110 0.9358877 0.0626888428 0.10622251 0.
13791879
30:        35  1950  21287.094  9.584667 0.9310442 0.0083468526 0.14009335 0.
13268845
31:        36  1950 239196.645 11.378953 0.9794752 0.0186808373 0.13489155 0.
13290118
32:        37  1950  40619.290 10.291156 0.7330219 0.2601333694 0.10261370 0.
07861946
33:        38  1950  11691.245  9.115934 0.9771115 0.0003605596 0.14706168 0.
09089487
34:        39  1950  90302.580 10.746475 0.9740519 0.0255674540 0.10452846 0.
10393450
35:        40  1950  29004.558  9.894147 0.9192986 0.0497647799 0.12336741 0.
11235643
36:        41  1950  42259.472  9.931195 0.9868103 0.0020258131 0.16672170 0.
13857049
37:        42  1950 156686.746 11.201141 0.9832335 0.0164644437 0.09880657 0.
10191069
38:        44  1950 158379.200 11.323776 0.9815226 0.0165957107 0.12487487 0.
14649727
39:        45  1950  46022.326 10.485139 0.5460489 0.4531223014 0.10659796 0.
08473043
40:        46  1950   9599.118  8.818552 0.9162242 0.0008204528 0.16266163 0.
09660700
41:        47  1950  34649.663  9.943050 0.9027344 0.0970398866 0.07093708 0.
06132804
42:        48  1950  30359.031  9.488845 0.8997961 0.0996547326 0.12533567 0.
11145467
43:        49  1950  23753.862  9.041635 0.9766023 0.0013029992 0.17386874 0.
17461763
44:        50  1950  26981.929  9.971937 0.9987343 0.0009841553 0.15254648 0.
11416354
45:        51  1950  26131.339  9.767200 0.7525941 0.2461513355 0.12547731 0.
10170689
46:        53  1950  60999.051 10.111218 0.9787415 0.0059768839 0.17066479 0.
14830325
47:        54  1950  36464.582 10.131046 0.9658604 0.0339343342 0.08786474 0.
08138517
48:        55  1950  48374.296 10.272613 0.9902752 0.0016169929 0.12680407 0.
09185237
49:        56  1950  12105.375  9.014432 0.9866761 0.0038670217 0.22437836 0.
16409135
50:        01  2010  71339.343 10.617986 0.6826625 0.2842627266 0.45205584 0.
40646737
51:        04  2010 426134.467 11.729611 0.7308264 0.0187691029 0.54642924 0.
52423678
52:        05  2010  38878.907 10.075687 0.7950857 0.1607631904 0.44166252 0.
38063936
```

```
53:        06  2010 642309.586 12.028139 0.7521886 0.0327303708 0.61324170 0.
57547000
54:        08  2010  78581.188  9.752102 0.8999102 0.0157306582 0.64690719 0.
58984001
55:        09  2010 446762.125 12.676402 0.8410686 0.0654977699 0.62141187 0.
58659285
56:        10  2010 299311.333 12.461821 0.7205420 0.2022584776 0.54528697 0.
52302735
57:        11  2010 601723.000 13.307552 0.3911717 0.5139623880 0.68489724 0.
68450892
58:        12  2010 280616.567 11.546377 0.7931614 0.1445937485 0.51530959 0.
46939916
59:        13  2010  60928.635 10.185328 0.6685962 0.2826094516 0.44959590 0.
39669291
60:        16  2010  35626.864  9.644912 0.9284080 0.0033466445 0.55067939 0.
53154062
61:        17  2010 125790.510 10.407480 0.9088049 0.0504928378 0.54126240 0.
49697140
62:        18  2010  70476.109 10.596420 0.9399534 0.0251067275 0.46421510 0.
43690816
63:        19  2010  30771.263  9.817908 0.9539552 0.0108657320 0.54735255 0.
48786584
64:        20  2010  27172.552  9.152584 0.9250112 0.0181830928 0.58079177 0.
51657183
65:        21  2010  36161.392  9.972578 0.9373957 0.0374206849 0.43401717 0.
35322500
66:        22  2010  70833.938 10.570677 0.6470182 0.3172600771 0.43419798 0.
36304328
67:        23  2010  83022.562 11.030065 0.9594105 0.0068331458 0.56261819 0.
50159801
68:        24  2010 240564.667 11.721885 0.7288400 0.1997083166 0.58591667 0.
54084187
69:        25  2010 467687.786 12.411849 0.8512564 0.0565535860 0.65394322 0.
60942622
70:        26  2010 119080.000 10.747791 0.9104921 0.0392844387 0.53547967 0.
49950534
71:        27  2010  60964.655 10.134909 0.9291702 0.0129051826 0.58063604 0.
52820917
72:        28  2010  36186.549 10.103454 0.5652806 0.4103268474 0.47714137 0.
41013774
73:        29  2010  52077.626 10.025186 0.9307597 0.0349667583 0.46338318 0.
40928609
74:        30  2010  17668.125  8.854630 0.8908092 0.0033174800 0.58645555 0.
52780129
75:        31  2010  19638.075  8.714136 0.9517046 0.0072362931 0.58365989 0.
52421188
76:        32  2010 158855.941  9.858459 0.8646370 0.0200078356 0.54117767 0.
51322618
77:        33  2010 131647.000 11.463352 0.9540404 0.0080624546 0.61464202 0.
56520862
78:        34  2010 418661.619 12.708203 0.7381942 0.1210667229 0.57539905 0.
57337725
79:        35  2010  62399.364 10.113584 0.7615998 0.0139211765 0.53002190 0.
49337286
80:        36  2010 312550.032 11.734789 0.8624579 0.0619746435 0.56249246 0.
52027533
81:        37  2010  95354.830 10.889573 0.7220023 0.2058400095 0.53758010 0.
46371740
82:        38  2010  12690.396  8.612037 0.9085952 0.0041324303 0.58591617 0.
53029856
83:        39  2010 131096.636 11.174915 0.9258888 0.0410519423 0.46891495 0.
43565610
84:        40  2010  48718.844 10.011326 0.7716325 0.0346743788 0.49261890 0.
45358639
```

```
85:        41  2010 106418.722 10.582472 0.8964359 0.0068550922 0.58987685 0.
56926283
86:        42  2010 189587.746 11.469826 0.9183383 0.0435406324 0.44903290 0.
43016168
87:        44  2010 210513.400 11.852633 0.8954383 0.0323337601 0.63593308 0.
62435424
88:        45  2010 100551.391 11.010357 0.5964981 0.3614539392 0.48612418 0.
43467645
89:        46  2010  12336.061  8.716507 0.8274787 0.0041615919 0.56355612 0.
48735563
90:        47  2010  66801.105 10.446955 0.8939731 0.0724246249 0.41804853 0.
37213112
91:        48  2010  98998.272  9.871143 0.8501636 0.0644352548 0.47845516 0.
44648444
92:        49  2010  95306.379 10.033336 0.9207001 0.0048743037 0.61428488 0.
59487461
93:        50  2010  44695.786 10.427185 0.9603398 0.0067392826 0.60730523 0.
51605824
94:        51  2010  59709.134 10.273523 0.7574385 0.1923639480 0.53210603 0.
48178793
95:        53  2010 172424.103 10.953999 0.8470212 0.0123878645 0.61492294 0.
58761997
96:        54  2010  33690.800 10.078839 0.9567417 0.0219574150 0.40386589 0.
34374636
97:        55  2010  78985.917 10.643134 0.9240841 0.0151321462 0.53971975 0.
49355419
98:        56  2010  24505.478  9.737264 0.9313790 0.0043544397 0.63170972 0.
56433788
    statefips  year        cv1   log.cv1        s1           s2         s3
s4

> # ------------------------------------------------------
> # ----- Question 2 -------------------------------------
> # ------------------------------ .... [TRUNCATED]

> d2010p <- d2010[,c("statefips","countyfips","d2010")]

> # merge the two small datasets
> mds <- merge(x = d1950p,
+              y = d2010p,
+              by = c("statefips","countyfips"),
+              .... [TRUNCATED]

> setDT(mds)

> # re-code missings to zeros
> mds[,
+     `:=`
+     (d1950 = ifelse(is.na(d1950) == TRUE, yes = 0, no = d1950),
+      d2010 = ifelse(is.na(d2010)  .... [TRUNCATED]

> # find types
> mds.sum <- mds[,
+                .(counties = .N),
+                by = c("d1950","d2010")]

> print(mds.sum)
   d1950 d2010 counties
   <num> <num>    <int>
1:     1     1     3090
2:     0     1       19
3:     1     0       12

> # ------------------------------------------------------
```

```
> # ----- Question 3 -----------------------------------
> # ------------------------------ .... [TRUNCATED]

> # ---- Question 3(b) ---------------------------------
>
> r1 <- lm(log.cv1 ~ d2010 + s1 + s2 + s3 + s4,
+          data = dyears)

> print(summary(r1))

Call:
lm(formula = log.cv1 ~ d2010 + s1 + s2 + s3 + s4, data = dyears)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5580 -0.6563  0.0414  0.7055  4.7814

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.74145    0.22349  48.063  < 2e-16 ***
d2010       -0.60543    0.08065  -7.507 6.93e-14 ***
s1          -1.17212    0.22021  -5.323 1.06e-07 ***
s2           0.30301    0.23501   1.289    0.197
s3          -8.91323    0.39837 -22.374  < 2e-16 ***
s4          12.21275    0.34508  35.392  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.156 on 6205 degrees of freedom
Multiple R-squared:  0.2373,   Adjusted R-squared:  0.2366
F-statistic:   386 on 5 and 6205 DF,  p-value: < 2.2e-16


> # ---- Question 3(d) ---------------------------------
>
> r2 <- lm(log.cv1 ~ d2010 + s1 + s2 + s3 + s4 + statefips,
+          data = dyears)

> print(summary(r2))

Call:
lm(formula = log.cv1 ~ d2010 + s1 + s2 + s3 + s4 + statefips,
    data = dyears)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5104 -0.5499  0.0106  0.5744  4.6101

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.858702   0.227679  52.085  < 2e-16 ***
d2010       -1.504120   0.080573 -18.668  < 2e-16 ***
s1          -2.061972   0.204355 -10.090  < 2e-16 ***
s2          -1.015039   0.238708  -4.252 2.15e-05 ***
s3          -3.828948   0.381938 -10.025  < 2e-16 ***
s4           9.004915   0.310026  29.046  < 2e-16 ***
statefips04 -0.129305   0.204894  -0.631 0.528010
statefips05 -0.329858   0.115967  -2.844 0.004464 **
statefips06  0.269523   0.131036   2.057 0.039741 *
statefips08 -1.491355   0.128313 -11.623  < 2e-16 ***
statefips09  1.289960   0.258321   4.994 6.09e-07 ***
statefips10  0.894751   0.403548   2.217 0.026645 *
statefips11  1.229112   0.689295   1.783 0.074612 .
statefips12 -0.052345   0.119087  -0.440 0.660276
```

```
statefips13 -0.628350   0.099507  -6.315 2.90e-10 ***
statefips16 -1.456276   0.138723 -10.498  < 2e-16 ***
statefips17 -0.238008   0.112774  -2.110 0.034857 *
statefips18 -0.065421   0.115033  -0.569 0.569569
statefips19 -0.561787   0.115328  -4.871 1.14e-06 ***
statefips20 -1.392142   0.114320 -12.178  < 2e-16 ***
statefips21 -0.187081   0.109091  -1.715 0.086413 .
statefips22 -0.027843   0.119455  -0.233 0.815708
statefips23  0.381443   0.194297   1.963 0.049668 *
statefips24  0.267840   0.163913   1.634 0.102303
statefips25  1.041260   0.204966   5.080 3.88e-07 ***
statefips26 -0.092880   0.117753  -0.789 0.430278
statefips27 -0.495652   0.118824  -4.171 3.07e-05 ***
statefips28 -0.567053   0.113362  -5.002 5.83e-07 ***
statefips29 -0.371497   0.110219  -3.371 0.000755 ***
statefips30 -1.830542   0.132411 -13.825  < 2e-16 ***
statefips31 -1.658266   0.117212 -14.148  < 2e-16 ***
statefips32 -1.773152   0.190120  -9.326  < 2e-16 ***
statefips33  0.296327   0.235437   1.259 0.208214
statefips34  1.119596   0.173898   6.438 1.30e-10 ***
statefips35 -1.037524   0.153028  -6.780 1.31e-11 ***
statefips36  0.784394   0.125622   6.244 4.55e-10 ***
statefips37  0.004109   0.109094   0.038 0.969959
statefips38 -1.689869   0.133139 -12.693  < 2e-16 ***
statefips39  0.515035   0.115539   4.458 8.43e-06 ***
statefips40 -0.736568   0.120347  -6.120 9.91e-10 ***
statefips41 -0.642202   0.147520  -4.353 1.36e-05 ***
statefips42  0.872980   0.122601   7.121 1.20e-12 ***
statefips44  0.427336   0.319218   1.339 0.180719
statefips45  0.028613   0.131452   0.218 0.827695
statefips46 -1.772112   0.128294 -13.813  < 2e-16 ***
statefips47  0.011266   0.112211   0.100 0.920028
statefips48 -0.895341   0.097999  -9.136  < 2e-16 ***
statefips49 -1.565636   0.158099  -9.903  < 2e-16 ***
statefips50 -0.266214   0.205832  -1.293 0.195937
statefips51 -0.673794   0.104024  -6.477 1.01e-10 ***
statefips53 -0.491853   0.144726  -3.399 0.000682 ***
statefips54  0.036111   0.128760   0.280 0.779140
statefips55 -0.027310   0.122545  -0.223 0.823652
statefips56 -1.390629   0.171505  -8.108 6.14e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9659 on 6157 degrees of freedom
Multiple R-squared:  0.472,    Adjusted R-squared:  0.4674
F-statistic: 103.8 on 53 and 6157 DF,  p-value: < 2.2e-16


> # ------------------------------------------------------
> # ----- Question 4 -------------------------------------
> # ------------------------------ .... [TRUNCATED]

> # ---- Question 4(a) ----------------------------------
>
> # women are s3, men are s4
>
> # Eq 1:
> e1 <- lm(E ~ s4 + s3,
+          data = d20 .... [TRUNCATED]

> print(summary(e1))

Call:
lm(formula = E ~ s4 + s3, data = d2010)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.32757 -0.03017  0.00299  0.03264  0.39698

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.253717   0.004892  51.858  < 2e-16 ***
s4          0.146891   0.018034   8.145 5.42e-16 ***
s3          0.220850   0.021195  10.420  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04992 on 3106 degrees of freedom
Multiple R-squared:  0.3774,   Adjusted R-squared:  0.377
F-statistic: 941.4 on 2 and 3106 DF,  p-value: < 2.2e-16


> # Eq 2:
> e2 <- lm(E ~ s4,
+          data = d2010)

> print(summary(e2))

Call:
lm(formula = E ~ s4, data = d2010)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34798 -0.03156  0.00087  0.03303  0.48428

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.287909   0.003691   78.00   <2e-16 ***
s4          0.317586   0.007669   41.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05078 on 3107 degrees of freedom
Multiple R-squared:  0.3556,   Adjusted R-squared:  0.3554
F-statistic:  1715 on 1 and 3107 DF,  p-value: < 2.2e-16


> # Eq 4:
> e4 <- lm(s3 ~ s4,
+          data = d2010)

> print(summary(e4))

Call:
lm(formula = s3 ~ s4, data = d2010)

Residuals:
     Min       1Q   Median       3Q      Max
-0.31339 -0.02520 -0.00174  0.02213  0.39532

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.154820   0.003072    50.4   <2e-16 ***
s4          0.772901   0.006382   121.1   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04226 on 3107 degrees of freedom
```

```
Multiple R-squared:  0.8252,   Adjusted R-squared:  0.8251
F-statistic: 1.467e+04 on 1 and 3107 DF,  p-value: < 2.2e-16


> # ---- Question 4(b) -----------------------------------
>
> # copying coefficients by hand into here
> # beta_s - beta_l = pi * gamma
>
> beta_s  .... [TRUNCATED]

> beta_l <- 0.146891

> pi <- 0.772901

> gamma <- 0.220850

> print("beta_s - beta_l")
[1] "beta_s - beta_l"

> beta_s - beta_l
[1] 0.170695

> print("pi * gamma")
[1] "pi * gamma"

> pi * gamma
[1] 0.1706952
```