Stata Summer Series

Stata 301 – Regression Analysis in Stata

What you will get out of this session:

- How do I run a simple regression in Stata and read the results?
- How can I output the regression results outside of Stata?
- What are some other regression models similar to the linear `regress`?

Remember: Regressions are a powerful tool. Try to understand the statistical assumptions that go into your regressions to ensure that your analysis is meaningful and accurate. Even if the program runs without errors or crashing, your analysis could be wrongly suited to your data. You should be able to understand every component of the regression output in the results window.

Regression commands

- `Regress`
- `Logit`
- `Probit`
- `Mlogit`
- `Ologit`
- `tobit`
- `xi: glm`
- `streg`

Helpful resources for regressions

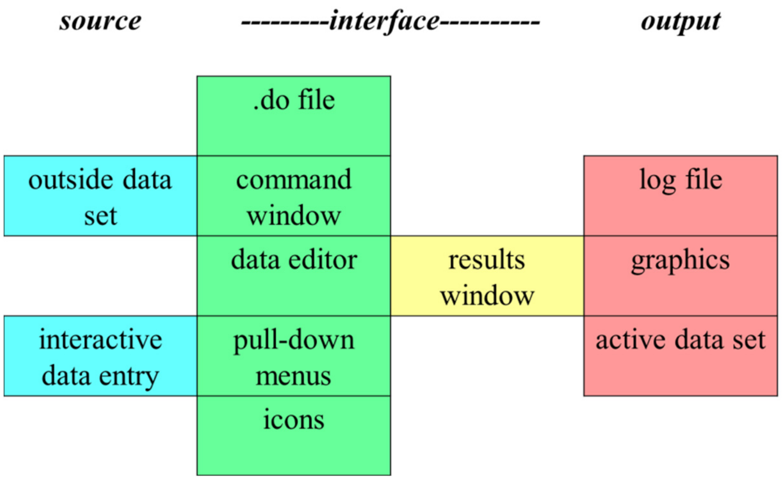
- Stata manual:
  - https://www.stata.com/manuals13/rregress.pdf
  - https://www.stata.com/manuals13/rlogit.pdf
  - Type "help command" for more manuals or google, e.g., "probit stata"
- UCLA IDRE Regression Handbook:
  - https://stats.idre.ucla.edu/stata/webbooks/reg/chapter1/regressionwith-statachapter-1-simple-and-multiple-regression/

Last class in the summer series:

- Stata 302 – **Additional Topics in Regression Analysis – Time Series** (Friday, August 10, 2018 12:00 pm-1:30 pm, 6A)

Review of STATA basics

STATA as a conceptual map:



A few basic commands for getting data from a source:

`use` reads data from a data file that has been created by STATA

```
use "\\Stata2\PopulationProjections\2015\Pop_Baseline\Pop2000", clear
```

(or click on the OPEN (use) icon in the main interface)

`infix` and `infile` read data from an ascii (.txt) file.

```
infile caseid edyears using D:\StataDemos\example01dat.txt
infix caseid 1-2 edyears 3-5 using D:\StataDemos\example01dat.txt
```

`import` (among other commands) reads data from an external spreadsheet

```
import delimited "D:\Martin_UI\STATA users group\2015 CHR Analytic Data (2).csv"
```

`input` reads data that you write in the .do file.

`clear` removes any currently active data file to make room for new ones.

Things you put into the interface, in addition to statistical commands

```
rename v261 edyears
generate edmonths = edyears*12
egen edmean = mean(edyears)
sort FIPS
by FIPS: egen c_births_p = total(pop2010*(birthrate))
replace c_births_p = . if (gender=="m" | (age<10 | age>45))
```

Commonly used relational and logical operators:

```
==              ~=
>       >=      <       <=
&               |
```

Note that == is a logical test, while = is an assignment

```
* file stataclass08032018.do
* STATA commands for getting to know regressions and other simple analyses
* created for the 3rd STATA intro class at Urban Institute, 08/03/2018
* by Smartin, with thanks to Ekalish and Dhanson

* log the results if you wish
* log using "D:\Martin_UI\STATA users group\stataclass0803.log", replace

* first, import a data set:
* the Robert Wood Johnson 2015 County Health Rankings Analytic Data
import delimited "D:\Martin_UI\STATA users group\2015 CHR Analytic Data (2).csv"

* take a look at what we have
summarize

* county code 0 refers to US states, so drop those
drop if countycode==0

. * start with a simple two variable Ordinary Least Squares regression
. * does higher air pollution predict more low birthweight babies?
. regress lowbirthweightvalue airpollutionparticulatematterval

      Source |       SS           df       MS      Number of obs   =     3,016
-------------+----------------------------------   F(1, 3014)      =    167.75
       Model |  .071302809         1  .071302809   Prob > F        =    0.0000
    Residual |  1.28107534     3,014  .000425042   R-squared       =    0.0527
-------------+----------------------------------   Adj R-squared   =    0.0524
       Total |  1.35237814     3,015   .00044855   Root MSE        =    .02062

------------------------------------------------------------------------------------------------
          lowbirthweightvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------------------+-----------------------------------------------------------------
airpollutionparticulatematte~l |   .0031875   .0002461    12.95   0.000      .002705    .0036701
                        _cons |   .0451406   .0028905    15.62   0.000     .0394731    .0508082
------------------------------------------------------------------------------------------------

.
. * maybe we should weight this by the population of each county
. regress lowbirthweightvalue airpollutionparticulatematterval /*
> */ [w = populationestimatevalue]
(analytic weights assumed)
(sum of wgt is 313,785,289)

      Source |       SS           df       MS      Number of obs   =     3,016
-------------+----------------------------------   F(1, 3014)      =    448.72
       Model |  .088356855         1  .088356855   Prob > F        =    0.0000
    Residual |   .59347899     3,014  .000196907   R-squared       =    0.1296
-------------+----------------------------------   Adj R-squared   =    0.1293
       Total |  .681835845     3,015  .000226148   Root MSE        =    .01403

------------------------------------------------------------------------------------------------
          lowbirthweightvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------------------+-----------------------------------------------------------------
airpollutionparticulatematte~l |   .0031133    .000147    21.18   0.000     .0028251    .0034014
                        _cons |   .0460358   .0016656    27.64   0.000     .0427701    .0493016
------------------------------------------------------------------------------------------------

.
. * there are at least three big concerns with regressions which we will mention
. * at different points in this activity
.
. * CONCERN 1: unequally influential observations
. * OLS assumes extreme values are very rare, and it gets squirrely when it sees them
. * so let's try a standard robust variance estimator
.
. regress lowbirthweightvalue airpollutionparticulatematterval /*
```

```
> */ [w = populationestimatevalue], vce(robust)
(analytic weights assumed)
(sum of wgt is 313,785,289)

Linear regression                               Number of obs     =      3,016
                                                F(1, 3014)        =      94.85
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1296
                                                Root MSE          =     .01403

------------------------------------------------------------------------------------------------
                               |               Robust
           lowbirthweightvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------------------+----------------------------------------------------------------
airpollutionparticulatematte~l |   .0031133   .0003197     9.74   0.000     .0024865      .00374
                         _cons |   .0460358   .0037245    12.36   0.000      .038733    .0533387
------------------------------------------------------------------------------------------------

.
. * we will keep this as our simplest regression finding
. est store simple

.
. * maybe water pollution is what we should worry about instead of air pollution?
. regress lowbirthweightvalue airpollutionparticulatematterval /*
> */ drinkingwaterviolationsvalue /*
> */ [w = populationestimatevalue], vce(robust)
(analytic weights assumed)
(sum of wgt is 302,225,771)

Linear regression                               Number of obs     =      2,966
                                                F(2, 2963)        =      46.38
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1340
                                                Root MSE          =     .01393

------------------------------------------------------------------------------------------------
                               |               Robust
           lowbirthweightvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------------------+----------------------------------------------------------------
airpollutionparticulatematte~l |   .0030709   .0003193     9.62   0.000     .0024448     .003697
  drinkingwaterviolationsvalue |   .0063103   .0046743     1.35   0.177    -.0028549    .0154755
                         _cons |   .0457851   .0037668    12.15   0.000     .0383993    .0531709
------------------------------------------------------------------------------------------------

.
. * no need to keep this finding - it was a dead end
.
. * maybe the counties with the most air pollution are simply the poorest
. regress lowbirthweightvalue airpollutionparticulatematterval /*
> */ childreninpovertyvalue  medianhouseholdincomevalue /*
> */ [w = populationestimatevalue], vce(robust)
(analytic weights assumed)
(sum of wgt is 313,785,289)

Linear regression                               Number of obs     =      3,016
                                                F(3, 3012)        =     165.15
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4230
                                                Root MSE          =     .01143

------------------------------------------------------------------------------------------------
                               |               Robust
           lowbirthweightvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------------------+----------------------------------------------------------------
airpollutionparticulatematte~l |   .0035261   .0003411    10.34   0.000     .0028572     .004195
```

```
        childreninpovertyvalue |   .1363497   .0099138    13.75   0.000     .1169112    .1557882
    medianhouseholdincomevalue |   3.00e-07   5.33e-08     5.64   0.000     1.96e-07    4.05e-07
                         _cons |  -.0053713   .0062129    -0.86   0.387    -.0175533    .0068108
------------------------------------------------------------------------------------------------

.
. * keep this finding - people might ask about this
. est store plusincome

.
. * maybe the counties with the most air pollution have poor access to health care
. regress lowbirthweightvalue airpollutionparticulatematterval /*
> */ childreninpovertyvalue  medianhouseholdincomevalue /*
> */ primarycarephysiciansvalue couldnotseedoctorduetocostvalue /*
> */ [w = populationestimatevalue], vce(robust)
(analytic weights assumed)
(sum of wgt is 303,319,296)

Linear regression                               Number of obs     =      2,291
                                                F(5, 2285)        =     100.88
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4439
                                                Root MSE          =     .01112

------------------------------------------------------------------------------------------------
                               |               Robust
           lowbirthweightvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------------------+----------------------------------------------------------------
airpollutionparticulatematte~l |   .0037637    .000337    11.17   0.000     .0031028    .0044245
        childreninpovertyvalue |   .1175812   .0113366    10.37   0.000       .09535    .1398124
    medianhouseholdincomevalue |   2.47e-07   5.62e-08     4.39   0.000     1.36e-07    3.57e-07
    primarycarephysiciansvalue |   .0000598    .000014     4.27   0.000     .0000324    .0000873
couldnotseedoctorduetocostva~e |   .0460961   .0141486     3.26   0.001     .0183506    .0738416
                         _cons |  -.0118681   .0061682    -1.92   0.054    -.0239638    .0002277
------------------------------------------------------------------------------------------------

.
. * let's keep this finding too
. est store plushealthcare

.
. * maybe the counties with the most air pollution also have high black populations
. regress lowbirthweightvalue airpollutionparticulatematterval /*
> */ childreninpovertyvalue  medianhouseholdincomevalue /*
> */ primarycarephysiciansvalue couldnotseedoctorduetocostvalue /*
> */ percentofpopulationthatisnonhisp /*
> */ [w = populationestimatevalue], vce(robust)
(analytic weights assumed)
(sum of wgt is 303,319,296)

Linear regression                               Number of obs     =      2,291
                                                F(6, 2284)        =     169.39
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6677
                                                Root MSE          =      .0086

------------------------------------------------------------------------------------------------
                               |               Robust
           lowbirthweightvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------------------+----------------------------------------------------------------
airpollutionparticulatematte~l |   .0021062   .0002595     8.12   0.000     .0015974    .0026151
        childreninpovertyvalue |   .0330439   .0094177     3.51   0.000     .0145759     .051512
    medianhouseholdincomevalue |   6.58e-09   4.30e-08     0.15   0.878    -7.77e-08    9.08e-08
    primarycarephysiciansvalue |   .0000243    .000012     2.02   0.043     7.52e-07    .0000479
couldnotseedoctorduetocostva~e |    .052408   .0116247     4.51   0.000      .029612     .075204
percentofpopulationthatisnon~p |   .0681445   .0039085    17.44   0.000       .06048     .075809
```

```
                        _cons |    .031799   .0048979     6.49   0.000     .0221943    .0414038
------------------------------------------------------------------------------------------------

.
. * keep this
. est store plusrace

.
. * maybe the counties with the most air pollution also have systematic differences in behavior
. regress lowbirthweightvalue airpollutionparticulatematterval /*
> */ childreninpovertyvalue  medianhouseholdincomevalue /*
> */ primarycarephysiciansvalue couldnotseedoctorduetocostvalue /*
> */ percentofpopulationthatisnonhisp /*
> */ [w = populationestimatevalue], vce(robust)
(analytic weights assumed)
(sum of wgt is 303,319,296)

Linear regression                               Number of obs     =      2,291
                                                F(6, 2284)        =     169.39
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6677
                                                Root MSE          =      .0086

------------------------------------------------------------------------------------------------
                               |               Robust
           lowbirthweightvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------------------+----------------------------------------------------------------
airpollutionparticulatematte~l |   .0021062   .0002595     8.12   0.000     .0015974    .0026151
        childreninpovertyvalue |   .0330439   .0094177     3.51   0.000     .0145759     .051512
    medianhouseholdincomevalue |   6.58e-09   4.30e-08     0.15   0.878    -7.77e-08    9.08e-08
    primarycarephysiciansvalue |   .0000243    .000012     2.02   0.043     7.52e-07    .0000479
couldnotseedoctorduetocostva~e |    .052408   .0116247     4.51   0.000      .029612     .075204
percentofpopulationthatisnon~p |   .0681445   .0039085    17.44   0.000       .06048     .075809
                         _cons |    .031799   .0048979     6.49   0.000     .0221943    .0414038
------------------------------------------------------------------------------------------------

.
. * definitely keep this
. est store plusbehav

.
.
. * make a table of the analyses we have built up
. outreg2 [simple plusincome plushealthcare plusrace plusbehav] using myfile, replace see
Hit Enter to continue.
dir : seeout

.
. * this is too wide, so here is a simpler table of the analyses we have built up
. outreg2 [simple plushealthcare plusbehav] using myfile, replace see
Hit Enter to continue.
dir : seeout

.
. * examine the predicted levels of low birth weight at county pollution extremes
. * (net of county income, health services, and demographics)
. margins, at(airpollutionparticulatematterval=(7 14)) atmeans vsquish

Adjusted predictions                            Number of obs     =      2,291
Model VCE    : Robust

Expression   : Linear prediction, predict()
1._at        : airpolluti~l    =           7
               childreni~ue    =    .2225566 (mean)
               medianhous~e    =    54884.95 (mean)
               primarycar~e    =    75.48096 (mean)
```

```
               couldnotse~e    =    .1424128 (mean)
               percentof~sp    =    .1267389 (mean)
2._at        : airpolluti~l    =          14
               childreni~ue    =    .2225566 (mean)
               medianhous~e    =    54884.95 (mean)
               primarycar~e    =    75.48096 (mean)
               couldnotse~e    =    .1424128 (mean)
               percentof~sp    =    .1267389 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .0721949   .0013113    55.05   0.000     .0696234    .0747664
          2  |   .0869385   .0006569   132.35   0.000     .0856504    .0882266
------------------------------------------------------------------------------

.
. * and why not graph the predicted relationship?
. predict plowbirthweightvalue
(option xb assumed; fitted values)
(837 missing values generated)

. twoway (scatter plowbirthweightvalue airpollutionparticulatematterval)

.
. * the graph shows evidence of another concern
. * CONCERN 2: nonlinear relationships
.
. * one approach: look for non-linearities in the residuals of
. * the main model without pollution,
. regress lowbirthweightvalue /*
> */ childreninpovertyvalue medianhouseholdincomevalue/*
> */ primarycarephysiciansvalue uninsuredvalue /*
> */ percentofpopulationthatisnonhisp  percentofpopulationthatishispani /*
> */ teenbirthsvalue somecollegevalue/*
> */ adultsmokingvalue excessivedrinkingvalue physicalinactivityvalue/*
> */ if airpollutionparticulatematterval ~=. [w = populationestimatevalue], vce(robust)
(analytic weights assumed)
(sum of wgt is 298,959,926)

Linear regression                               Number of obs     =      2,073
                                                F(11, 2061)       =     162.45
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7168
                                                Root MSE          =     .00783

------------------------------------------------------------------------------------------------
                               |               Robust
           lowbirthweightvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------------------+----------------------------------------------------------------
        childreninpovertyvalue |    .020754   .0121387     1.71   0.087    -.0030515    .0445595
    medianhouseholdincomevalue |   1.12e-07   4.59e-08     2.45   0.014     2.23e-08    2.02e-07
    primarycarephysiciansvalue |   .0000513   .0000133     3.86   0.000     .0000252    .0000774
                uninsuredvalue |   .0038858   .0103917     0.37   0.708    -.0164935    .0242651
percentofpopulationthatisnon~p |   .0706107   .0039302    17.97   0.000     .0629031    .0783183
percentofpopulationthatishis~i |   .0089242   .0042358     2.11   0.035     .0006173    .0172311
               teenbirthsvalue |   .0001219   .0000454     2.69   0.007      .000033    .0002109
              somecollegevalue |   .0233632   .0059157     3.95   0.000     .0117617    .0349646
             adultsmokingvalue |    .075996   .0115168     6.60   0.000     .0534101    .0985818
        excessivedrinkingvalue |  -.0286934   .0089884    -3.19   0.001    -.0463207   -.0110661
       physicalinactivityvalue |   .0672259   .0112663     5.97   0.000     .0451313    .0893205
                         _cons |   .0119696   .0069717     1.72   0.086    -.0017026    .0256419
------------------------------------------------------------------------------------------------
```

```
. predict plbw1
(option xb assumed; fitted values)
(1,034 missing values generated)

. predict rlbw1, residuals
(1,046 missing values generated)

. lowess rlbw1 airpollutionparticulatematterval


.
. * another approach: break the key independent variable into discrete categories
.
. gen airpollutionparticulatematterint = int(airpollutionparticulatematterval)
(35 missing values generated)

. fvset base 11 airpollutionparticulatematterint


.
. regress lowbirthweightvalue i.airpollutionparticulatematterint /*
> */ childreninpovertyvalue  medianhouseholdincomevalue /*
> */ primarycarephysiciansvalue couldnotseedoctorduetocostvalue /*
> */ [w = populationestimatevalue], vce(robust)
(analytic weights assumed)
(sum of wgt is 303,319,296)

Linear regression                               Number of obs     =      2,291
                                                F(11, 2279)       =      51.30
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4550
                                                Root MSE          =     .01102

------------------------------------------------------------------------------------------------
                               |               Robust
           lowbirthweightvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------------------+----------------------------------------------------------------
airpollutionparticulatematte~t |
                            7  |  -.0202838   .0031488    -6.44   0.000    -.0264587   -.0141089
                            8  |  -.0121942   .0032742    -3.72   0.000     -.018615   -.0057734
                            9  |  -.0119895   .0018157    -6.60   0.000    -.0155501   -.0084288
                           10  |  -.0045001   .0015727    -2.86   0.004    -.0075842   -.0014161
                           12  |   .0034743   .0013539     2.57   0.010     .0008193    .0061293
                           13  |    .002548   .0013449     1.89   0.058    -.0000894    .0051854
                           14  |   .0014275   .0014365     0.99   0.320    -.0013896    .0042445
                               |
        childreninpovertyvalue |   .1153426   .0111496    10.35   0.000     .0934783     .137207
    medianhouseholdincomevalue |   2.30e-07   5.52e-08     4.17   0.000     1.22e-07    3.38e-07
    primarycarephysiciansvalue |    .000063   .0000144     4.38   0.000     .0000348    .0000911
couldnotseedoctorduetocostva~e |   .0510972   .0131323     3.89   0.000     .0253447    .0768497
                         _cons |   .0338464    .004665     7.26   0.000     .0246983    .0429945
------------------------------------------------------------------------------------------------


.
. * yet another issue: the possibility of "hot-spots"
.
. * according to summary stats, low birthweight has a sample mean of 8.2%
. generate lowbirthweightcounty_yn = .
(3,143 missing values generated)

. replace lowbirthweightcounty_yn = 0 if lowbirthweightvalue > 0 & lowbirthweightvalue < .082
(1,915 real changes made)

. replace lowbirthweightcounty_yn = 1 if lowbirthweightvalue >= 0.082 & lowbirthweightvalue < .24
(1,127 real changes made)

.
. * run the full model on the dichotomous outcome to see if the relationship still shows
```

```
. logit lowbirthweightcounty_yn airpollutionparticulatematterval /*
> */ childreninpovertyvalue  medianhouseholdincomevalue /*
> */ primarycarephysiciansvalue couldnotseedoctorduetocostvalue /*
> */ percentofpopulationthatisnonhisp, vce(robust)

Iteration 0:   log pseudolikelihood = -1525.4273
Iteration 1:   log pseudolikelihood = -907.05788
Iteration 2:   log pseudolikelihood = -893.80137
Iteration 3:   log pseudolikelihood = -893.59567
Iteration 4:   log pseudolikelihood = -893.59562

Logistic regression                             Number of obs     =      2,291
                                                Wald chi2(6)      =     507.14
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -893.59562               Pseudo R2         =     0.4142

------------------------------------------------------------------------------------------------
                               |               Robust
       lowbirthweightcounty_yn |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------------+----------------------------------------------------------------
airpollutionparticulatematte~l |   .3765324   .0407742     9.23   0.000     .2966165    .4564483
        childreninpovertyvalue |   6.403843   1.503217     4.26   0.000     3.457591    9.350095
    medianhouseholdincomevalue |  -.0000191   .0000129    -1.48   0.139    -.0000444    6.19e-06
    primarycarephysiciansvalue |   .0055469   .0021423     2.59   0.010      .001348    .0097458
couldnotseedoctorduetocostva~e |   8.571171    1.52381     5.62   0.000     5.584558    11.55778
percentofpopulationthatisnon~p |     13.176   .8290623    15.89   0.000     11.55107    14.80094
                         _cons |   -8.46675   1.073781    -7.88   0.000    -10.57132   -6.362177
------------------------------------------------------------------------------------------------

.
. * according to summary stats, one in 20 counties has more than 12.5% low birthweight
.
. generate vlowbirthweightcounty_yn = .
(3,143 missing values generated)

. replace vlowbirthweightcounty_yn = 0 if lowbirthweightvalue > 0 & lowbirthweightvalue < .125
(2,907 real changes made)

. replace vlowbirthweightcounty_yn = 1 if lowbirthweightvalue >= .125 & lowbirthweightvalue < .24
(135 real changes made)

.
. * run the full model on the extreme dichotomous outcome to see if the relationship still shows
. logit vlowbirthweightcounty_yn airpollutionparticulatematterval /*
> */ childreninpovertyvalue  medianhouseholdincomevalue /*
> */ primarycarephysiciansvalue couldnotseedoctorduetocostvalue /*
> */ percentofpopulationthatisnonhisp, vce(robust)

Iteration 0:   log pseudolikelihood = -426.24847
Iteration 1:   log pseudolikelihood = -270.47893
Iteration 2:   log pseudolikelihood = -205.51287
Iteration 3:   log pseudolikelihood = -194.81515
Iteration 4:   log pseudolikelihood = -194.11118
Iteration 5:   log pseudolikelihood = -194.10955
Iteration 6:   log pseudolikelihood = -194.10955

Logistic regression                             Number of obs     =      2,291
                                                Wald chi2(6)      =     184.26
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -194.10955               Pseudo R2         =     0.5446

------------------------------------------------------------------------------------------------
                               |               Robust
      vlowbirthweightcounty_yn |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------------+----------------------------------------------------------------
airpollutionparticulatematte~l |   .0509528   .1183694     0.43   0.667    -.1810471    .2829526
```

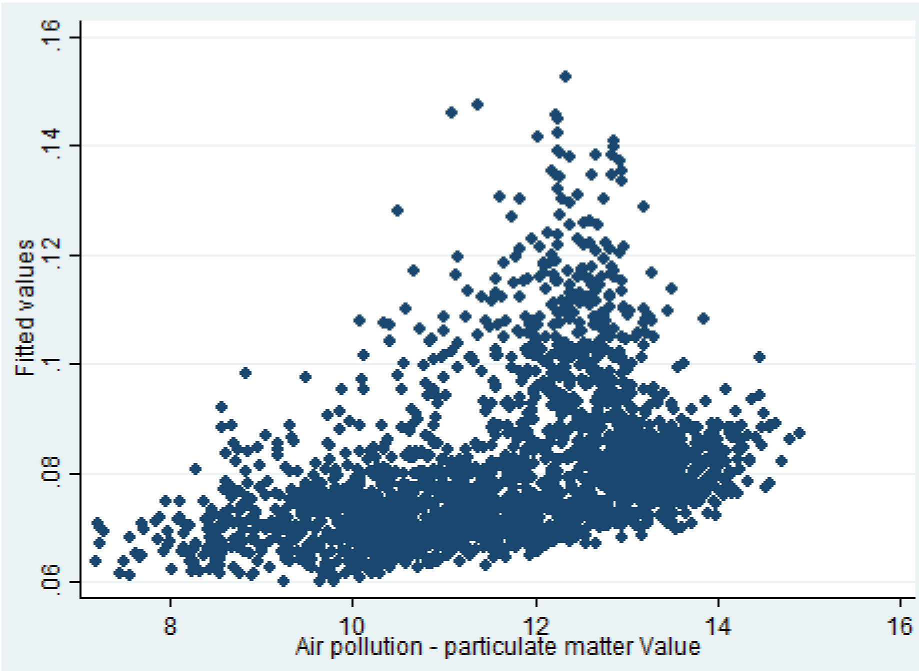
```
         childreninpovertyvalue |   7.979745   3.073515     2.60   0.009     1.955767    14.00372
    medianhouseholdincomevalue |    -.00008   .0000503    -1.59   0.111    -.0001785    .0000185
    primarycarephysiciansvalue |  -.0038778   .0058941    -0.66   0.511      -.01543    .0076745
couldnotseedoctorduetocostva~e |   7.020765   3.093142     2.27   0.023      .958319    13.08321
percentofpopulationthatisnon~p |   6.923033   .7382868     9.38   0.000     5.476017    8.370048
                         _cons |  -6.051736   3.466156    -1.75   0.081    -12.84528    .7418049
------------------------------------------------------------------------------------------------

.
end of do-file
```

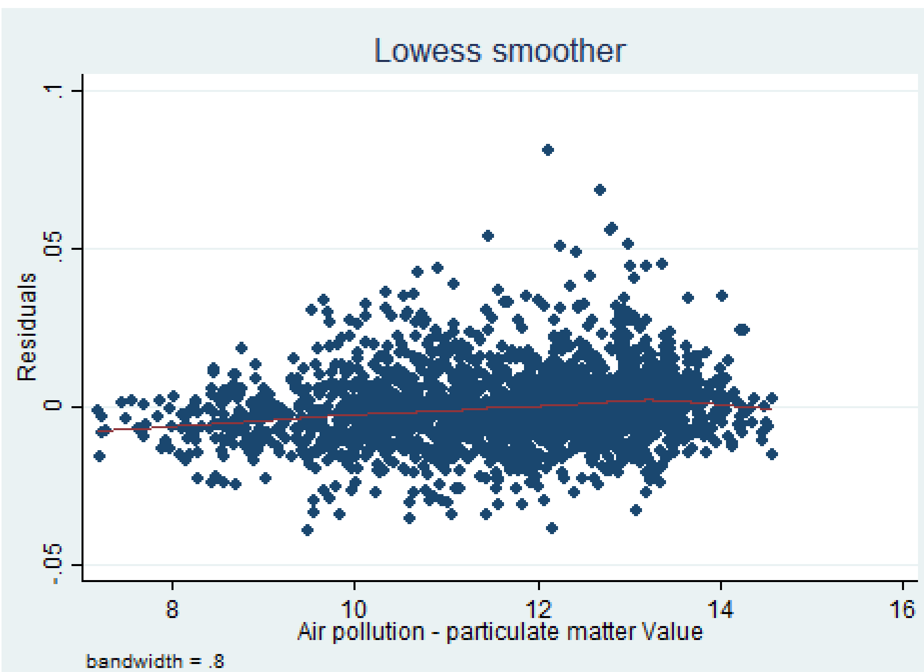
outreg2 [simple plushealthcare plusbehav] using myfile, replace see

	(1)	(2)	(3)	
	simple	plushealthcare	plusbehav	Robust standard errors in parentheses
VARIABLES	lowbirthweightvalue	Lowbirthweightvalue	lowbirthweightvalue	*** p<0.01, ** p<0.05, * p<0.1
airpollutionparticulatematterval	0.00311***	0.00376***	0.00211***	
	(0.000320)	(0.000337)	(0.000259)	
childreninpovertyvalue		0.118***	0.0330***	
		(0.0113)	(0.00942)	
medianhouseholdincomevalue		2.47e-07***	6.58e-09	
		(5.62e-08)	(4.30e-08)	
primarycarephysiciansvalue		5.98e-05***	2.43e-05**	
		(1.40e-05)	(1.20e-05)	
couldnotseedoctorduetocostvalue		0.0461***	0.0524***	
		(0.0141)	(0.0116)	
percentofpopulationthatisnonhisp			0.0681***	
			(0.00391)	
Constant	0.0460***	-0.0119*	0.0318***	
	(0.00372)	(0.00617)	(0.00490)	
Observations	3,016	2,291	2,291	

twoway (scatter plowbirthweightvalue airpollutionparticulatematterval)



lowess rlbw1 airpollutionparticulatematterval



Useful additional stuff:
Some commands for RCTs

```
. * if you want to compare a treatment and a control group values on a continuous variable

 ttest YEARSJOB, by(nonstandard) unequal
```

- Two-sample t test with unequal variances
- --
-    Group | Obs   Mean   Std. Err.   Std. Dev.   [95% Conf. Interval]
- ---------+--
-        0 | 980   9.430612   .2788544   8.729523   8.883391   9.977833
-        1 | 379   7.907652   .3880947   7.555398   7.144557   8.670747
- ---------+--
- combined | 1359   9.005887   .2290413   8.443521   8.556573   9.4552
- ---------+--
-     diff | 1.522961   .4778884   .5848756   2.461045
- --
-     diff = mean(0) - mean(1)   t =   3.1869
- Ho: diff = 0   Satterthwaite's degrees of freedom = 787.963
-     Ha: diff < 0   Ha: diff != 0   Ha: diff > 0
- Pr(T < t) = 0.9993   Pr(|T| > |t|) = 0.0015   Pr(T > t) = 0.0007

```
. * you can also do this with immediate commands if you are just handed the summary statistics

. * ttesti (Ntreat, meantreat, sdtreat, Ncont, meancont, sdcont)
. ttesti 4252 18.1 12.9 6764 32.6 18.2, unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |   4,252        18.1    .1978304        12.9    17.71215    18.48785
       y |   6,764        32.6     .221294        18.2    32.16619    33.03381
---------+--------------------------------------------------------------------
combined |  11,016    27.00323    .1697512     17.8166    26.67049    27.33597
---------+--------------------------------------------------------------------
    diff |               -14.5    .2968297               -15.08184   -13.91816
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t = -48.8496
Ho: diff = 0                     Satterthwaite's degrees of freedom =  10858.6

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000

.
. * to compare a treatment and a control group values on a categorical variable

    . prtest nonstandard if (RACECEN1==1 | RACECEN1==2), by(RACECEN1)
    Two-sample test of proportion                      1: Number of obs =     1389
                                                       2: Number of obs =      260
    ------------------------------------------------------------------------------
        Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
    -------------+----------------------------------------------------------------
               1 |   .2800576   .0120482                      .2564436    .3036716
               2 |   .3538462   .0296544                      .2957247    .4119676
    -------------+----------------------------------------------------------------
            diff |  -.0737886   .0320084                     -.1365239   -.0110532
                 |  under Ho:   .0307147    -2.40   0.016
    ------------------------------------------------------------------------------
            diff = prop(1) - prop(2)                                  z =  -2.4024
        Ho: diff = 0
        Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
     Pr(Z < z) = 0.0081         Pr(|Z| < |z|) = 0.0163          Pr(Z > z) = 0.9919
```

```
. * again, you can do this with immediate commands if you are just handed the summary statistics

. * prtesti (Ntreat, ptreat, Ncont, pcont)
. prtesti 345 .3536 1900 .1411

Two-sample test of proportions                    x: Number of obs =      345
                                                  y: Number of obs =     1900
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |      .3536   .0257393                      .3031518    .4040482
           y |      .1411   .0079865                      .1254467    .1567533
-------------+----------------------------------------------------------------
        diff |      .2125   .0269499                      .1596791    .2653209
             |  under Ho:   .0221741     9.58   0.000
------------------------------------------------------------------------------
        diff = prop(x) - prop(y)                                  z =   9.5833
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 1.0000         Pr(|Z| > |z|) = 0.0000          Pr(Z > z) = 0.0000
```

FYI: here are some other regression-style models you might be asked to run, with commands and outputs similar to `regress`

probit	probit model
mlogit	multinomial logit model
ologit	ordinal logit model
tobit	mixed regression and logit model
xi: glm	loglinear model
	(plus fixed effects and random effects models for categorical variables)

Lastly: here is a variant of a logit regression model that I have chosen for our STATA topic next week.
This is a variant that counts not only *whether* an event occurs, but also *when* an event occurs. This is often a useful approach in RCTs that involve treatments and outcomes measured at multiple time points. (Survival rates from cancer treatments, time it takes to find a job, criminal recidivism, etc.)

streg hazard model (aka rate/survival model)