Lecture 7: Review

Prof. Sharyn O'Halloran

Sustainable Development U9611

Econometrics II

Ex 1029: Wage and Race

- The dataset provided is designed to explore the relationship between wage and race (black_indicator), controlling for the region in the US, education, experience and weather they worked in a standard metropolitan statistical area.
- Model to be tested:

$$lwage = \beta_0 + \beta_1 \exp er + \beta_2 educ + \beta_3 smsa\_ind + \beta_4 region + \beta_5 black\_ind + u$$

Creating Dummy Variables and Interactive Terms

- We proceed by recoding region into 4 dummies:
- We rewrite our model including interaction terms as follows:

$$
\begin{aligned}
lwage = \beta_0 + \beta_1 \exp er + \beta_2 educ + \beta_3 smsa\_ind \\
+ \beta_4 regMW + \beta_5 regNE + \beta_6 regS \\
+ \beta_7 black\_ind + \beta_8 blackregMW \\
+ \beta_9 blackregNE + \beta_{10} blackregSE + u
\end{aligned}
$$

Hypotheses

- We expect positive coefficients for:
  - Education
  - Experience and
  - SMSA
- We expect a negative coefficient on:
  - black-indicator

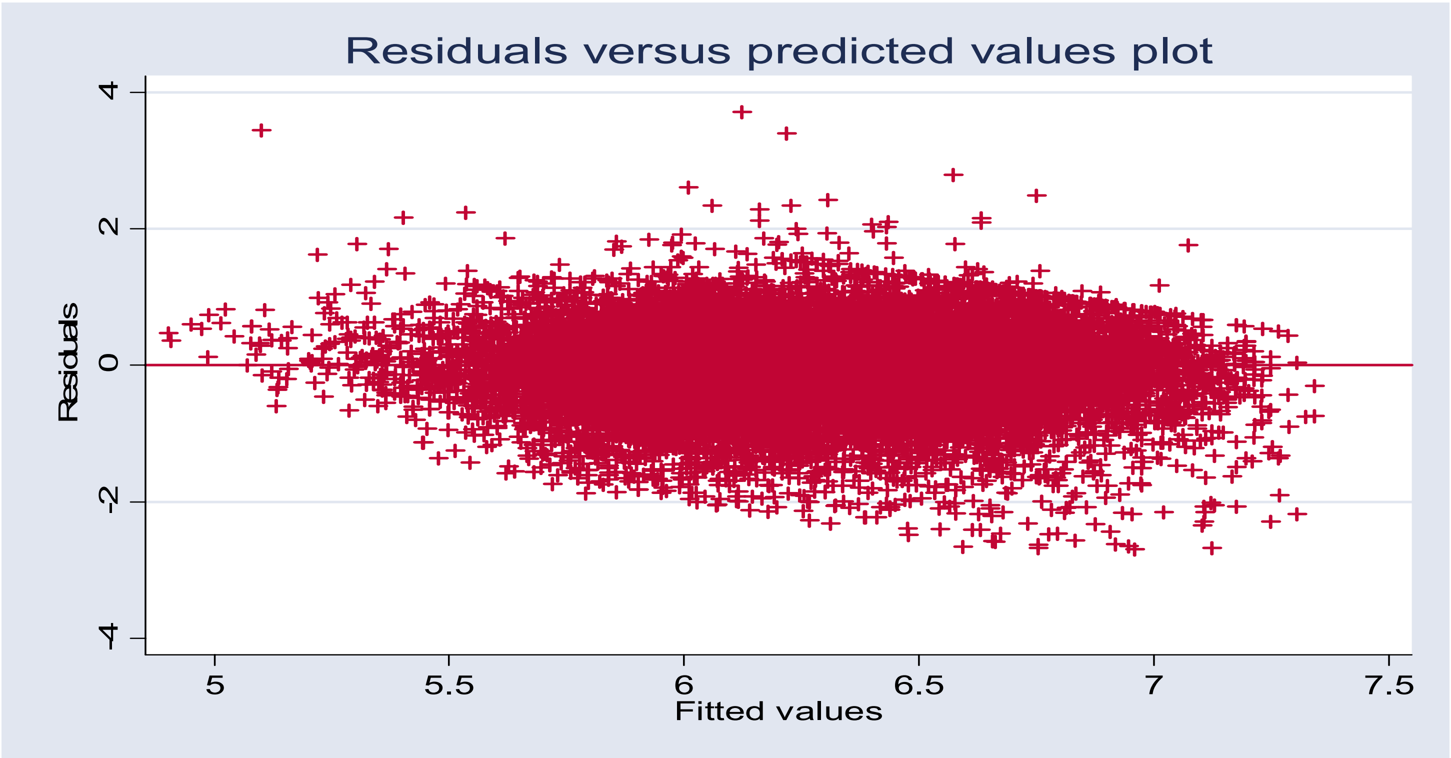
Results of Tentative Model

- **reg lwage exper educ smsa_ind regMW regNE regS black_ind blackregMW blackregNE blackregS**

```
      Source |       SS       df       MS              Number of obs =   25631
-------------+------------------------------           F( 10, 25620) = 1010.01
       Model |  2852.13293    10  285.213293           Prob > F      =  0.0000
    Residual |  7234.7208 25620  .282385667           R-squared     =  0.2828
-------------+------------------------------           Adj R-squared =  0.2825
       Total |  10086.8537 25630  .393556525           Root MSE      =   .5314
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0183495   .0002789    65.79   0.000     .0178028    .0188961
        educ |   .0969922   .0012015    80.72   0.000     .0946371    .0993473
    smsa_ind |   .1575999   .0077298    20.39   0.000     .1424491    .1727507
       regMW |   .0034929   .0100704     0.35   0.729    -.0162456    .0232315
       regNE |   .0382672   .0102318     3.74   0.000     .0182123    .0583221
        regS |  -.0571932   .0097345    -5.88   0.000    -.0762735    -.038113
   black_ind |  -.1937687    .040989    -4.73   0.000    -.2741094    -.113428
  blackregMW |  -.0468973    .051017    -0.92   0.358    -.1468935     .053099
  blackregNE |  -.0242864   .0508993    -0.48   0.633    -.1240519    .0754792
   blackregS |  -.0435901   .0443259    -0.98   0.325    -.1304714    .0432912
       _cons |   4.573932   .0196774   232.45   0.000     4.535363    4.612501
------------------------------------------------------------------------------
```

Fail to reject the null hypothesis that $\beta_i=0$ in favor of the alternatives that $\beta_i \neq 0$.

Residual Plots



- No obvious pattern

F-test

- Test the joint significance of the interactive terms
- Command:
  - test blackregMW blackregNE blackregS

```
( 1)  blackregMW = 0
( 2)  blackregNE = 0
( 3)  blackregS = 0
      F(  3, 25620) =    0.42
          Prob > F =    0.7408
```

- Results:
  - Variables not jointly significant
  - Remove from model

Re-run Results

- reg lwage exper educ smsa_ind regMW regNE regS black_ind

```
      Source |       SS       df       MS              Number of obs =   25631
-------------+------------------------------           F(  7, 25623) = 1442.80
       Model |  2851.77965     7  407.397093           Prob > F      =  0.0000
    Residual |  7235.07408 25623  .282366393           R-squared     =  0.2827
-------------+------------------------------           Adj R-squared =  0.2825
       Total |  10086.8537 25630  .393556525           Root MSE      =  .53138
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0183511   .0002789    65.81   0.000     .0178046    .0188977
        educ |   .0970151   .0012011    80.77   0.000     .0946609    .0993693
    smsa_ind |   .1578088   .0077105    20.47   0.000     .1426959    .1729218
       regMW |   .0017984   .0098616     0.18   0.855    -.0175308    .0211276
       regNE |   .0377502   .0100117     3.77   0.000     .0181268    .0573737
        regS |  -.0593619   .0094407    -6.29   0.000    -.0778662   -.0408576
   black_ind |   -.230438    .012657   -18.21   0.000    -.2552465   -.2056296
       _cons |   4.574619   .0196608   232.68   0.000     4.536083    4.613155
------------------------------------------------------------------------------
```

After removing interactive terms, black indicator variable remains significant

Interpretation

- Keeping all else constant:
  - This is a log-level problem; we're regressing the log of y on the level of x
  - So we use the formula: $\%\Delta y=(100\beta)\Delta x$
  - In this case, b=-.23, so a 1-unit change in x causes a 23% decrease in y.
  - That is, black workers on average have wages 23% lower than non-black workers.
- Also note salary differentials by region

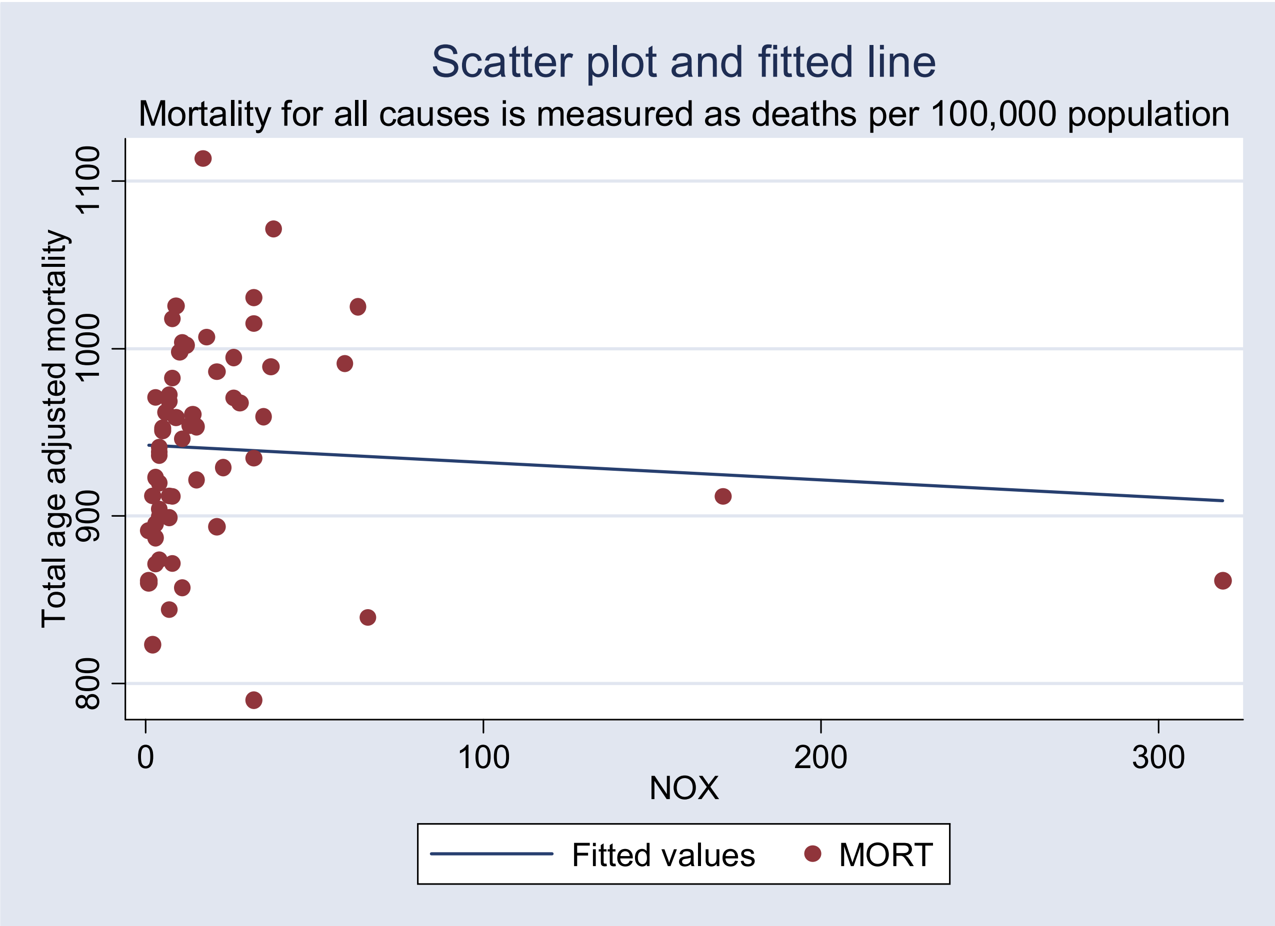
Ex 1123: Air Pollution and Mortality

- The dataset provided is designed to explore the relationship between mortality rate and concentrations in dangerous pollutants such as nitrogen oxides and sulfur dioxide.
- The model we would like to study is the following:

$$mortality = \beta_0 + \beta_1 \log(NO_x) + \beta_2 \log(SO_2) + \beta_3 precipitation + \beta_4 education + \beta_5 non-white + u$$

Transforming Variables

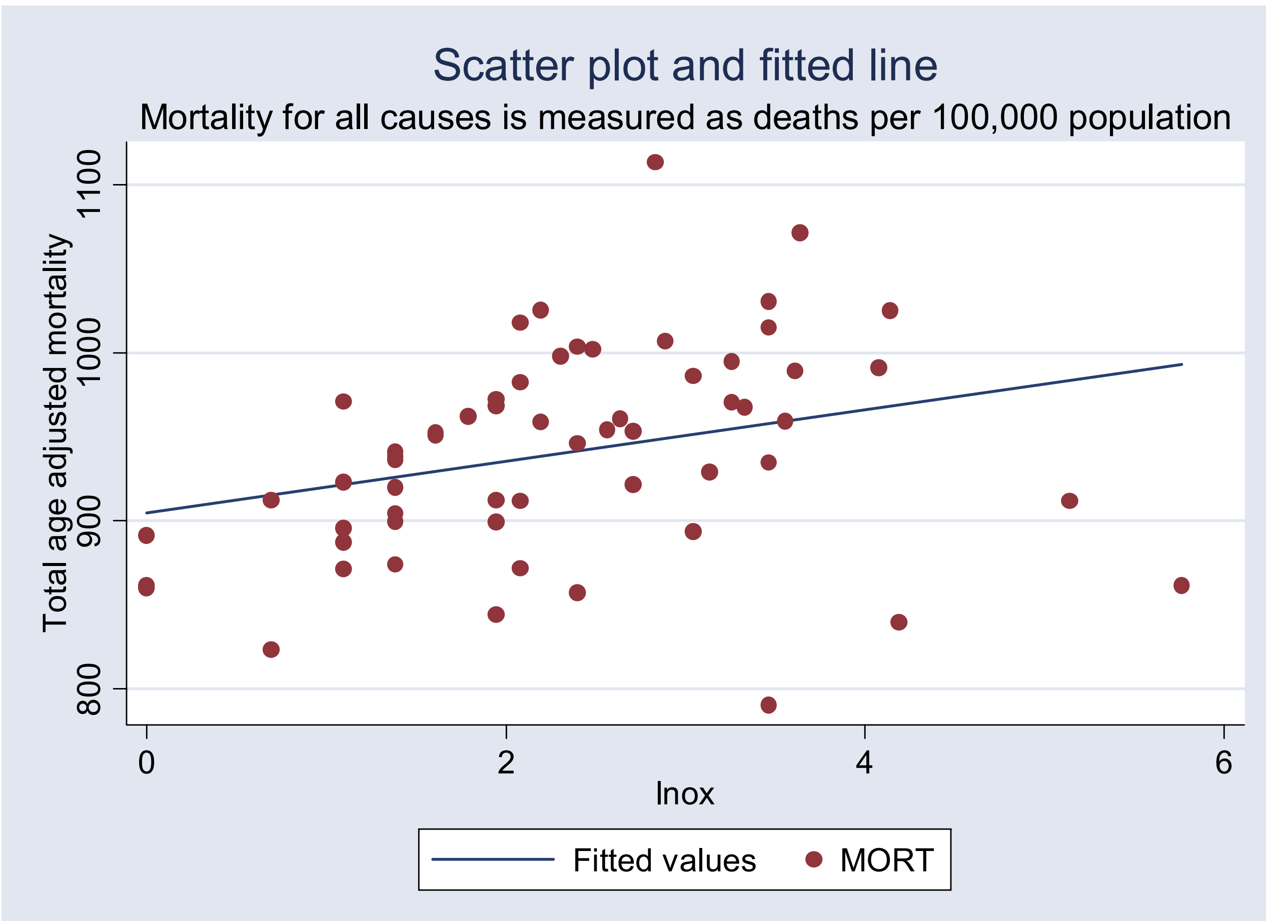
- It makes sense to log the independent variables for NO_x and SO_2



Scatterplot with NO_x

Transforming Variables

- It makes sense to log the independent variables for Nox and SO2



Scatterplot with log of No_x

Hypotheses

- We expect positive coefficients on:
  - log(NOx)
  - log(SO_2)
  - Precipitation (due to acid rain)
  - Non-white population
- We expect a negative coefficient on:
  - Education

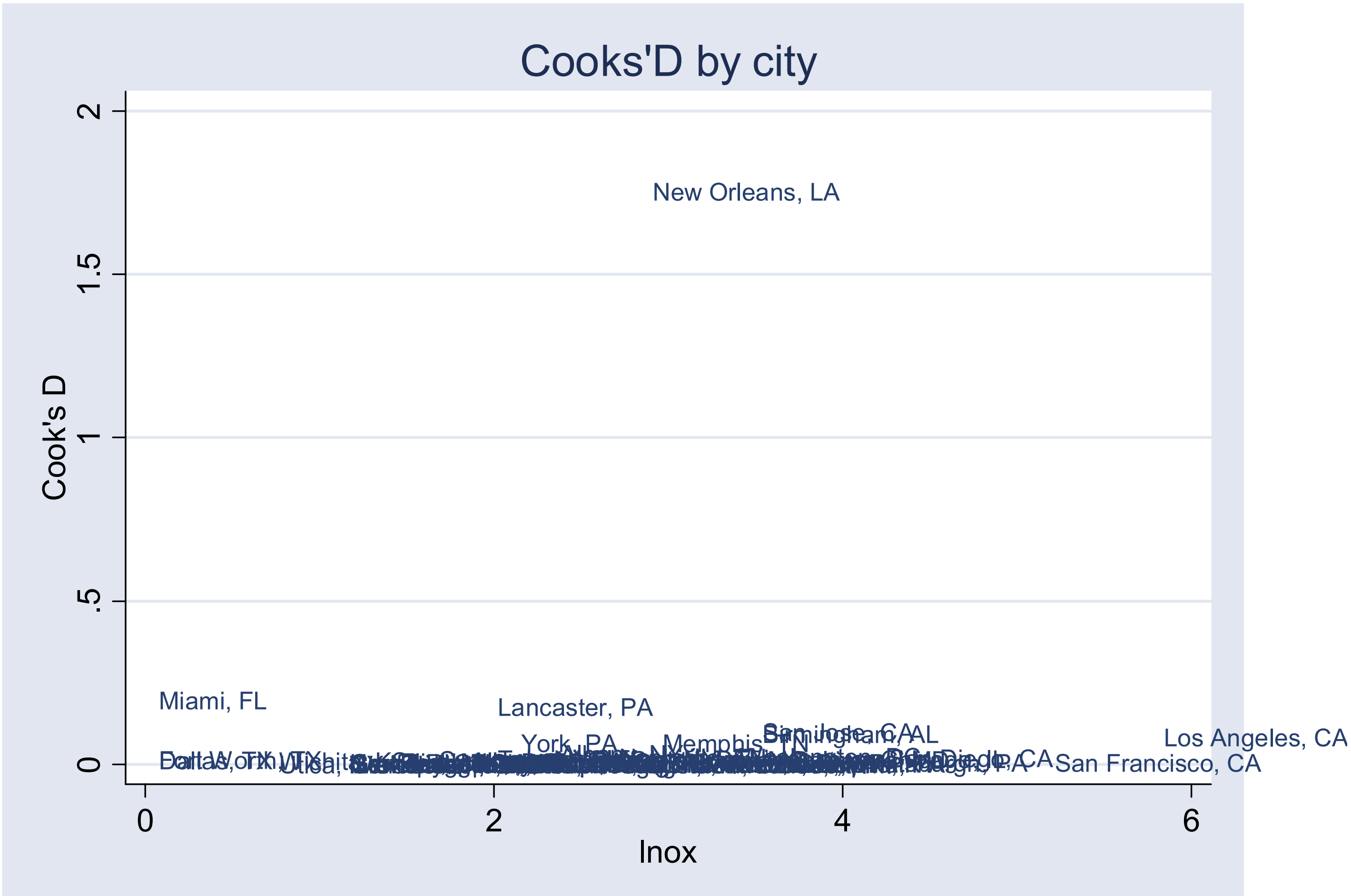
Results from Tentative Model

reg mort lnox lso2 precip educ nonwhite

```
      Source |       SS       df       MS              Number of obs =      60
-------------+------------------------------           F(  5,    54) =   23.85
       Model |  157116.254     5  31423.2507           Prob > F      =  0.0000
    Residual |  71159.1703    54  1317.76241           R-squared     =  0.6883
-------------+------------------------------           Adj R-squared =  0.6594
       Total |  228275.424    59  3869.07498           Root MSE      =  36.301
------------------------------------------------------------------------------
        mort |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lnox |   6.716442   7.399021     0.91   0.368    -8.117702    21.55059
        lso2 |   11.35782   5.295537     2.14   0.036     .7409073    21.97473
      precip |   1.946748   .7007028     2.78   0.008     .5419234    3.351573
        educ |  -14.66453   6.937913    -2.11   0.039    -28.57421   -.7548551
    nonwhite |   3.028928   .6685249     4.53   0.000     1.688616     4.36924
       _cons |   940.6586   94.05514    10.00   0.000     752.0894    1129.228
```

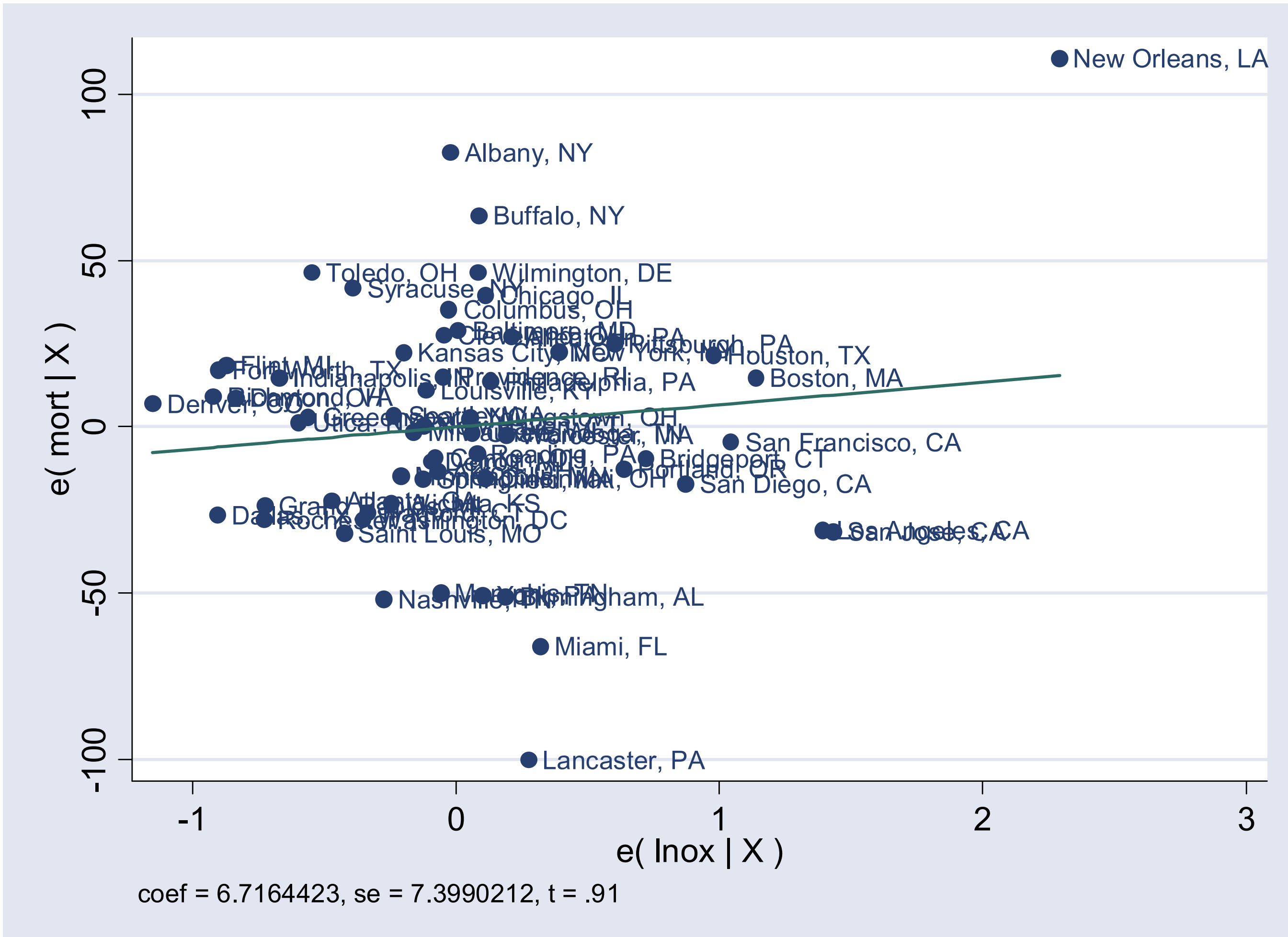
- All signs as expected
- Coefficient on NOx is insignificant, however

Checking Case Influence Statistics



- New Orleans has a high Cook's Distance

Checking for Problems



- It's also an outlier in the avplot for NOx…

Checking for Problems

- And for SO_2 as well.

Results Dropping New Orleans

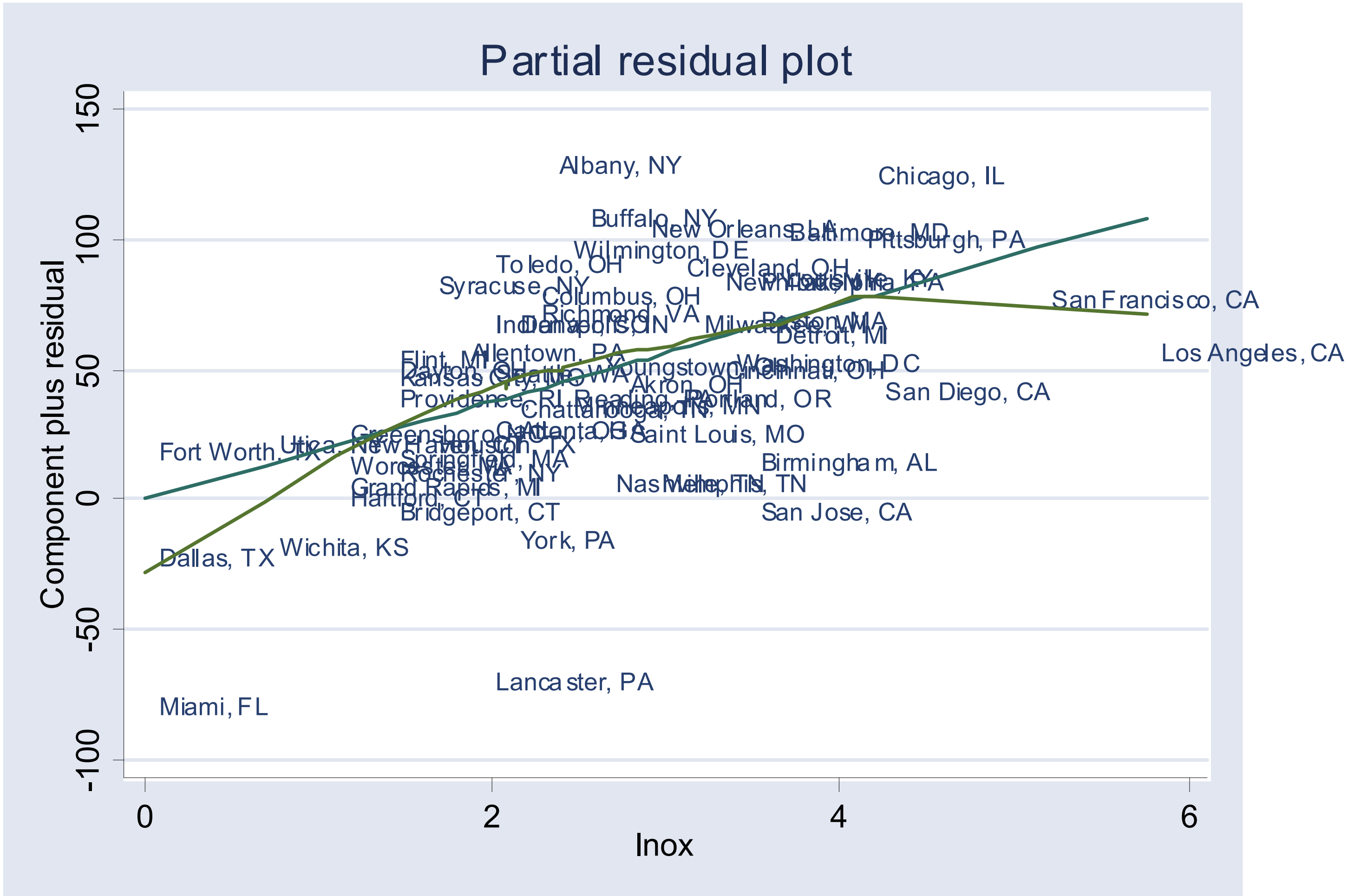
```
reg mort lnox lso2  precip educ nonwhite if city!="New Orleans"

      Source |       SS       df       MS              Number of obs =      59
-------------+------------------------------           F(  5,    53) =   27.90
       Model |  143441.648     5  28688.3296           Prob > F      =  0.0000
    Residual |  54501.7233    53   1028.3344           R-squared     =  0.7247
-------------+------------------------------           Adj R-squared =  0.6987
       Total |  197943.371    58  3412.81675           Root MSE      =  32.068

------------------------------------------------------------------------------
        mort |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lnox |   -9.89842   7.730678    -1.28   0.206    -25.4042    5.607357
        lso2 |   26.03266   5.931109     4.39   0.000    14.13636    37.92896
      precip |   1.363333   .6357352     2.14   0.037      .08821    2.638457
        educ |  -5.667182   6.523808    -0.87   0.389   -18.75228    7.417919
    nonwhite |   3.039655    .590569     5.15   0.000    1.855124    4.224186
       _cons |   852.3782   85.93317     9.92   0.000    680.0181    1024.738
```

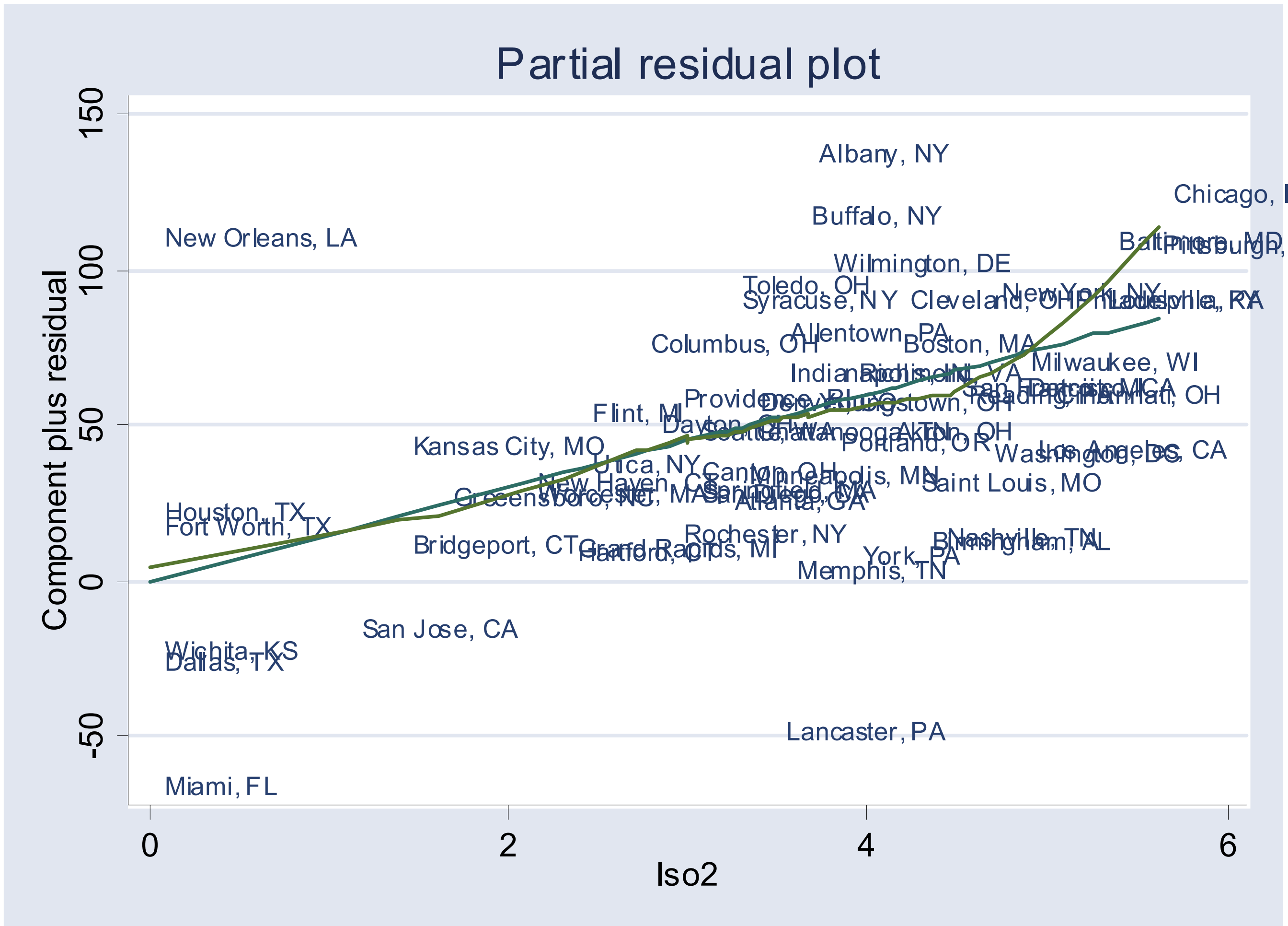
- Now education is no longer significant

Cprplot for NOx



- Some indications of non-linearity

Cprplot for SO_2



- This looks more or less linear

Review

1. “Regression”, “regression model”, “linear regression model”, “regression analysis”

2. Fitted values, residuals, least squares method of estimation

3. Properties of least squares; tests and confidence intervals for individual coefficients; prediction intervals; extra SS F-tests (full and reduced models)

4. Model building and refinement: transformation, indicator variables, x^2, interaction, variable selection

5. Influence and case-influence statistics

6. Variable selection

7. A note on the difference between "confounding variable" and "interaction"

a. Is there an association between gestation and mean brain weight after accounting for body weight?

$$\mu(\text{brain}) = \beta_0 + \beta_1 \text{ body} + \beta_2 \text{ gest}$$

(β_2 represents the association of gestation with mean brain weight after accounting for body weight.)

b. Is the association between gestation and brain weight Different for animals of different body sizes?

$$\mu(\text{brain}) = \beta_0 + \beta_1 \text{ body} + \beta_2 \text{ gest} + \beta_3 \text{ body*gest}$$

(There is an interactive effect of body and gest on brain)

8. What about all those F-tests?

a. All F-tests we've considered are special cases of the extra sum of squares F-test (Sect. 10.3)

b. F-test for overall significance of regression

Full: a model of interest

Reduced: model with β_0 only

c. F-test for lack-of fit

Full: one-way anova (separate means for each distinct combination of x's)

Reduced: a model of interest

d. Partial F-test is an F-test for a single β

e. One-way ANOVA F-test

Full: model with a separate mean for each group

i.e. β_0 and k-1 indicators to distinguish k groups

Reduced: b_0 only (single mean model)

f. “Type III” F-tests (a computer package term)

Full: model that has been specified

Reduced: model without a particular term

g. “Sequential” F-tests (depends on order that x’s are listed)

i. Full: intercept and x_1

Reduced: intercept

ii. Full: intercept, x_1, and x_2

Reduced: intercept and x_1

iii. Full: intercept, x_1, x_2, and x_3

Reduced: intercept, x_1, and x_2

9. In “linear regression,” what does “linear in b’s” mean?
a. β_0 *something + β_1 *something + β_2 *something + ...
b. Ex. of nonlinear regression: $\mu(y|x) = \beta_0 x^{\beta 1}$

10. A note about “mean response.” It is useful to explicitly write $\mu(y|x_1, x_2, x_3)$ to talk about the mean of y as a function of x_1, x_2, and x_3. Sometimes we abbreviate this to “the mean of the response” if it’s clear what x’s we’re talking about.

11. Partial residuals

a. You may find a plot of partial residuals vs. x_1 to be useful when it is desired to study the relationship between y and x_1 , after getting the effects of x_2 , x_3 . etc. out of the way, especially if the effect of x_1 is relatively small (in which case the plot of y versus x_1 does not reveal much).

b. For example: How is mammal brain weight related to litter size, after accounting for body weight?

c. Suppose $\mu(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. A plot of y versus x_1 won't show a linear relationship whose slope is β_1 if x_1 and x_2 are correlated. However, a plot of $y - (\beta_0 + \beta_2 x_2)$ versus x_1 will show a pattern whose slope is β_1 .

d. So, the partial residuals are yi - ( ), where the b's are the estimates from the regression of y on x_1 and x_1.