Lecture 4:
Regression Diagnostics

Modified slides from Prof. Sharyn O'Halloran

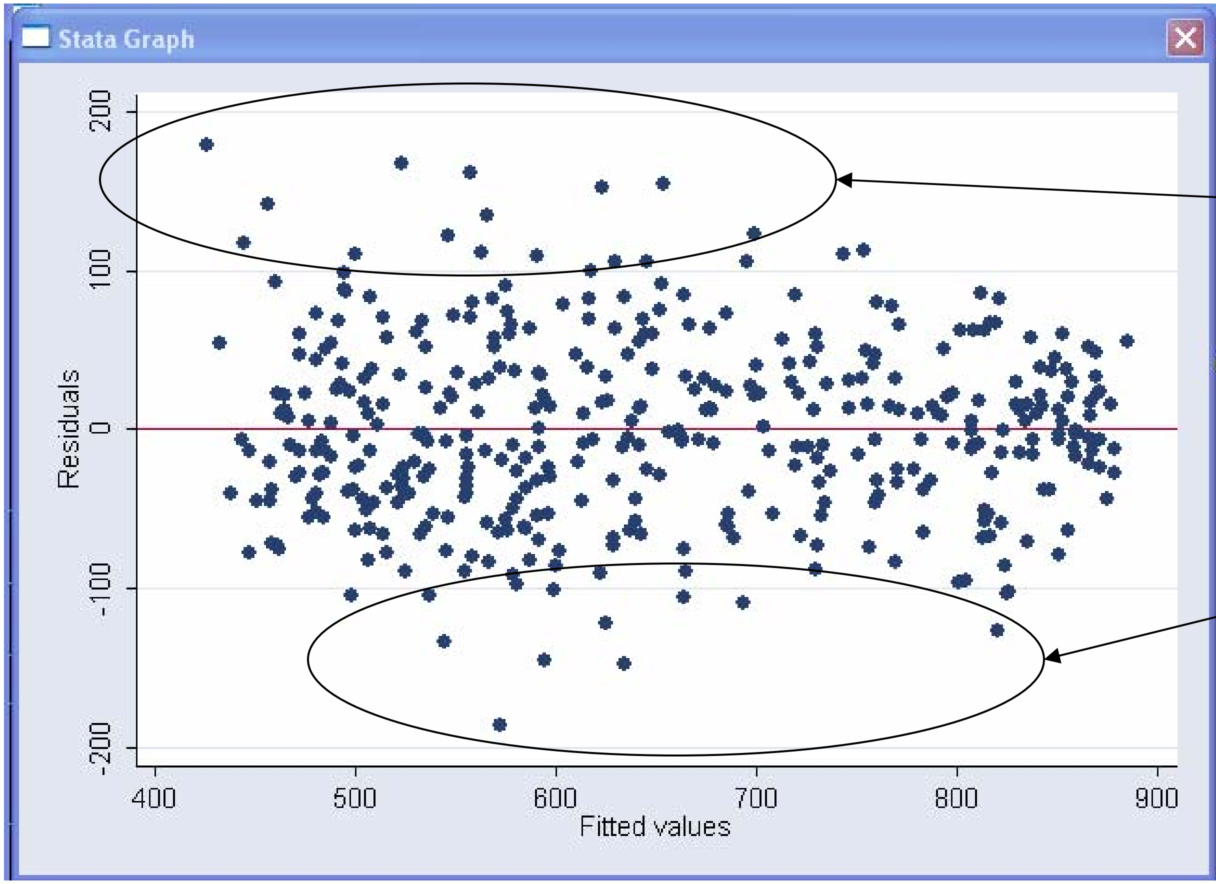
Regression Diagnostics

- Unusual and Influential Data
  - Outliers
  - Leverage
  - Influence
- Heteroscedasticity
  - Non-constant variance
- Multicollinearity
  - Non-independence of x variables

Unusual and Influential Data

- Outliers
  - An observation with large residual.
    - An observation whose dependent-variable value is unusual given its values on the predictor variables.
    - An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.

Outliers



Largest positive outliers

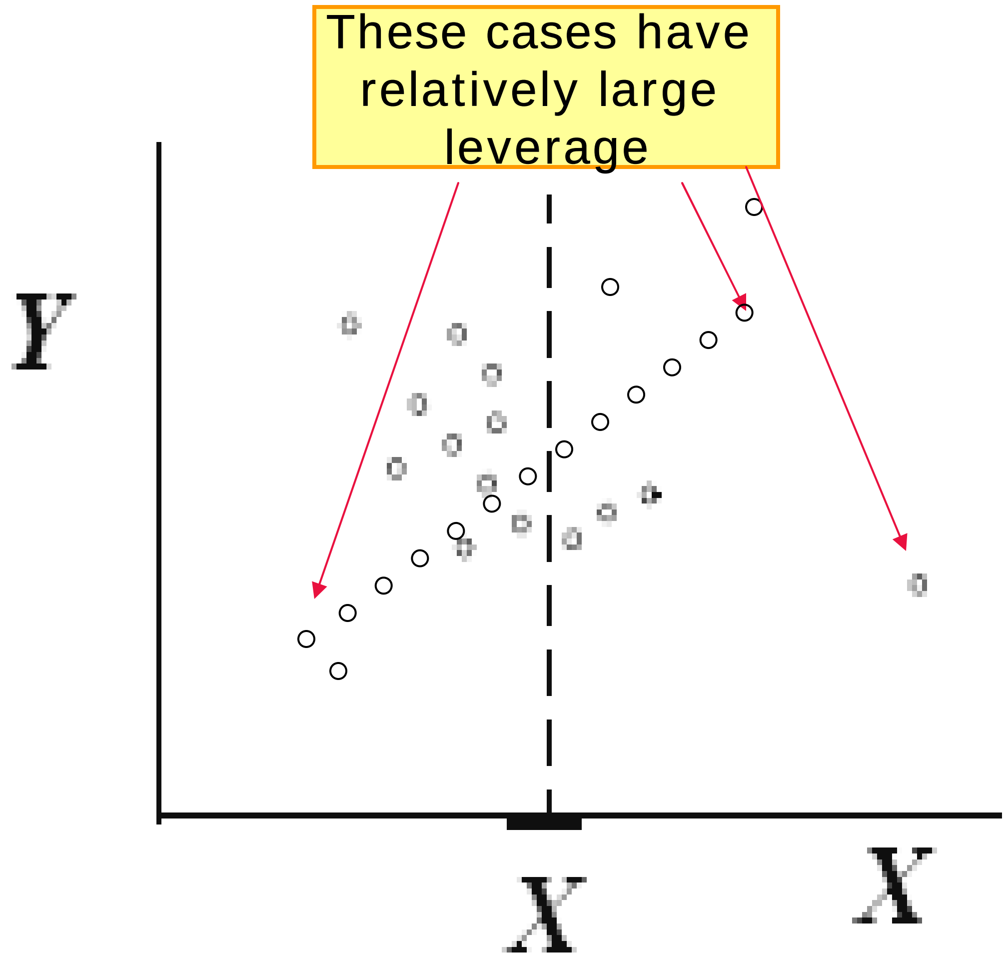
Largest negative outliers

```
reg api00 meals ell emer
rvfplot, yline(0)
```

Unusual and Influential Data

- Outliers
  - An observation with large residual.
    - An observation whose dependent-variable value is unusual given its values on the predictor variables.
    - An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.
- Leverage
  - An observation with an extreme value on a predictor variable
    - Leverage is a measure of how far an independent variable deviates from its mean.
    - These leverage points can have an effect on the estimate of regression coefficients.

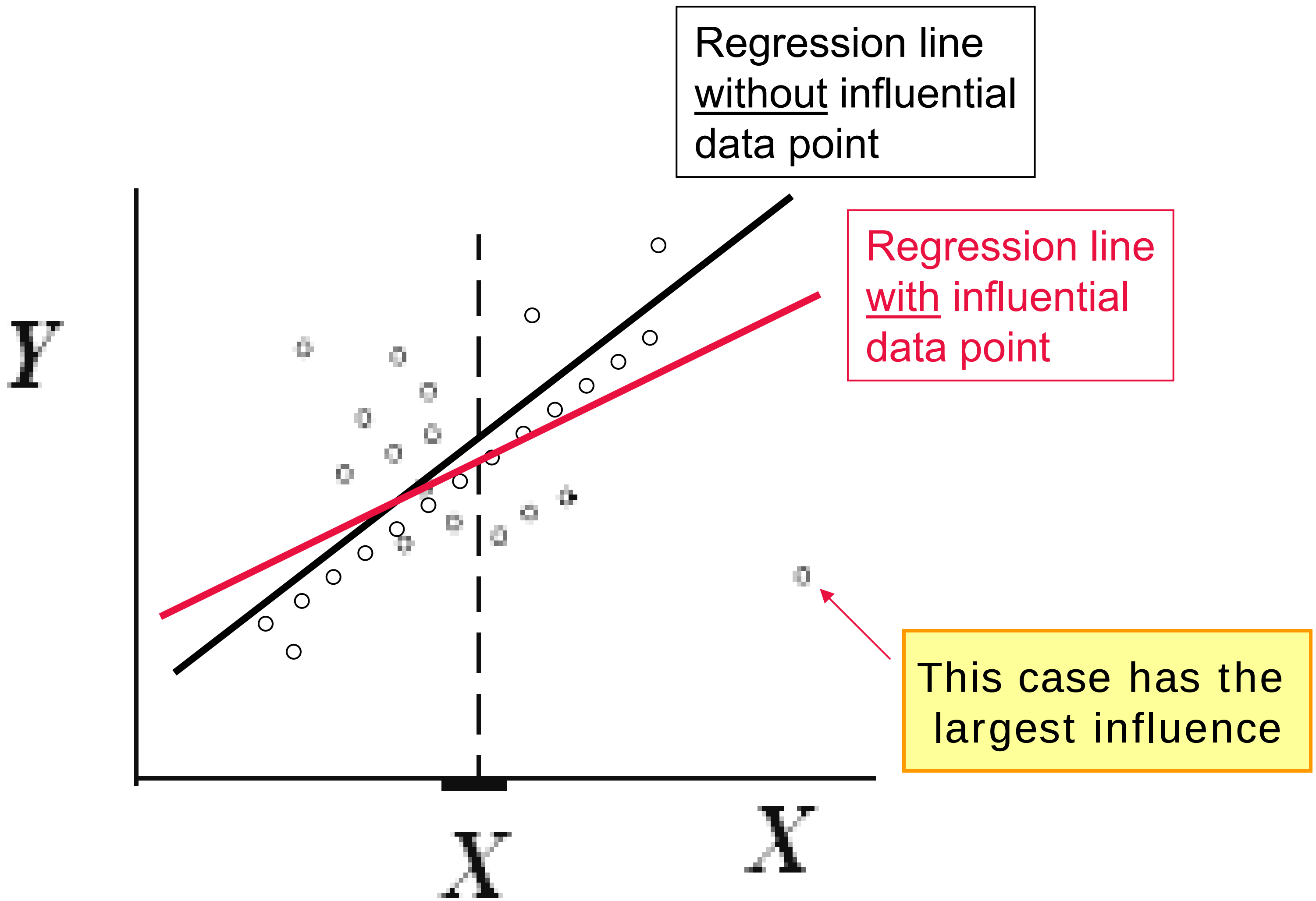
Leverage



Unusual and Influential Data

- Outliers
  - An observation with large residual.
    - An observation whose dependent-variable value is unusual given its values on the predictor variables.
    - An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.
- Leverage
  - An observation with an extreme value on a predictor variable
    - Leverage is a measure of how far an independent variable deviates from its mean.
    - These leverage points can have an effect on the estimate of regression coefficients.
- Influence
  - Influence can be thought of as the product of leverage and outlierness.
    - Removing the observation substantially changes the estimate of coefficients.

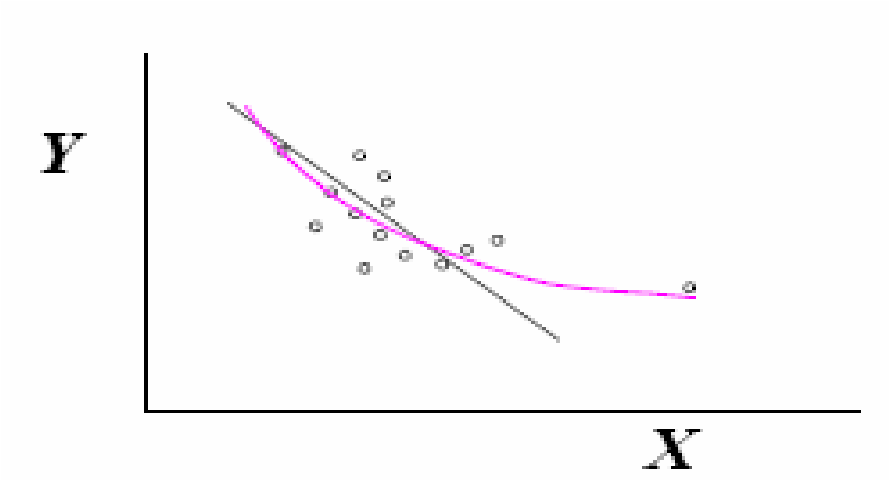
Influence



Regression line without influential data point
Regression line with influential data point
This case has the largest influence
Y
X
X

Introduction

- The problem: least squares is not resistant
  - One or several observations can have undue *influence* on the results



A quadratic-in-x term is significant here, but not when largest x is removed.

- Why is this a problem?
  - Conclusions that hinge on one or two data points must be considered extremely fragile and possibly misleading.

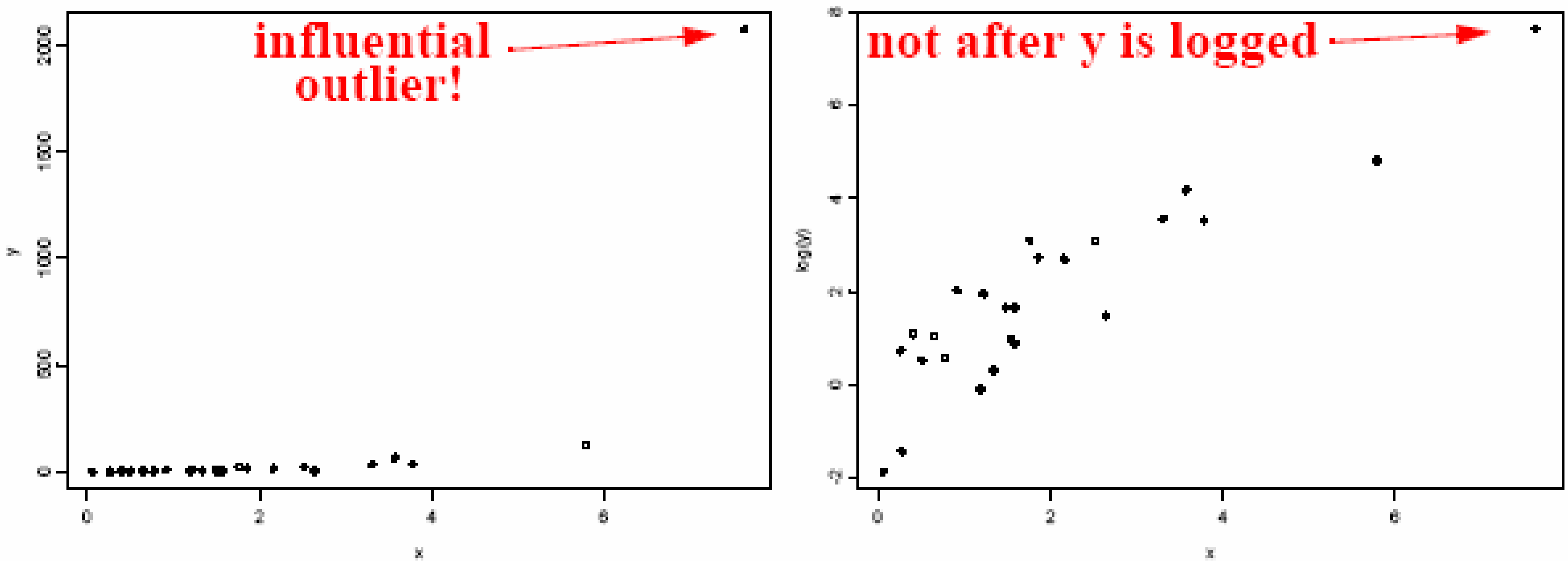
Tools

- Scatterplots
- Residuals plots
- Tentative fits of models with one or more cases set aside
- A strategy for dealing with influential observations
- Tools to help detect outliers and influential cases
  - Cook's distance
  - Leverage
  - Studentized residual

Difficulties to overcome

- Detection of influential observations depends on
  - Having determined a good scale for *y* (transformation) first
  - Having the appropriate *x*'s in the model,
- But assessment of appropriate functional form and *x*'s can be affected by influential observations (see previous page).

Example of Influential Outliers



Log transformation smoothes data and

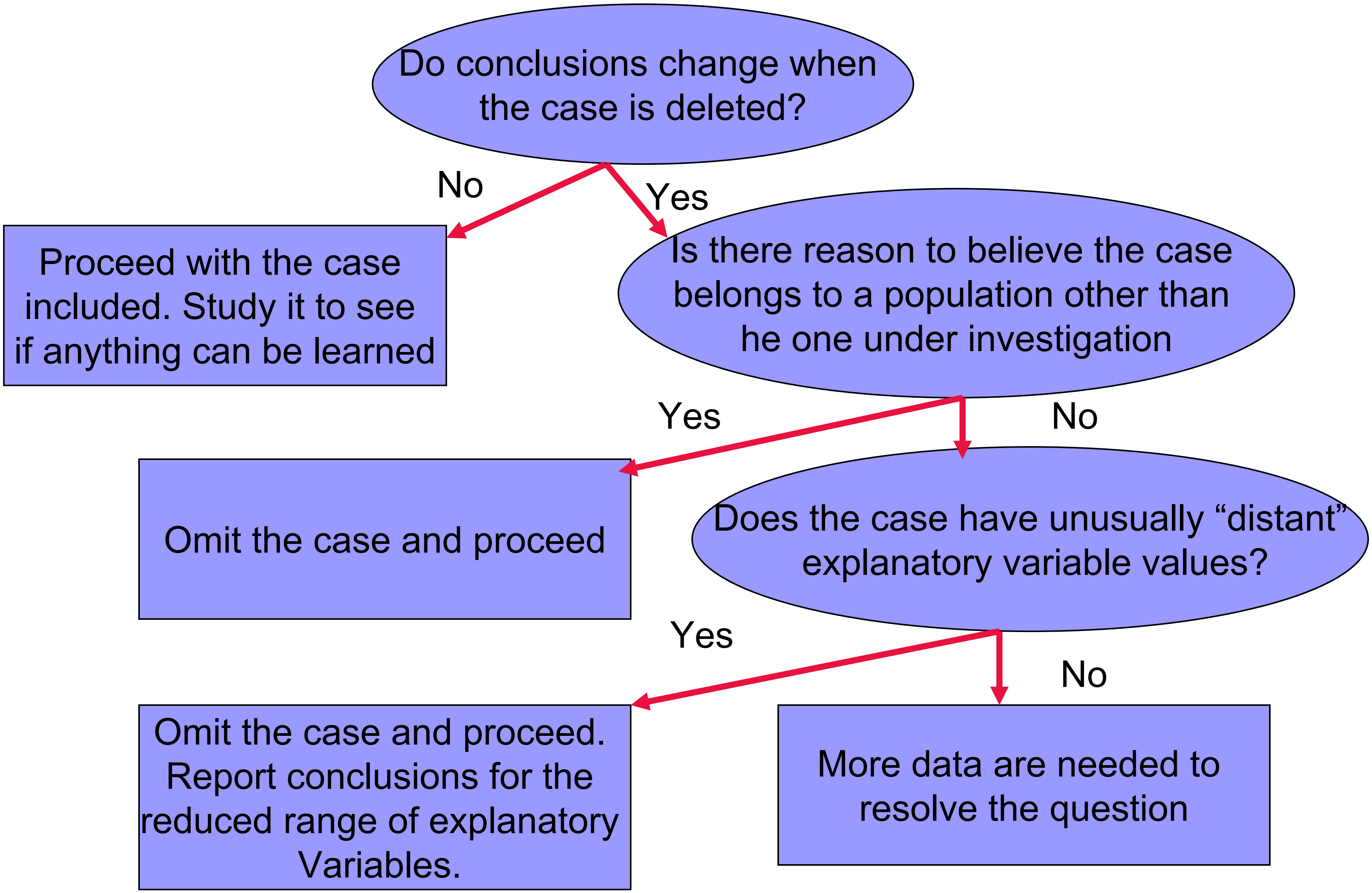
General strategy

- Start with a fairly *rich* model;
  - Include possible *x*'s even if you're not sure they will appear in the final model
  - Be careful about this with small sample sizes
- Resolve *influence* and transformation simultaneously, early in the data analysis
- In complicated problems, be prepared for dead ends.

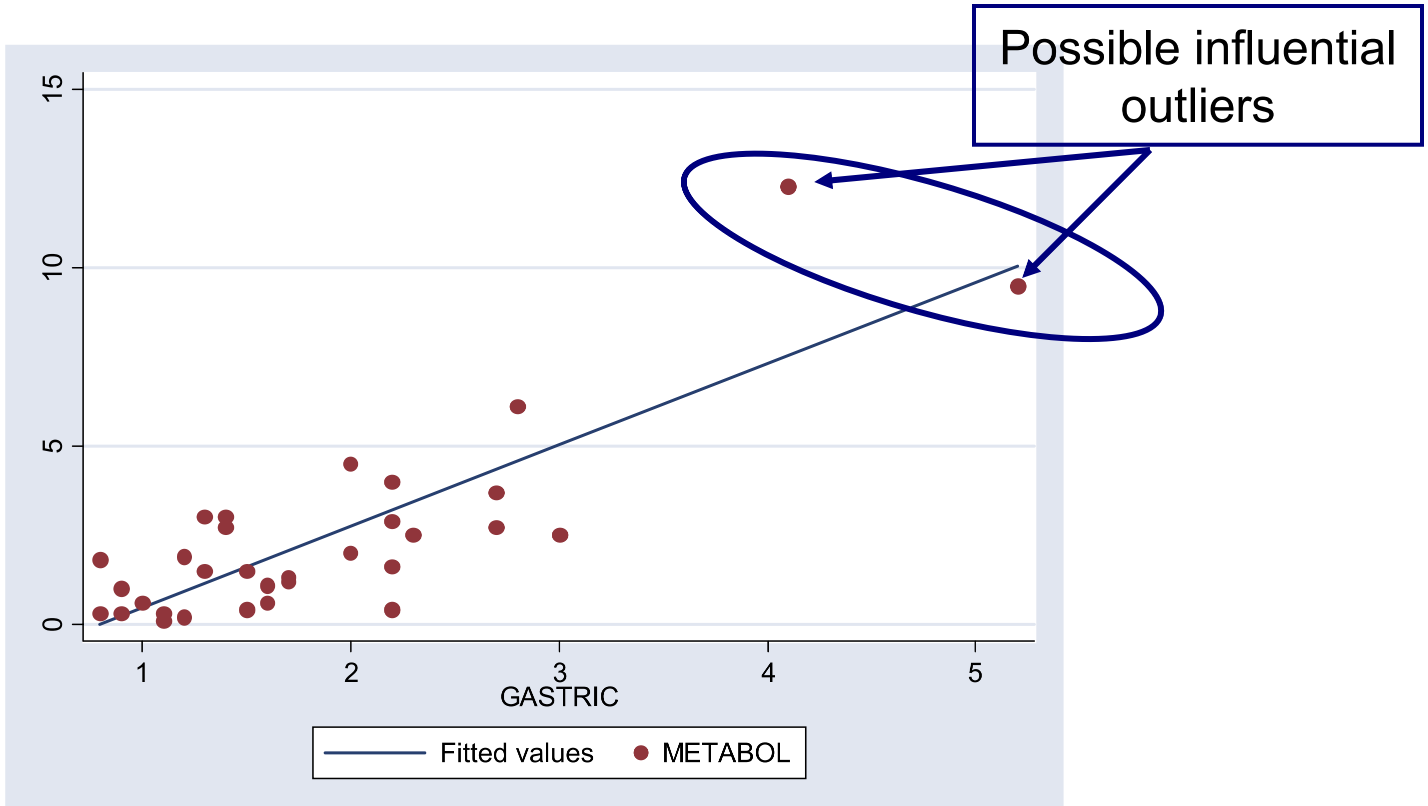
Influence

- By *influential observation(s)* we mean one or several observations whose removal causes a different conclusion in the analysis.
- Two strategies for dealing with the fact that least squares is not resistant:
  - Use an estimating procedure that is more resistant than least squares (and don’t worry about the influence problem)
  - Use least squares with the strategy defined below...

A strategy for dealing with influential cases



Alcohol Metabolism Example



- Does the fitted regression model change when the two isolated points are removed?

Example: Alcohol Metabolism

- Step 1: Create indicator variables and Interactive terms.
  - STATA commands to generate dummies for male and female:
    - gen female=gender if gender==1
      (14 missing values generated)
    - gen male=gender if gender==2
      (18 missing values generated)
    - replace female=0 if female!=1
      (14 real changes made)
    - replace male=0 if male!=2
      (18 real changes made)
  - Interactive Term
    - gen femgas=female*gastric

Example: Alcohol Metabolism (cont.)

Step 2: run initial regression model:

```
. reg metabol female gastric femgas

      Source |       SS       df       MS              Number of obs =      32
-------------+------------------------------           F(  3,    28) =   40.77
       Model |   178.28201     3  59.4273367           Prob > F      =  0.0000
    Residual |  40.8126802    28  1.45759572           R-squared     =  0.8137
-------------+------------------------------           Adj R-squared =  0.7938
       Total |   219.09469    31  7.06757066           Root MSE      =  1.2073

------------------------------------------------------------------------------
     metabol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |    .988497   1.072391     0.92   0.365    -1.208197     3.18519
     gastric |   2.343871    .280148     8.37   0.000     1.770014    2.917729
      femgas |  -1.506924   .5591376    -2.70   0.012    -2.652265   -.3615823
       _cons |  -1.185766   .7116847    -1.67   0.107    -2.643586    .2720539
------------------------------------------------------------------------------
```

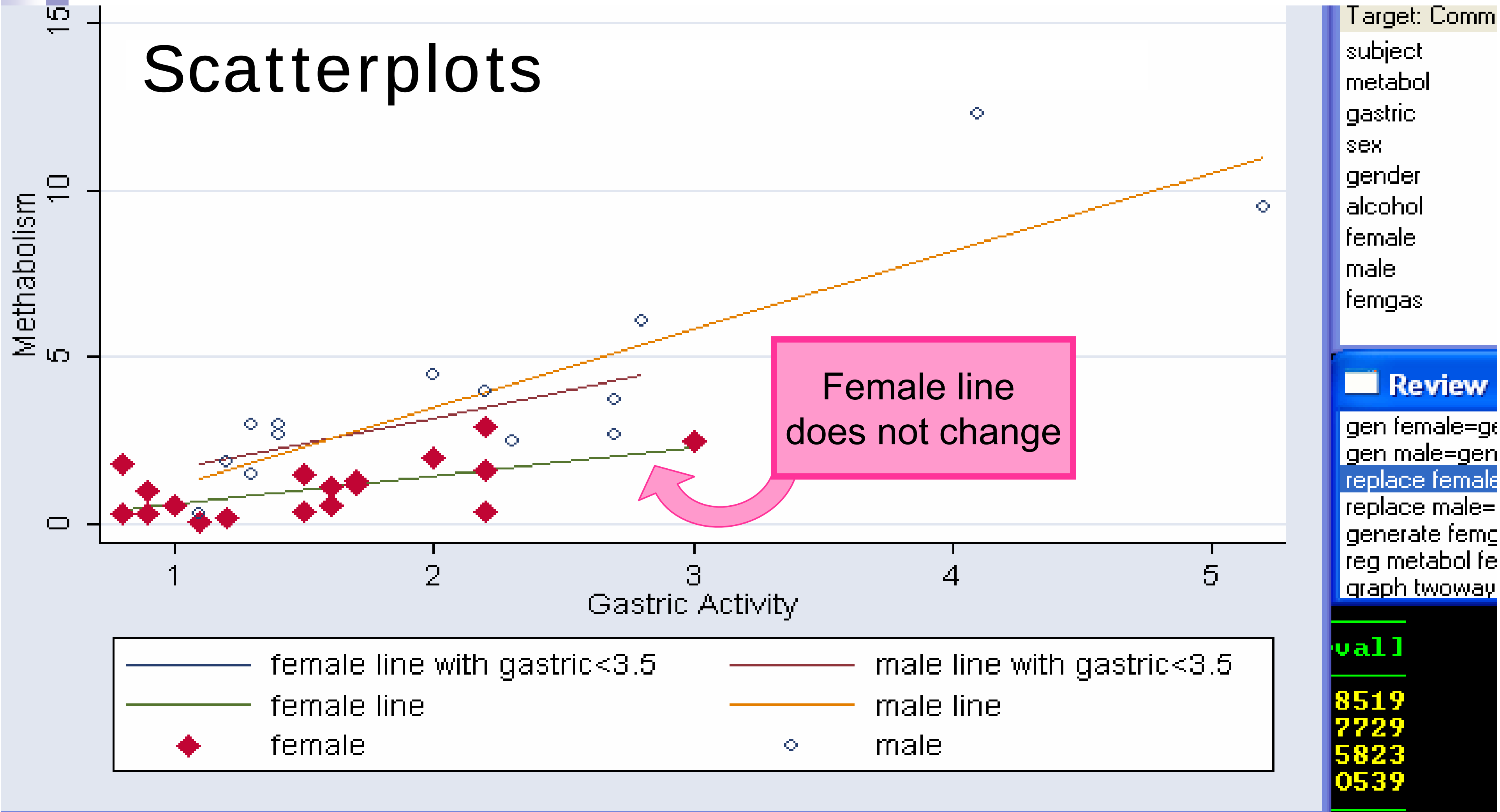
Example: Alcohol Metabolism (cont.)

Step 3: run initial regression model:

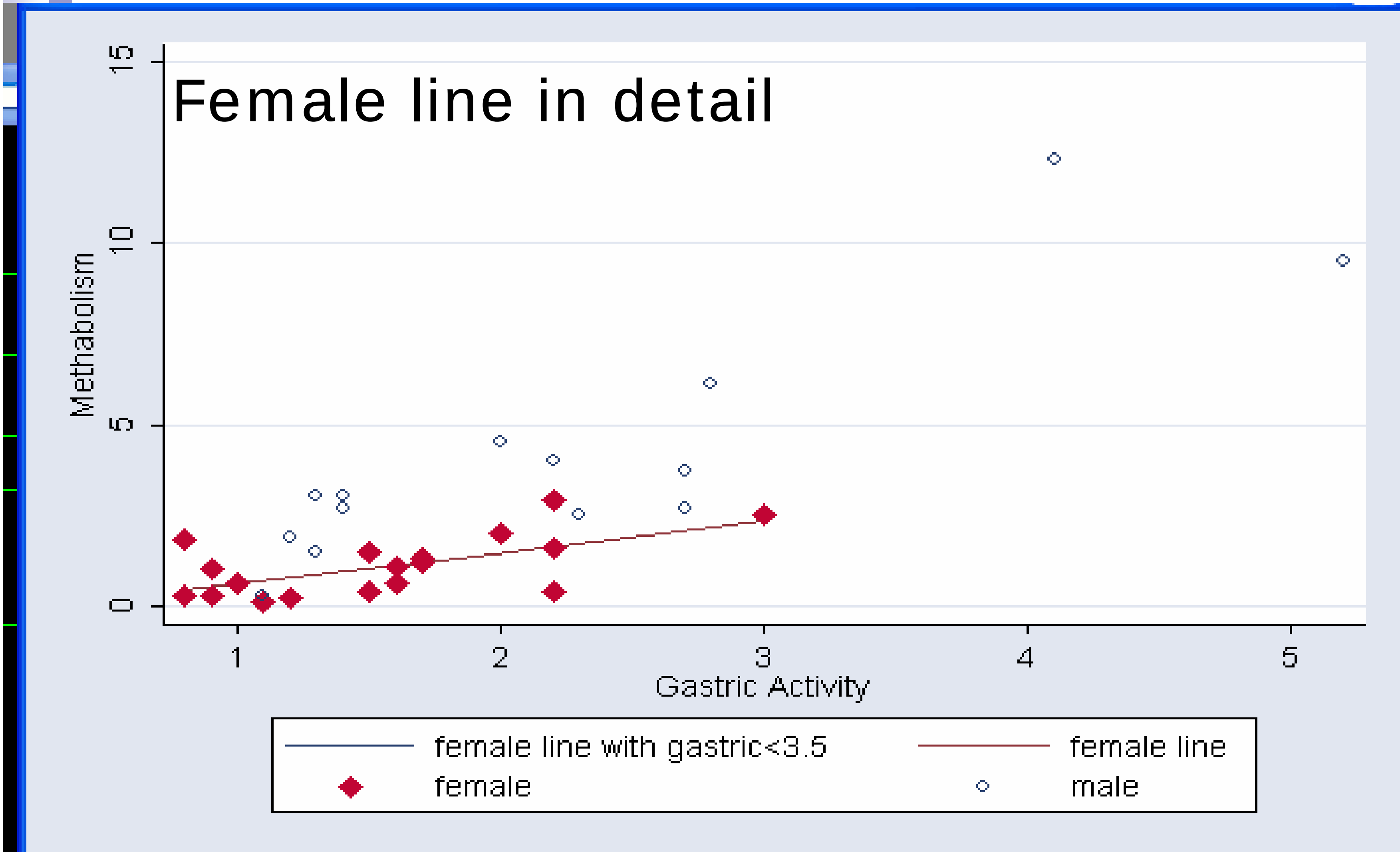
exclude the largest values of gastric, cases 31 and 32

```
. reg metabol female gastric femgas if gastric<3.5

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  3,    26) =   17.83
       Model |  41.6100636     3  13.8700212           Prob > F      =  0.0000
    Residual |  20.2236025    26  .777830864           R-squared     =  0.6729
-------------+------------------------------           Adj R-squared =  0.6352
       Total |  61.8336661    29  2.13219538           Root MSE      =  .88195

------------------------------------------------------------------------------
     metabol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2667927   .9932437    -0.27   0.790    -2.308434    1.774849
     gastric |   1.565434   .4073902     3.84   0.001     .7280313    2.402836
      femgas |   -.728486   .5393695    -1.35   0.188    -1.837176     .380204
       _cons |   .0695236   .8019484     0.09   0.932    -1.578905    1.717952
------------------------------------------------------------------------------
```

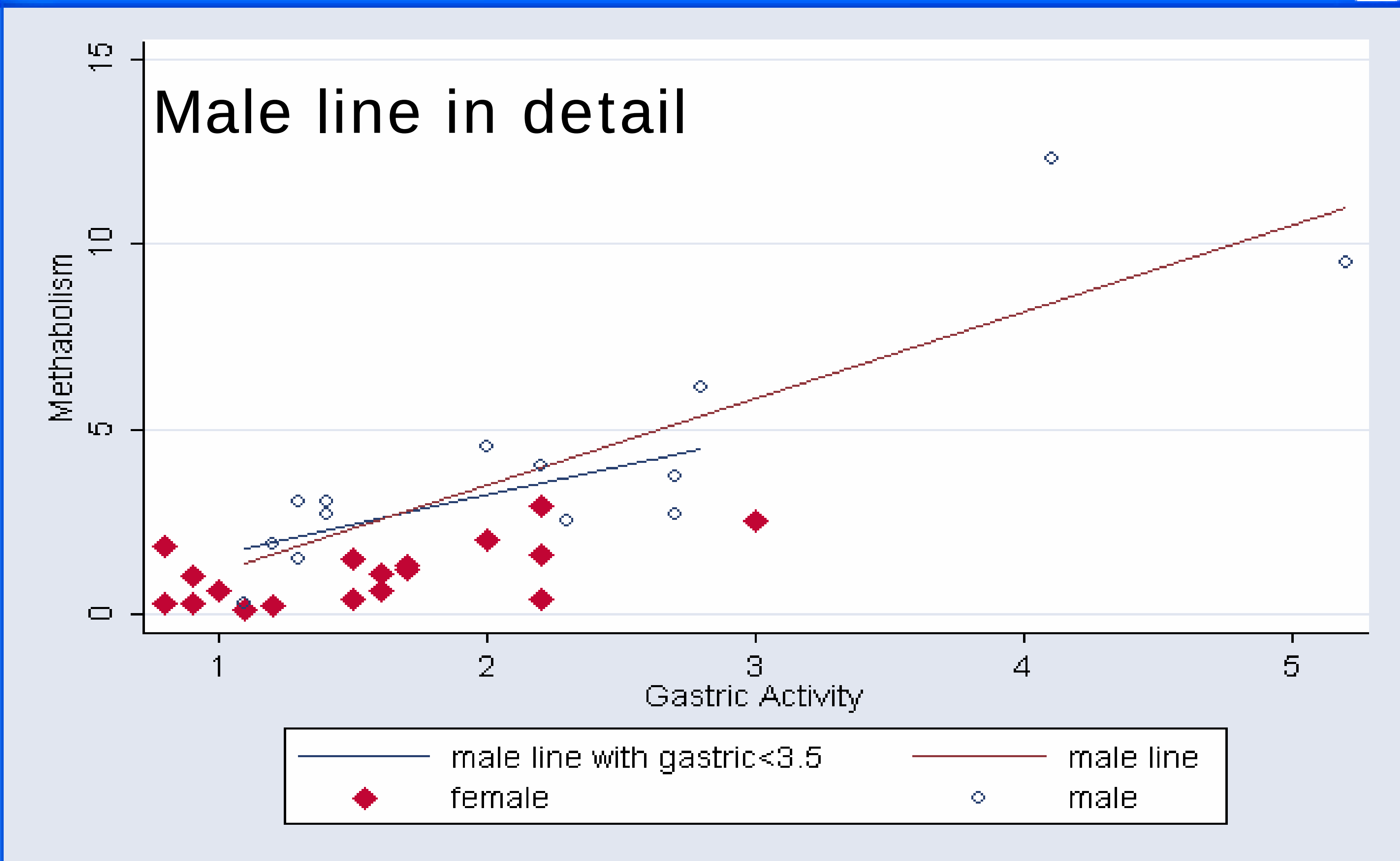
Scatterplots




```
. graph twoway lfit metabol gastric if female==1 & gastric<=3.5 || lfit metabol gas
> tric if female==0 & gastric<=3.5 || lfit metabol gastric if female==1 || lfit met
> abol gastric if female==0 || scatter metabol gastric if female==1, msymbol(D) mco
> lor(cranberry) || scatter metabol gastric if female==0, msymbol(Oh) mcolor(navy)
> legend(label(1 "female line with gastric<3.5") label(2 "male line with gastric<3.
> 5") label(3 "female line") label(4 "male line") label(5 "female") label(6 "male")
> ) ytitle("Methabolism") xtitle("Gastric Activity")
```




```
graph twoway lfit metabol gastric if female==1 & gastric<=3.5 || lfit metabol gastric if female==1 || scatter metabol gastric if female==1, msymbol(D) mcolor(cranberry) || scatter metabol gastric if female==0, msymbol(Oh) mcolor(navy) legend(label(1 "female line with gastric<3.5") label(2 "female line") label(3 "female") label(4 "male")) ytitle("Methabolism") xtitle("Gastric Activity")
```



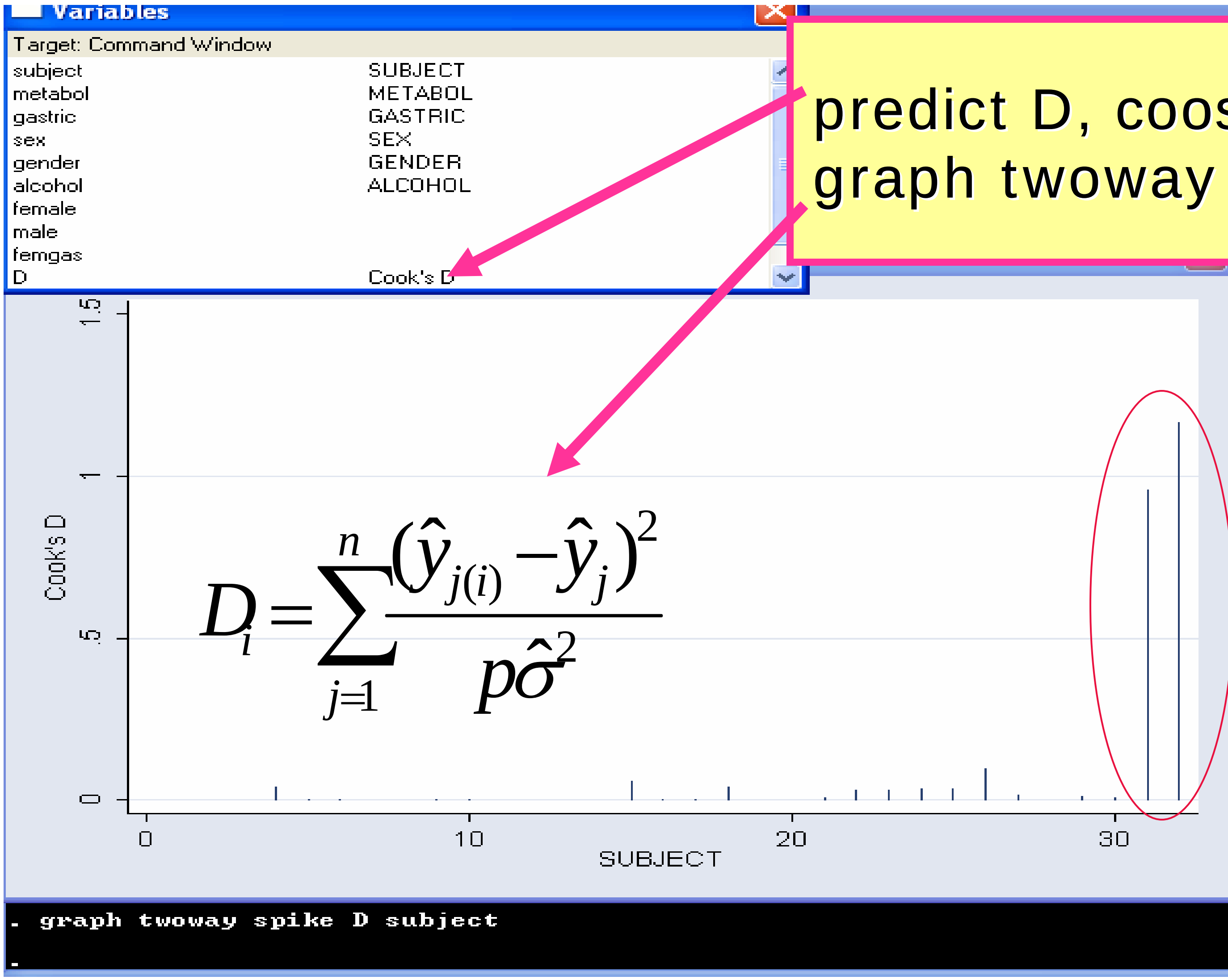

```
graph twoway lfit metabol gastric if female==0 & gastric<=3.5 || lfit metabol gastric if female==0 || scatter metabol gastric if female==1, msymbol(D) mcolor(cranberry) || scatter metabol gastric if female==0, msymbol(Oh) mcolor(navy) legend(label(1 "male line with gastric<3.5") label(2 "male line") label(3 "female") label(4 "male")) ytitle("Methabolism") xtitle("Gastric Activity")
```

Case influence statistics

- Introduction
  - These help identify influential observations and help to clarify the course of action.
  - Use them when:
    - you suspect influence problems and
    - when graphical displays may not be adequate
- One useful set of case influence statistics:
  - D_i: Cook's Distance - for measuring influence
  - h_i: Leverage - for measuring "unusualness" of x's
  - r_i: Studentized residual - for measuring "outlierness"
    - Note: $i = 1,2,\ldots, n$
- Sample use of influence statistics...

Cook's Distance: Measure of overall influence

```
predict D, cooskd
graph twoway spike D subject
```



$$D_i = \sum_{j=1}^{n} \frac{(\hat{y}_{j(i)} - \hat{y}_j)^2}{p\hat{\sigma}^2}$$

Note: observations 31 and 32 have large cooks distances.

The impact that omitting a case has on the estimated regression coefficients.

D_i: Cook's Distance for identifying influential cases

- One formula is: $$D_i = \sum_{j=1}^{n} \frac{(\hat{y}_{j(i)} - \hat{y}_j)^2}{p\hat{\sigma}^2}$$
  - Here the distance formula involves the estimated response of y at observation j, based on the reduced data set with observation i deleted and
  - included, and the number of regression coefficients
  - in estimated variance from the fit, based on all observations.
- An equivalent formula (admittedly mysterious) is:

$$D_i = \frac{1}{p}(studres_i)^2 \left(\frac{h_i}{1 - h_i} \right)$$

This term is big if case *i* is unusual in the y-direction

This term is big if case i is unusual in the x-direction

In the homework, you are asked to show the two formulae are equivalent.

Leverage: h_i for the single variable case

(also called: diagonal element of the hat matrix)

- It measures the multivariate distance between the x's for case i and the average x's, accounting for the correlation structure.

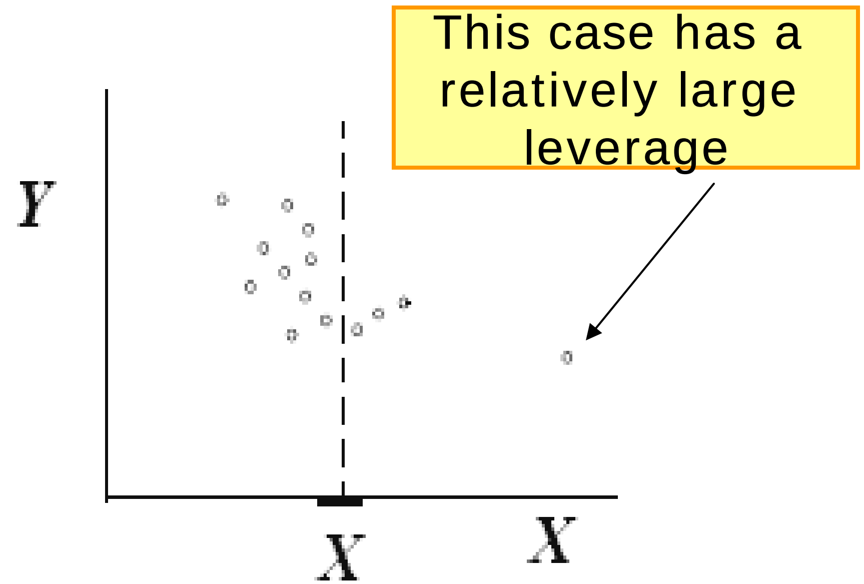
If there is only one x:

$$h_i = \frac{1}{(n-1)}\left(\frac{x_i - \bar{x}}{s_x}\right)^2 + \frac{1}{n}$$

Equivalently:

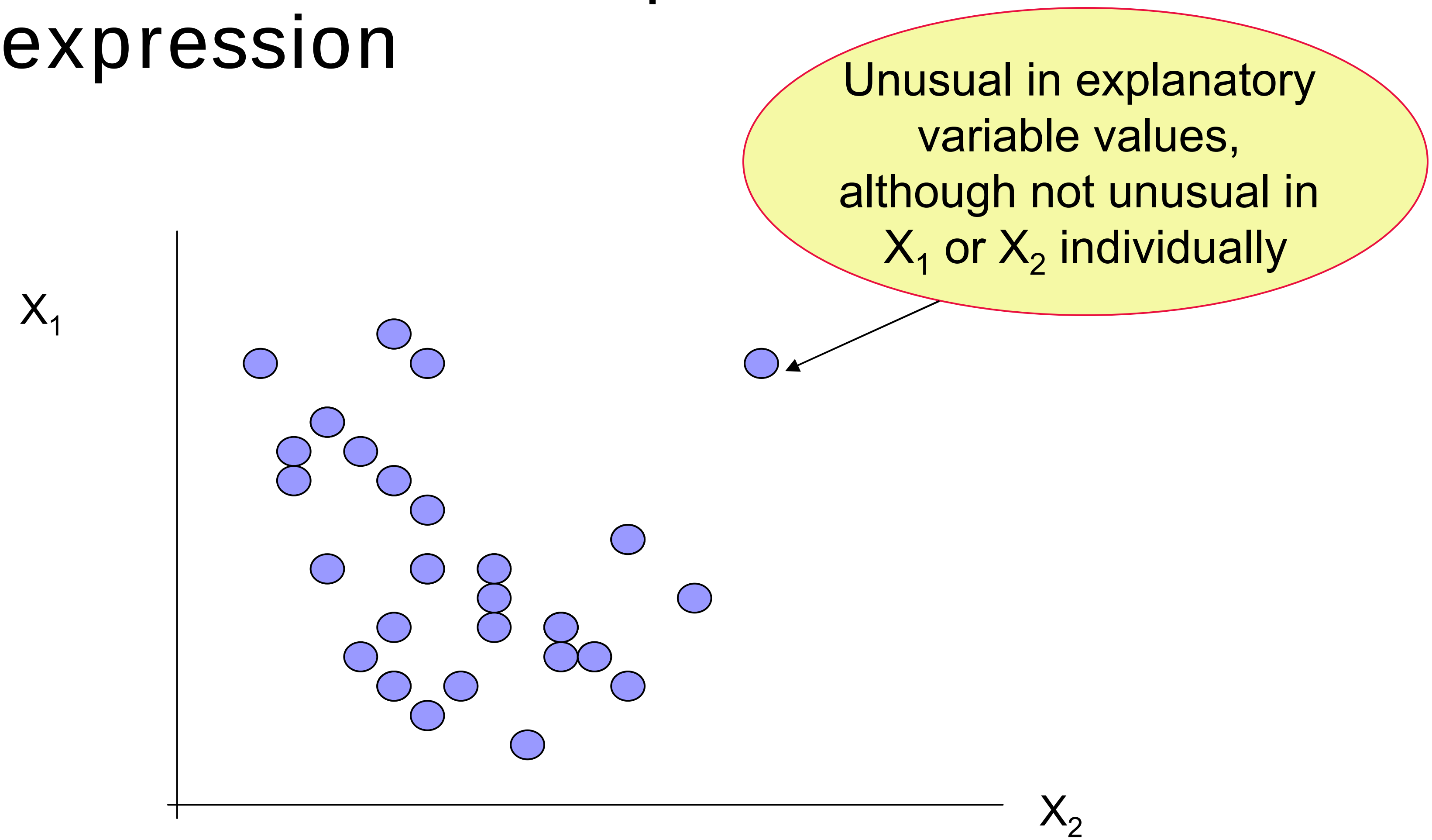
$$h_i = \frac{(x_i - \bar{x})^2}{\sum (x - \bar{x})^2} + \frac{1}{n}$$

Leverage is the proportion of the total sum of squares of the explanatory variable contributed by the i^{th} case.



Leverage: h_i for the multivariate case

- For several x's, h_i has a matrix expression



Studentized residual for detecting outliers

(in y direction)

- Formula: $$studres_i = \frac{res_i}{SE(res_i)}$$

- Fact: $$SE(res_i) = \hat{\sigma}\sqrt{1-h_i}$$

  - i.e. different residuals have different variances, and since $0 < h_i < 1$ those with largest h_i (unusual x's) have the smallest $SE(res_i)$.
  - For outlier detection use this type of residual (but use ordinary residuals in the standard residual plots).

How to use case influence statistics

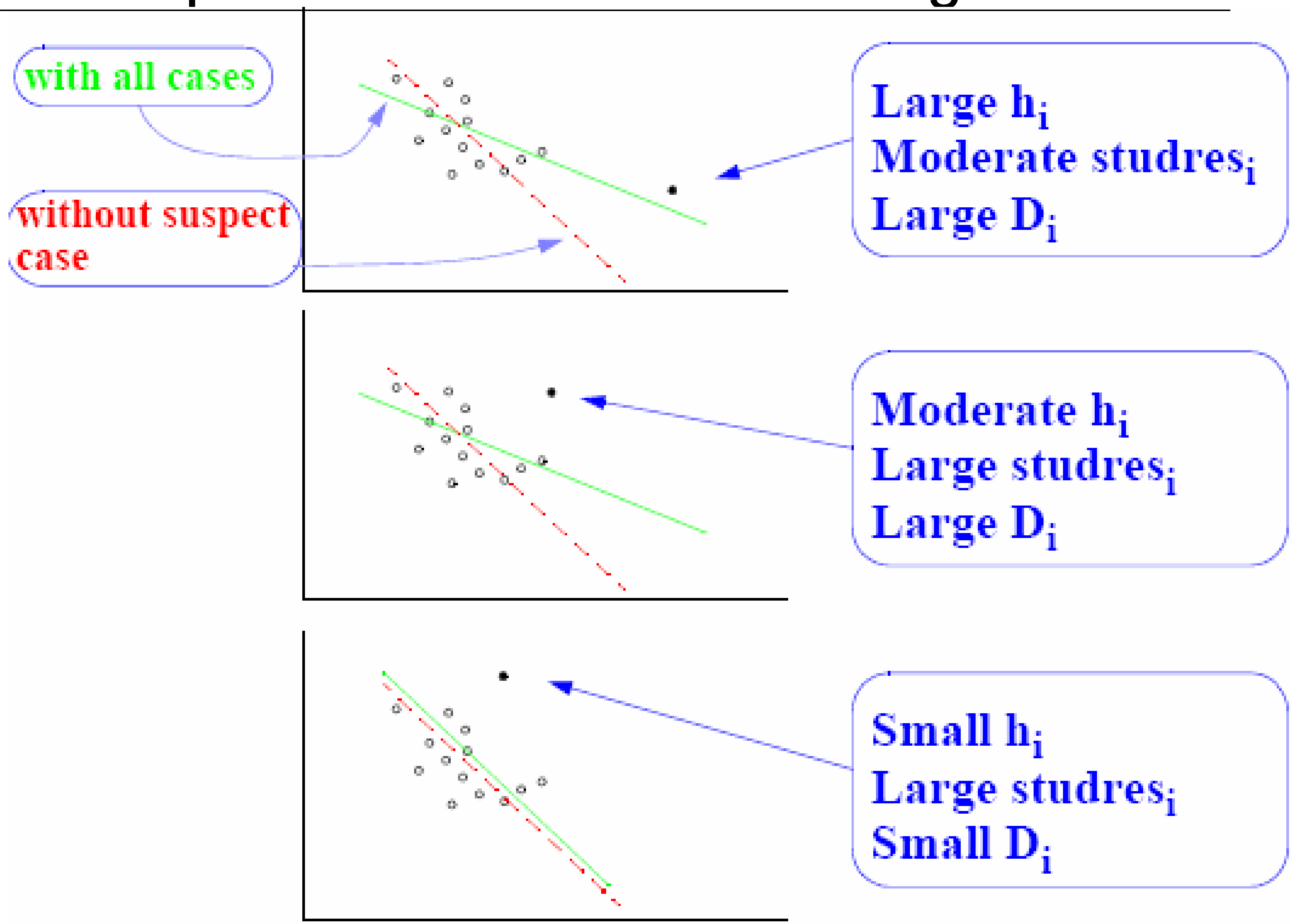
- Get the triplet (D_i, h_i, $studres_i$) for each i from 1 to n
- Look to see whether any D_i's are "large"
  - Large D_i's indicate influential observations
  - Note: you ARE allowed to investigate these more closely by manual case deletion.
- h_i and $studres_i$ help explain the reason for influence
  - unusual x-value, outlier or both;
  - helps in deciding the course of action outlined in the strategy for dealing with suspected influential cases.

ROUGH guidelines for "large"

(Note emphasis on ROUGH)

- D_i values near or larger than 1 are good indications of influential cases;
  - Sometimes a D_i much larger than the others in the data set is worth looking at.
- The average of h_i is always p/n (why?)
  - some people suggest using h_i>2p/n as "large"
- Based on normality, $|studres_i| > 2$ is considered "large"

Sample situation with a single x



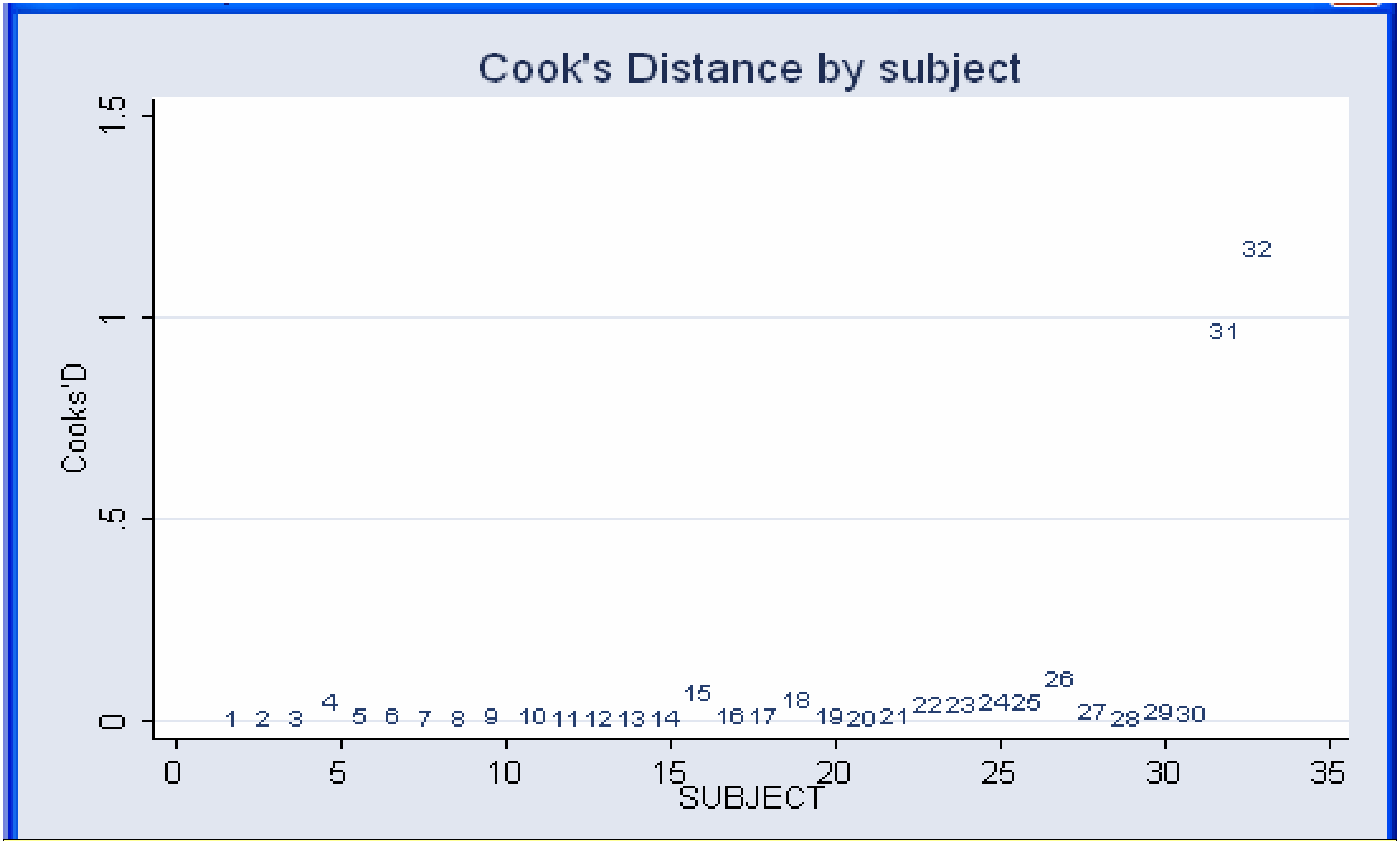
STATA commands:

- **`predict`** derives statistics from the most recently fitted model.
- Some **`predict`** options that can be used after anova or regress are:

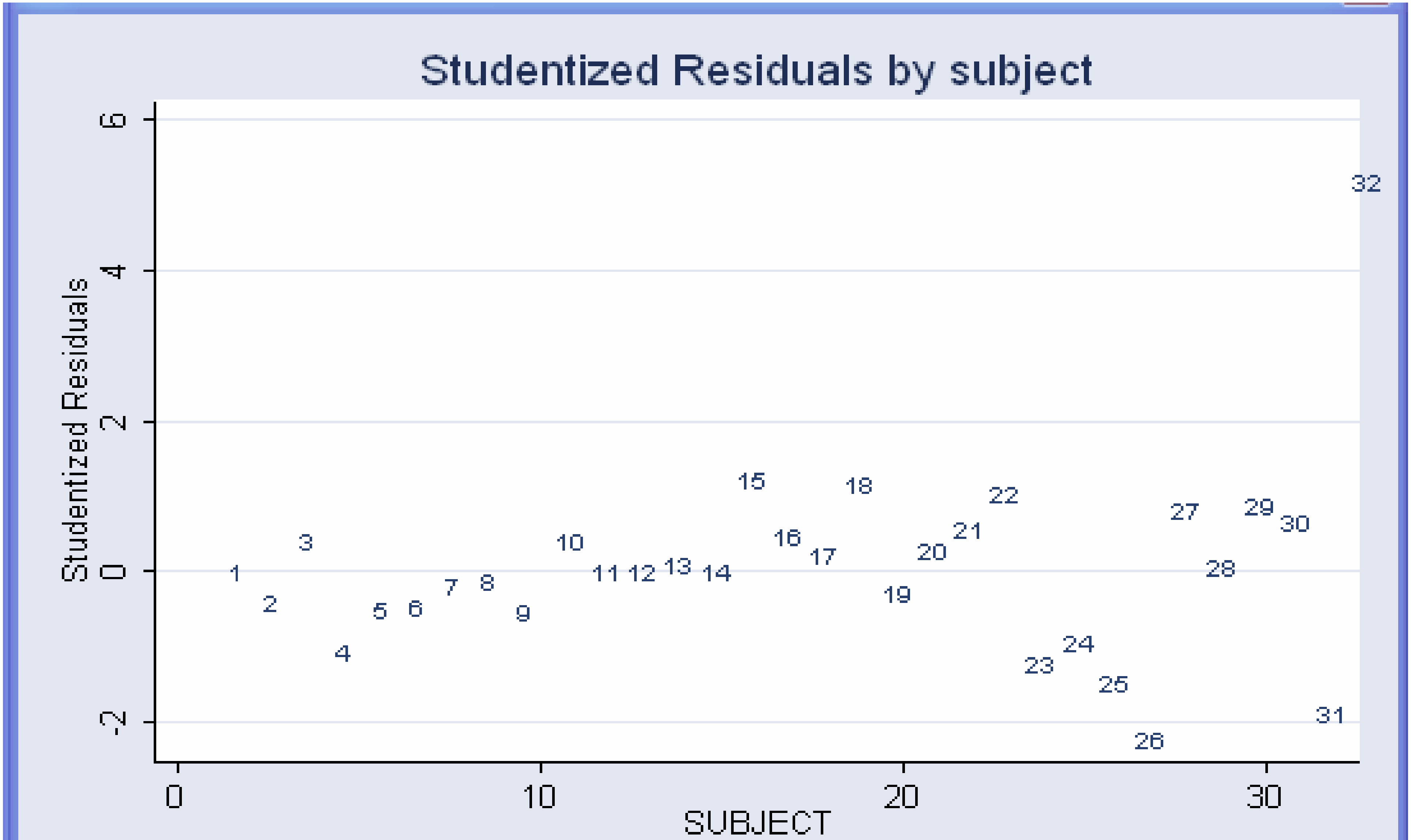
`predict newvariable, cooksd`	Cook's distance
`predict newvariable, rstudent`	Studentized residuals
`Predict newvariable, hat`	Leverage

`Predict newvariable, r` residuals

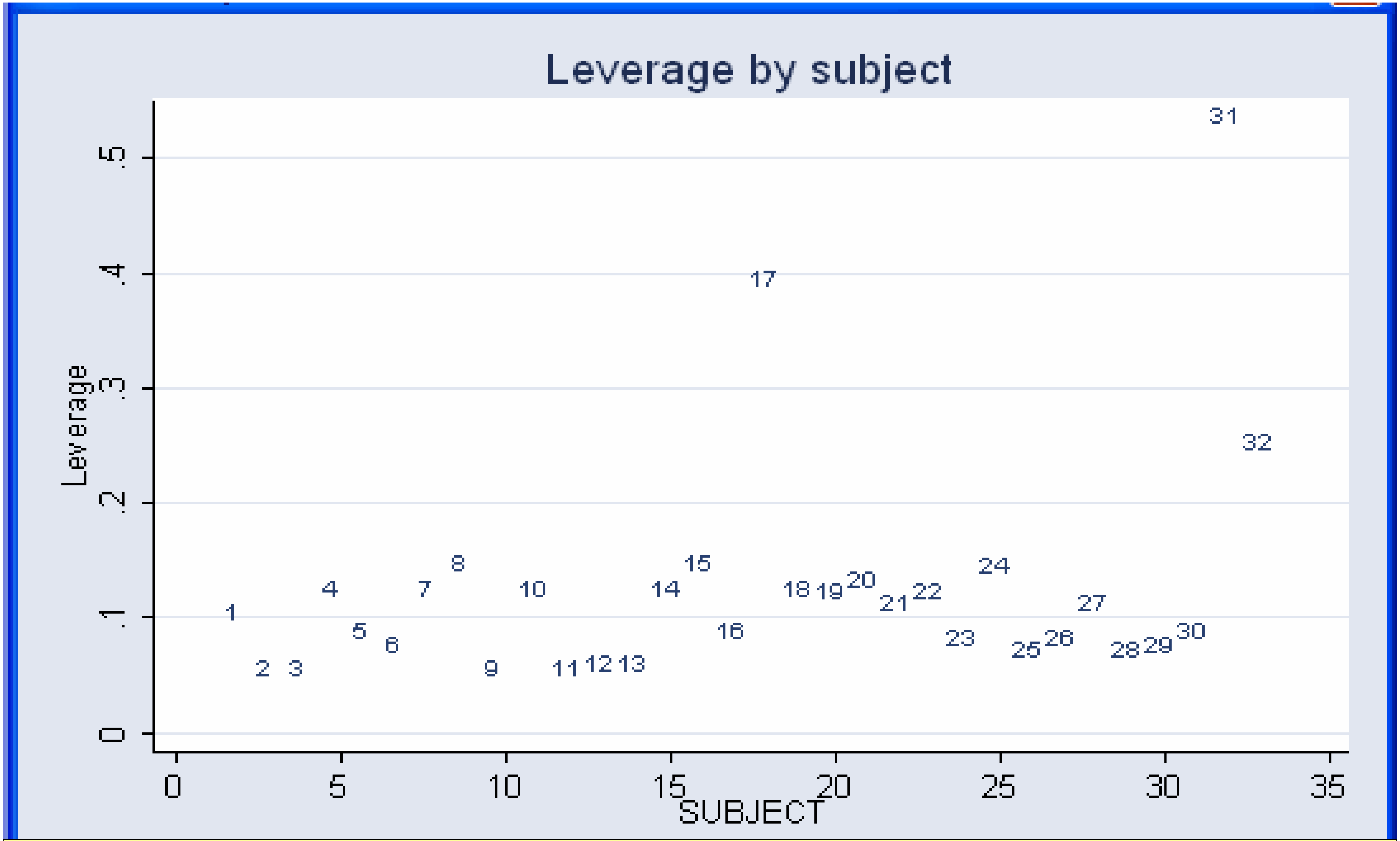
`In each case, you provide name of the newvariable to store the statistic`




```
1. predict D, cooksd
2. graph twoway scatter D subject, msymbol(i) mlabel(subject
   ytitle("Cooks'D") xlabel(0(5)35) ylabel(0(0.5)1.5)
   title("Cook's Distance by subject")
```




```
1. predict studres, rstudent
2. graph twoway scatter studres subject, msymbol(i)
   mlabel(subject) ytitle("Studentized Residuals")
   title("Studentized Residuals by subject")
```




```
1. predict leverage, hat
2. graph twoway scatter leverage subject, msymbol(i)
   mlabel(subject) ytitle(“Leverage") ylabel(0(.1).5)
   xlabel(0(5)35) title(“Leverage by subject")
```

Alternative case influence statistics

- Alternative to D_i: $dffits_i$ (and others)
- Alternative to studresi: externally-studentized residual
  - Suggestion: use whatever is convenient with the statistical computer package you're using.
- Note: D_i only detects influence of single-cases; influential pairs may go undetected.

Partial Residual Plots

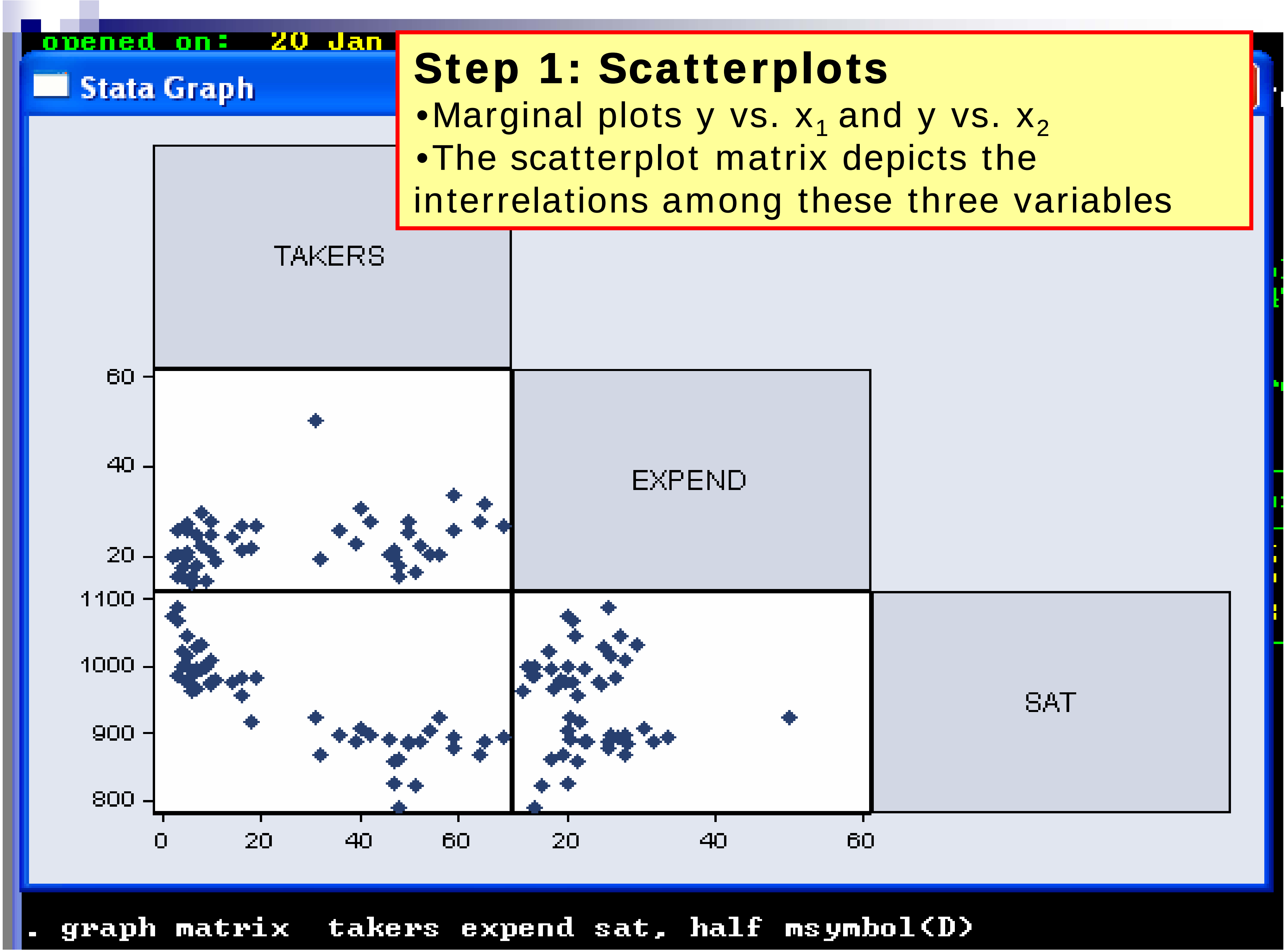
- A problem: a scatterplot of y vs x_2 gives information regarding $\mu(y|\ x_2)$ about
  - (a) whether x_2 is a useful predictor of y,
  - (b) nonlinearity in x_2 and
  - (c) outliers and influential observations.
- We would like a plot revealing (a), (b), and (c) for $\mu(y|x_1, x_2, x_3)$
  - e.g. what is the effect of x_2, after accounting for x_1 and x_3?

Example: SAT Data

- Question:
  - Is the distribution of state average SAT scores associated with state expenditure on public education, after accounting for percentage of high school students who take the SAT test?
- We would like to visually explore the function f(expend) in:
  - $\mu(\text{SAT}|\text{takers,expend}) = \beta_0 + \beta_1 \text{ takers} + f(\text{expend})$
  - After controlling for the number of students taking the test, does expenditures impact performance?

Step 1: Scatterplots

- Marginal plots y vs. x_1 and y vs. x_2
- The scatterplot matrix depicts the interrelations among these three variables

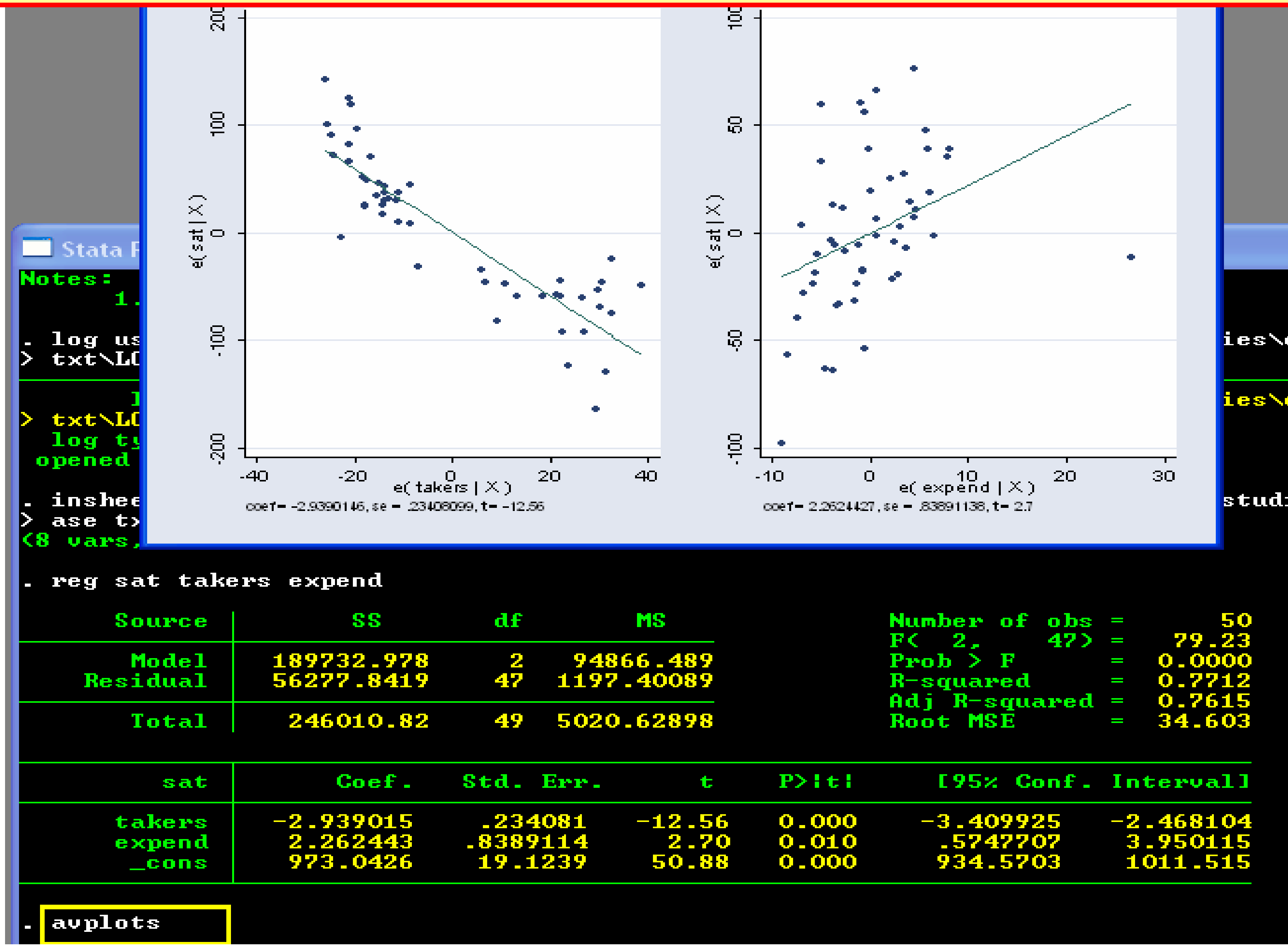


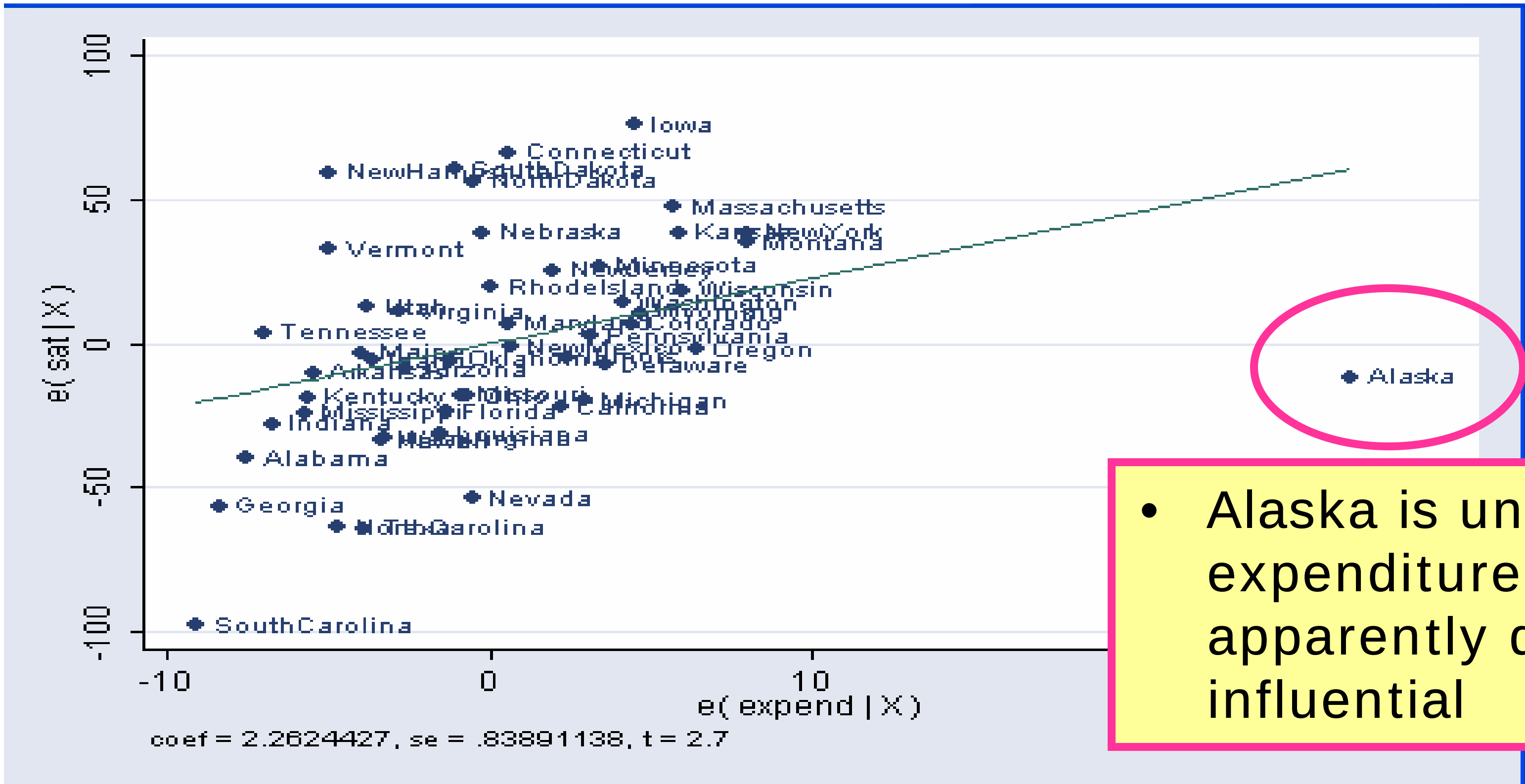
Stata Commands: avplot

- The **added variable plot** is also known as partial-regression leverage plots, adjusted partial residuals plots or adjusted variable plots.
  - The AVPlot depicts the relationship between y and one x variable, adjusting for the effects of other x variables
- **Avplots** help to uncover observations exerting a disproportionate influence on the regression model.
  - High leverage observations show in added variable plots as points horizontally distant from the rest of the data.

Added variable plots

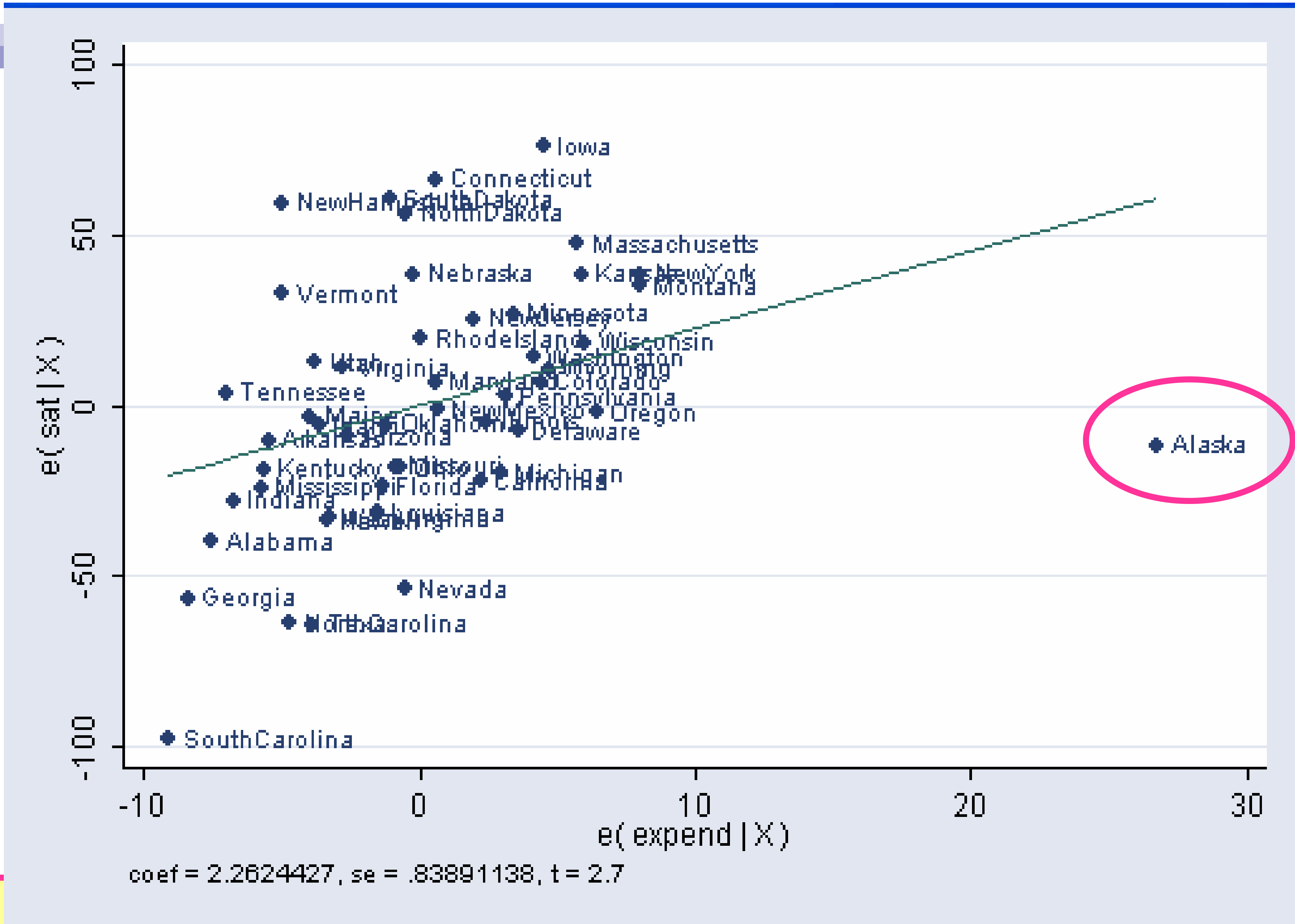
- Is the state with largest expenditure influential?
- Is there an association of expend and SAT, after accounting for takers?





- Alaska is unusual in its expenditure, and is apparently quite influential

```
reg sat takers expend

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =   79.23
       Model |  189732.978     2  94866.489            Prob > F      =  0.0000
    Residual |  56277.8419    47  1197.40089           R-squared     =  0.7712
-------------+------------------------------           Adj R-squared =  0.7615
       Total |   246010.82    49  5020.62898           Root MSE      =  34.603

------------------------------------------------------------------------------
         sat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      takers |  -2.939015    .234081   -12.56   0.000    -3.409925   -2.468104
      expend |   2.262443   .8389114     2.70   0.010     .5747707    3.950115
       _cons |   973.0426    19.1239    50.88   0.000     934.5703    1011.515
------------------------------------------------------------------------------

avplots

avplot expend, mlabel(state)
```



After accounting for % of students who take SAT, there is a positive association between expenditure and mean SAT scores.

Component plus Residual

- We'd like to plot *y* versus x_2 but with the effect of x_1 subtracted out;

  i.e. plot $y - \beta_0 + \beta_1 x_1$ versus x_2

- To approximate this, get the *partial residual for* x_2 :

  a. Get $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ in $\mu(y \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

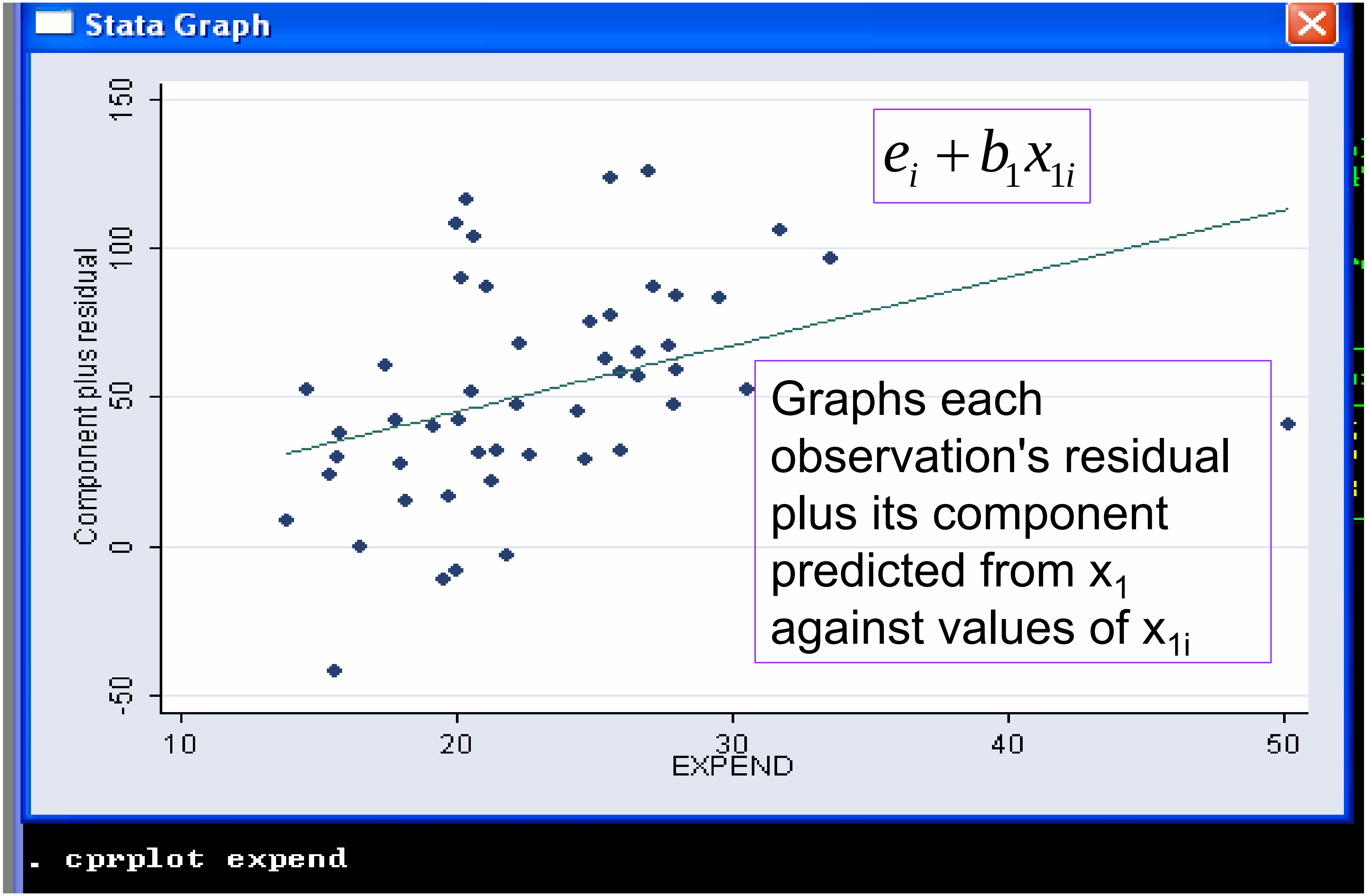
  b. Compute the partial residual as $pres = y - \hat{\beta}_0 + \hat{\beta}_1 x_1$

- This is also called a *component plus residual*; if res is the residual from 3a: $pres = res + \hat{\beta}_2 x_2$

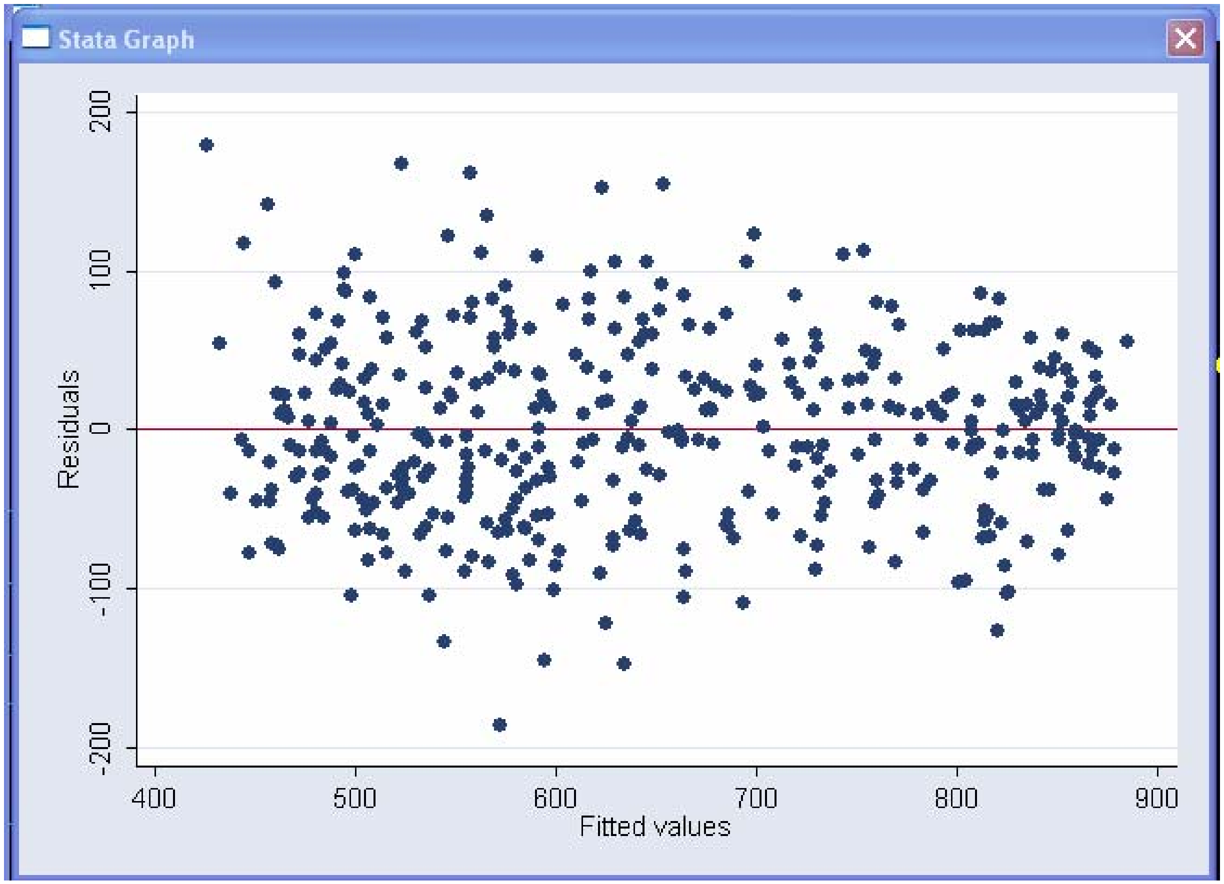
Stata Commands: cprplot

- The *component plus residual plot* is also known as partial-regression leverage plots, adjusted partial residuals plots or adjusted variable plots.
- The command “cprplot x” graph each obervation’s residual plus its component predicted from x against values of x.
- Cprplots help diagnose non-linearities and suggest alternative functional forms.

Graph cprplot x_1



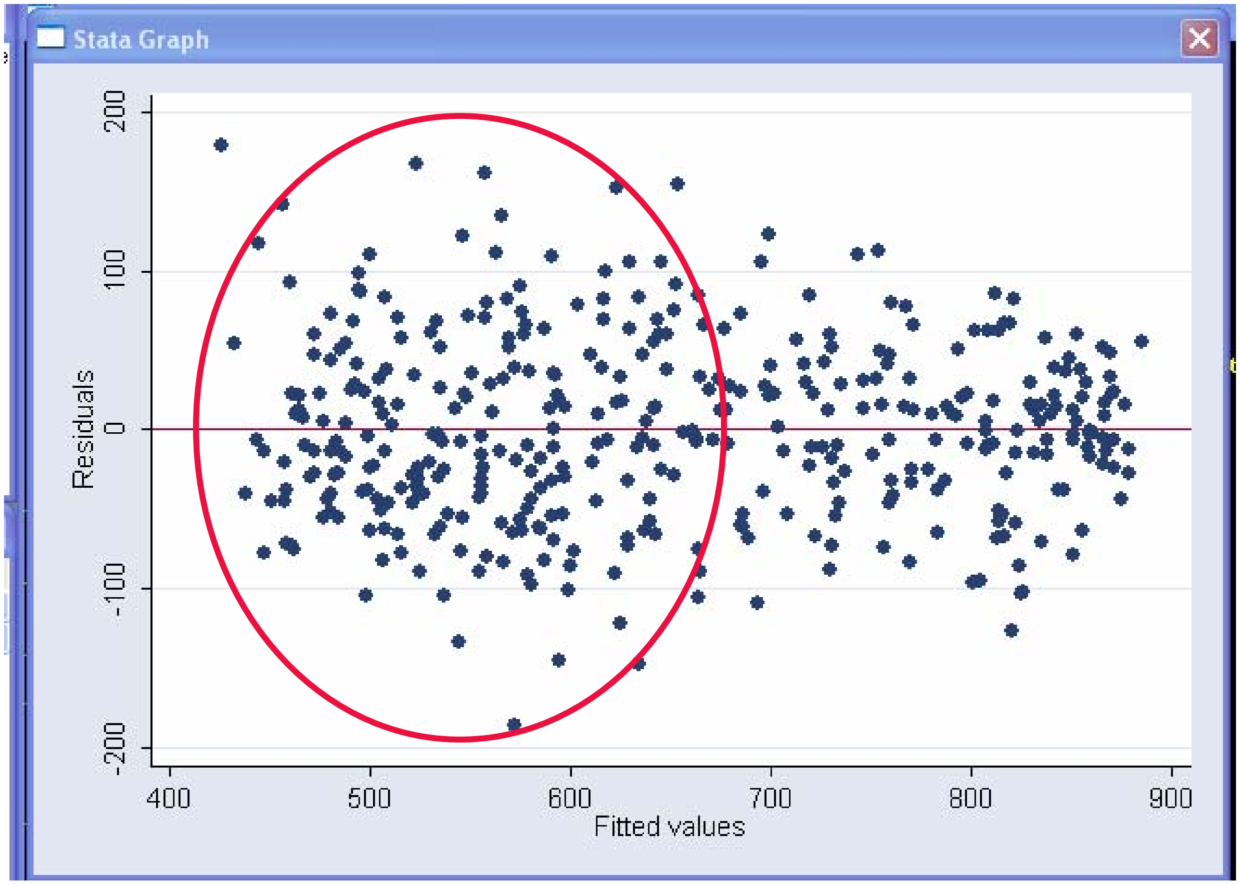
Heteroscedasticity—(Non-constant Variance)




```
reg api00 meals ell emer
rvfplot, yline(0)
```

- Heteroscedasticity is systematic variation in the size of the residuals
- Here, for instance, the variance for smaller fitted values is greater than for larger ones

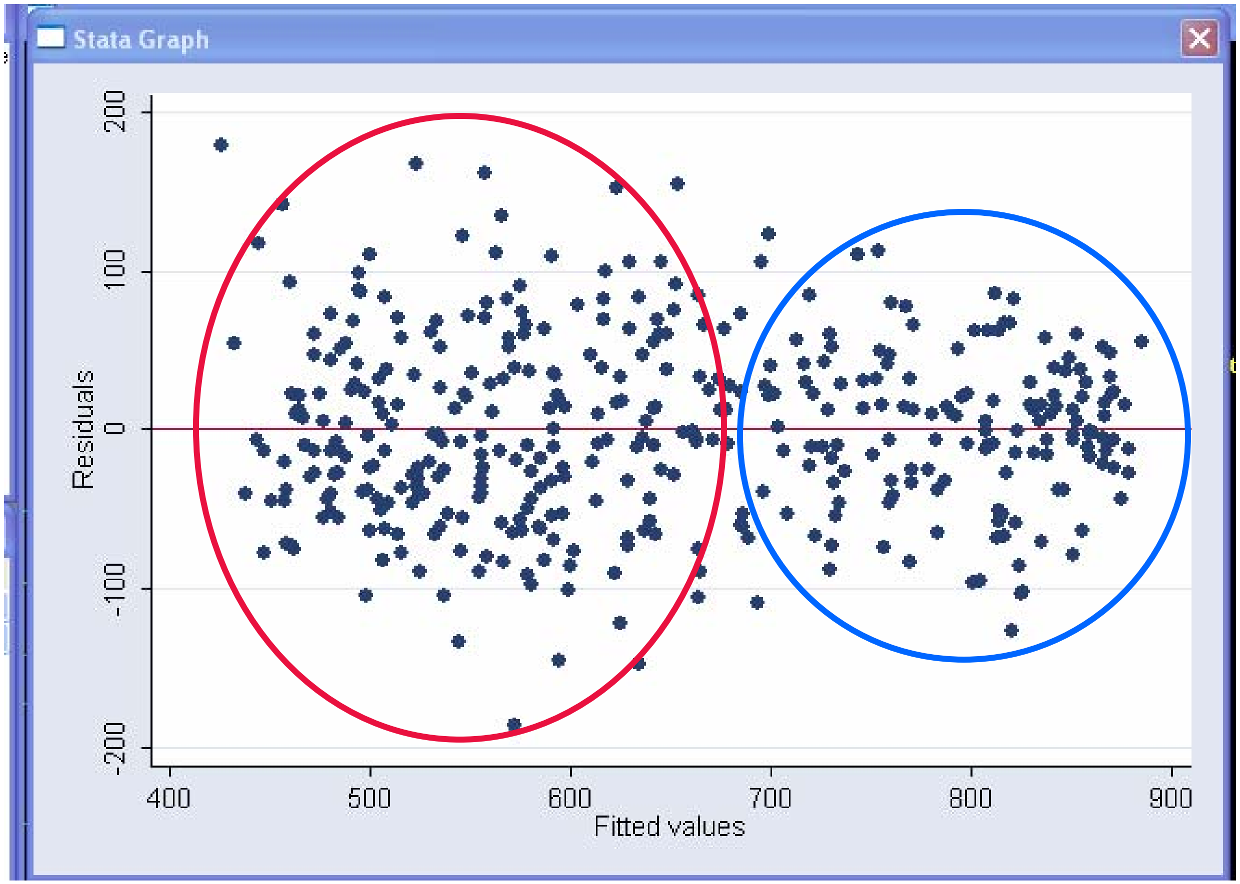
Hetroskedasticity




```
reg api00 meals ell emer
rvfplot, yline(0)
```

- Heteroskastic: Systematic variation in the size of the residuals
- Here, for instance, the variance for variance for smaller fitted values is greater than for larger ones

Hetroskedasticity




```
reg api00 meals ell emer
rvfplot, yline(0)
```

- Heteroskastic: Systematic variation in the size of the residuals
- Here, for instance, the variance for variance for smaller fitted values is greater than for larger ones

Tests for Heteroskedasticity

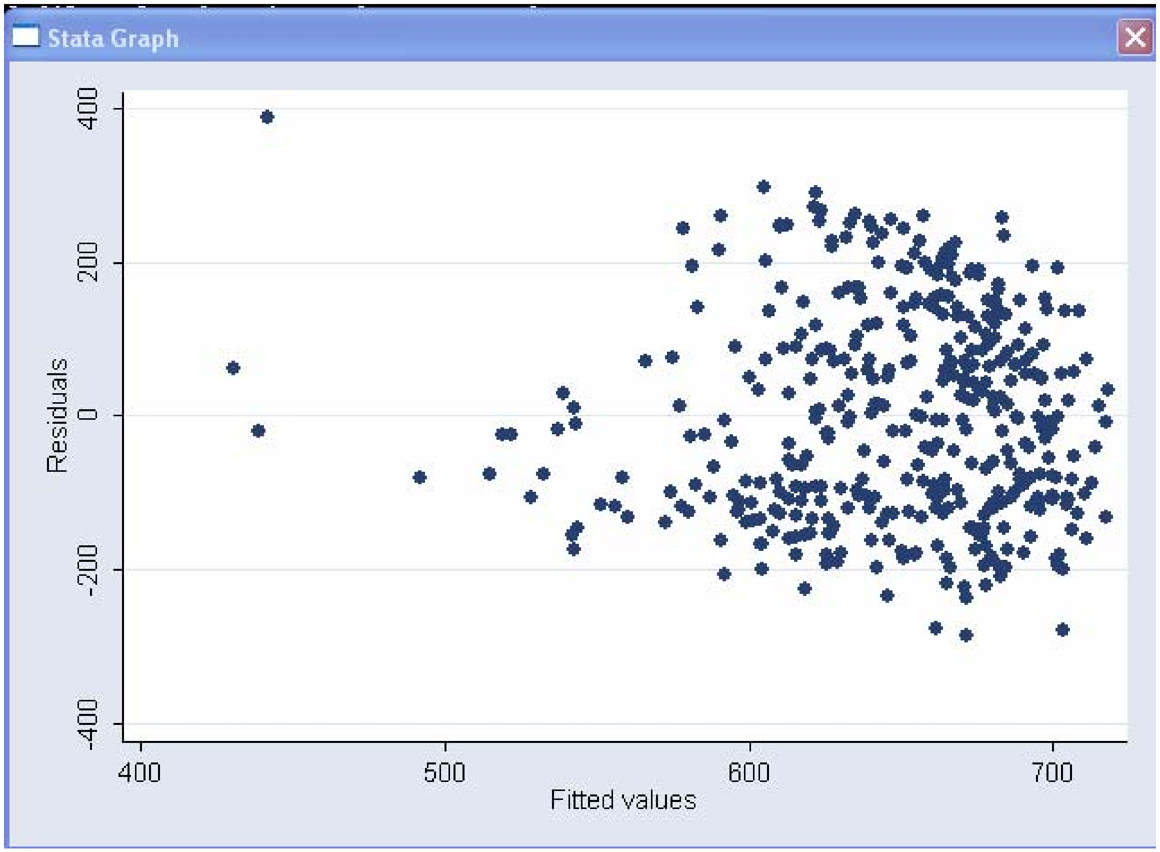
Grabbed whitetst from the web

```
. net install whitetst
checking whitetst consistency and verifying not already installed...
installing into c:\ado\plus\...
installation complete.

. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of api00

         chi2(1)      =     8.75
         Prob > chi2  =   0.0031

. whitetst

White's general test statistic :  18.35276  Chi-sq( 9)  P-value =  .0313
```

Fails hettest

Fails whitetst

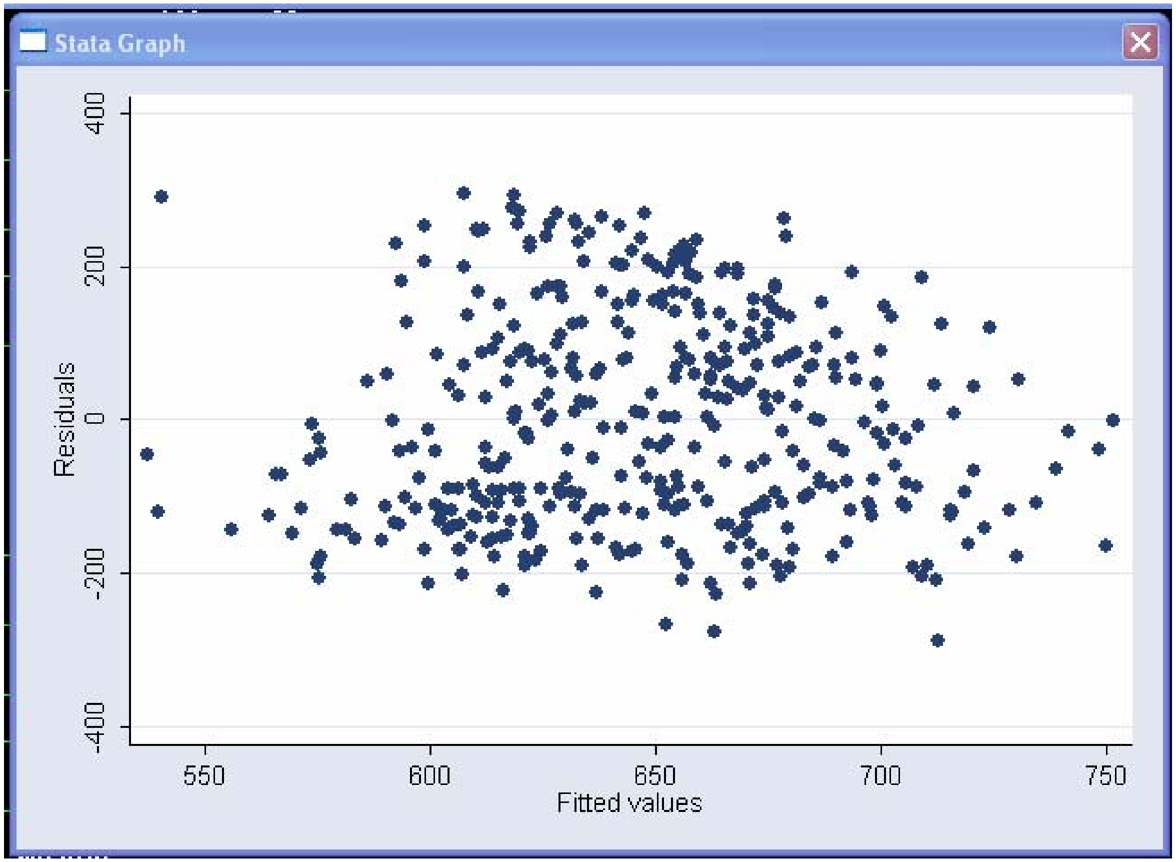
Another Example




```
reg api00 enroll
rvfplot
```

- These error terms are <u>really</u> bad!
- Previous analysis suggested logging enrollment to correct skewness

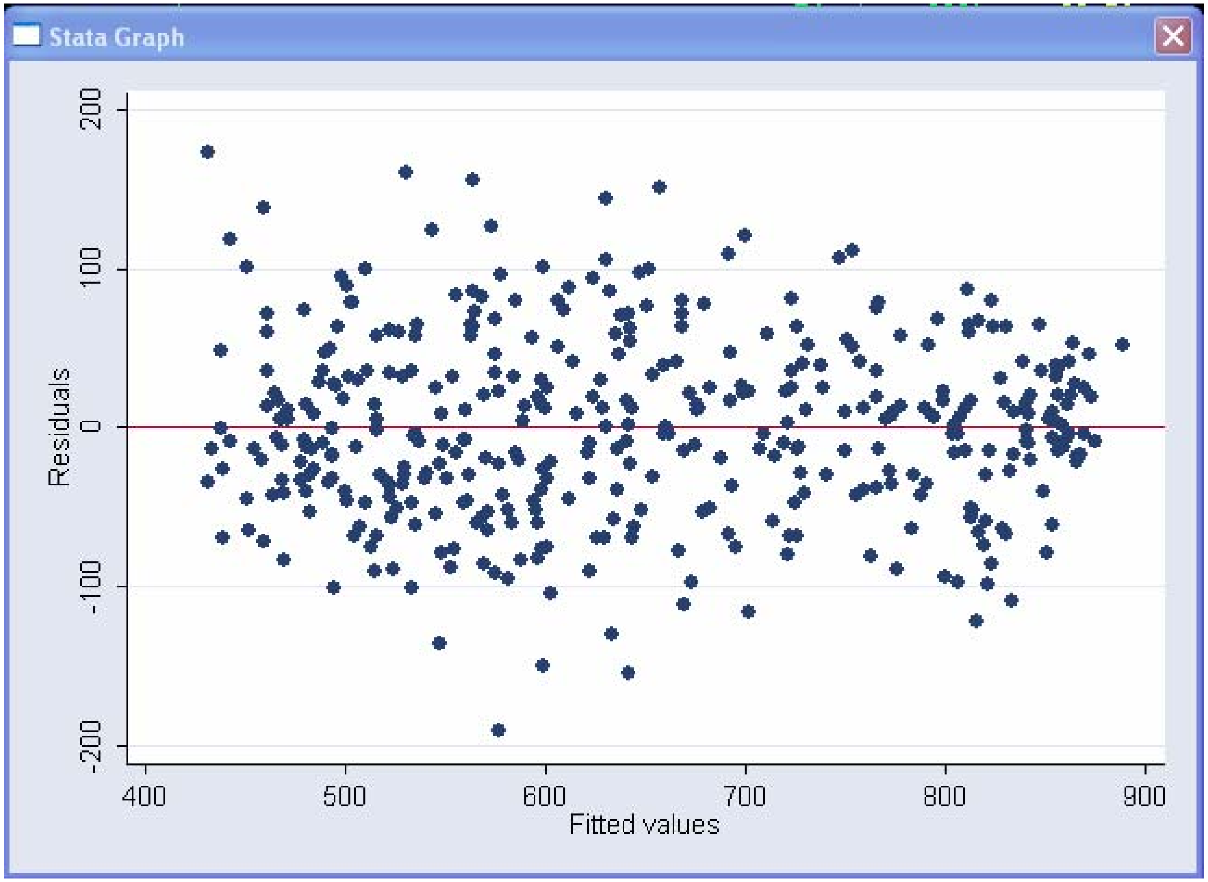
Another Example



- Much better
- Errors look more-or-less normal now

```
gen lenroll = log(enroll)
reg api00 lenroll
rvfplot
```

Back To First Example



- Adding enrollment keeps errors normal
- Don’t need to take the log of enrollment this time

```
reg api00 meals ell emer enroll
rvfplot, yline(0)
```

Weighted regression for certain types of non-constant variance (cont.)

1. Suppose: $\mu(y \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

$$\text{var}(y \mid x_1, x_2) = \sigma^2 / \omega_i$$

and the w_i's are known

2. *Weighted least squares* is the appropriate tool for this model; it minimizes the weighted sum of squared residuals

$$\sum_{i=1}^{n} \omega_i (y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2$$

3. In statistical computer programs: use linear regression in the usual way, specify the column *w* as a *weight*, read the output in the usual way

Weighted regression for certain types of non-constant variance

4. Important special cases where this is useful:

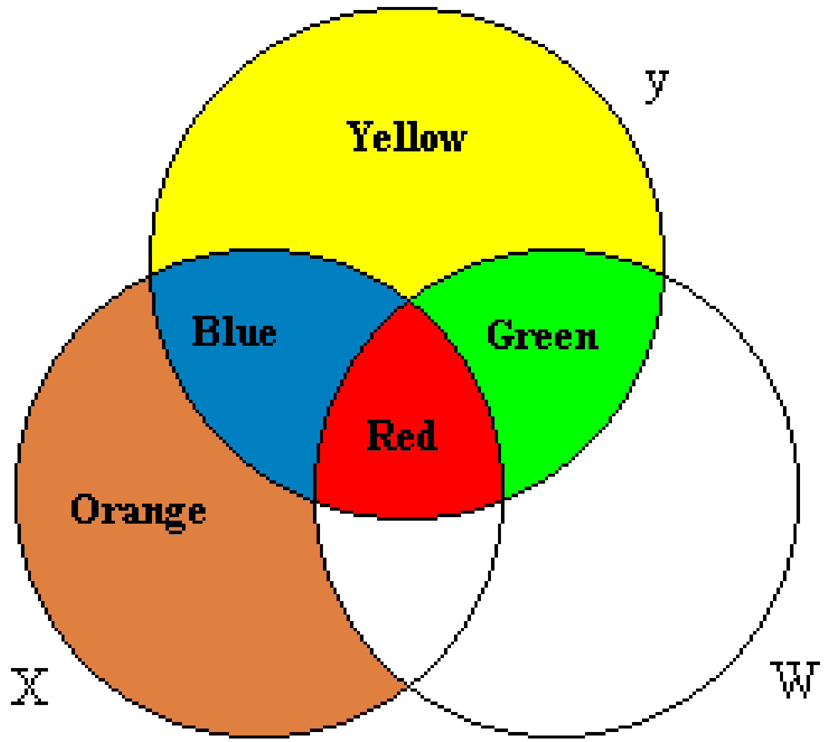
a. y_i is an average based on a sample of size m_i
In this case, the weights are $w_i = 1/m_i$

b. the variance is proportional to x; so $w_i = 1/x_i$

Multicollinearity

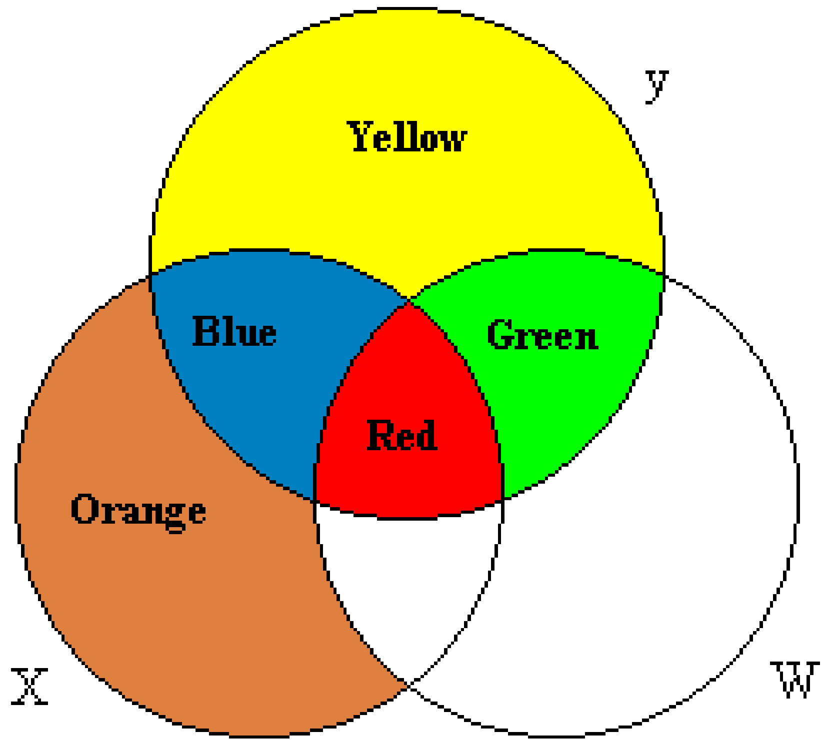
- This means that two or more regressors are highly correlated with each other.
- <u>Doesn't</u> bias the estimates of the dependent variable
  - So not a problem if all you care about is the predictive accuracy of the model
- But it <u>does</u> affect the inferences about the significance of the collinear variables
  - To understand why, go back to Venn diagrams

Multicollinearity



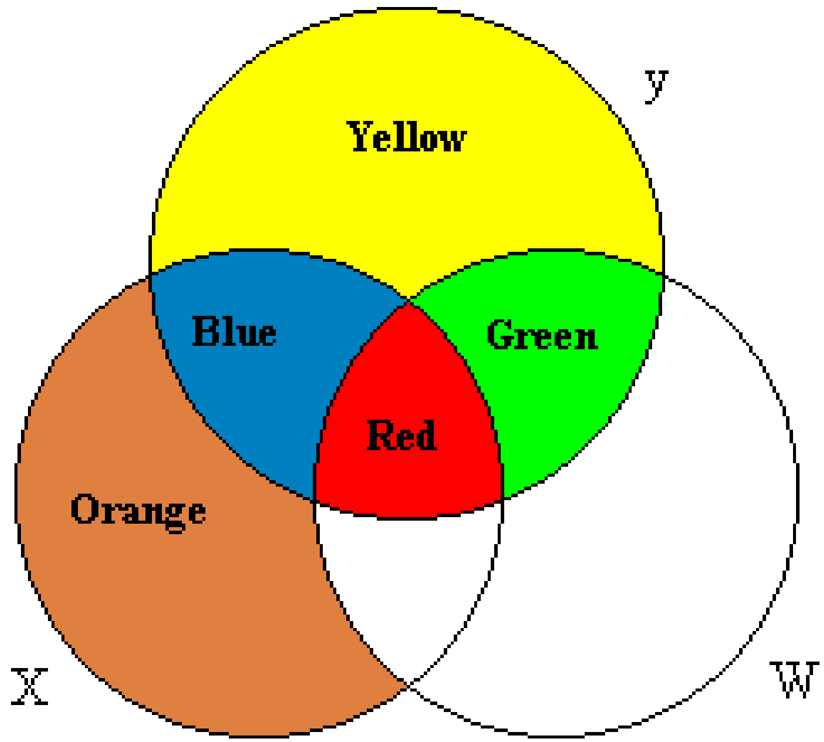
- Variable X explains Blue + Red
- Variable W explains Green + Red
- So how should Red be allocated?

Multicollinearity



- We could:
  1. Allocate Red to <u>both</u> X and W
  2. Split Red between X and W (using some formula)
  3. Ignore Red entirely

Multicollinearity

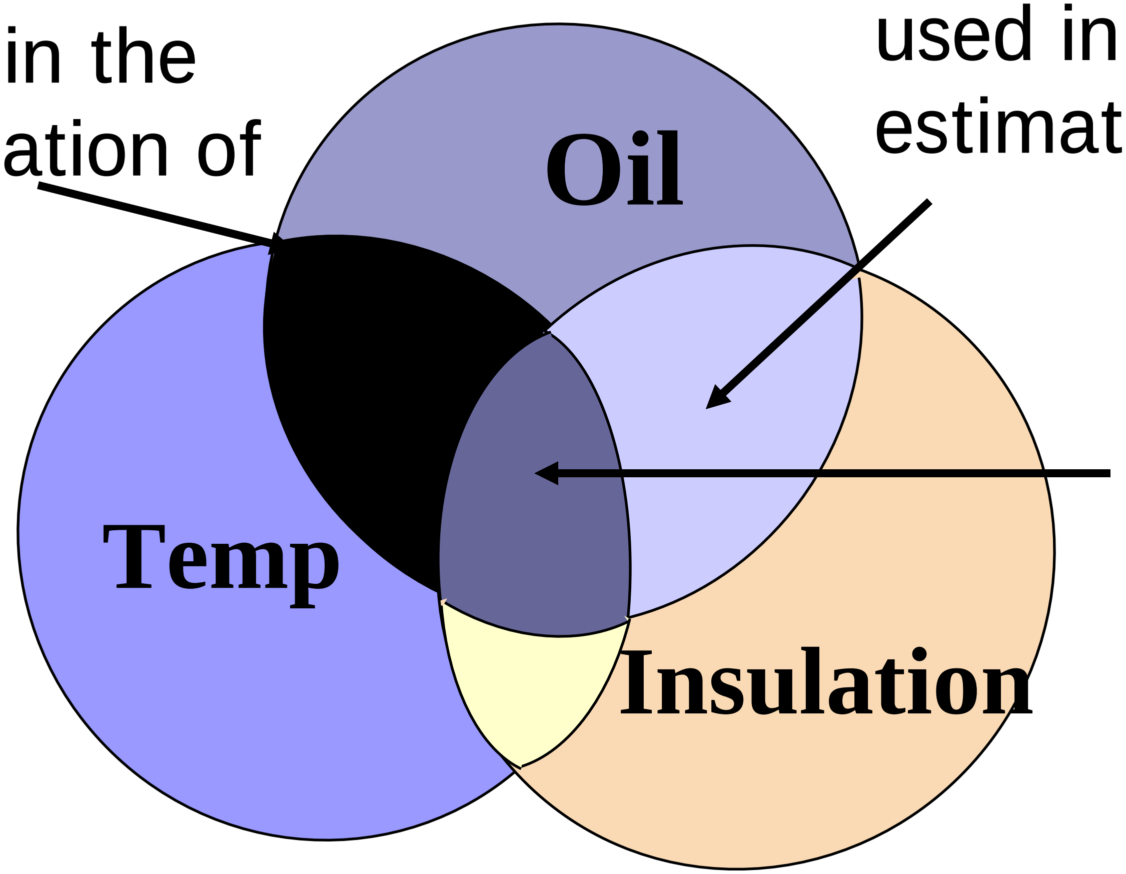


- In fact, only the information in the Blue and Green areas is used to predict Y.
- Red area is ignored when estimating β_x and β_w

Venn Diagrams and Estimation of Regression Model

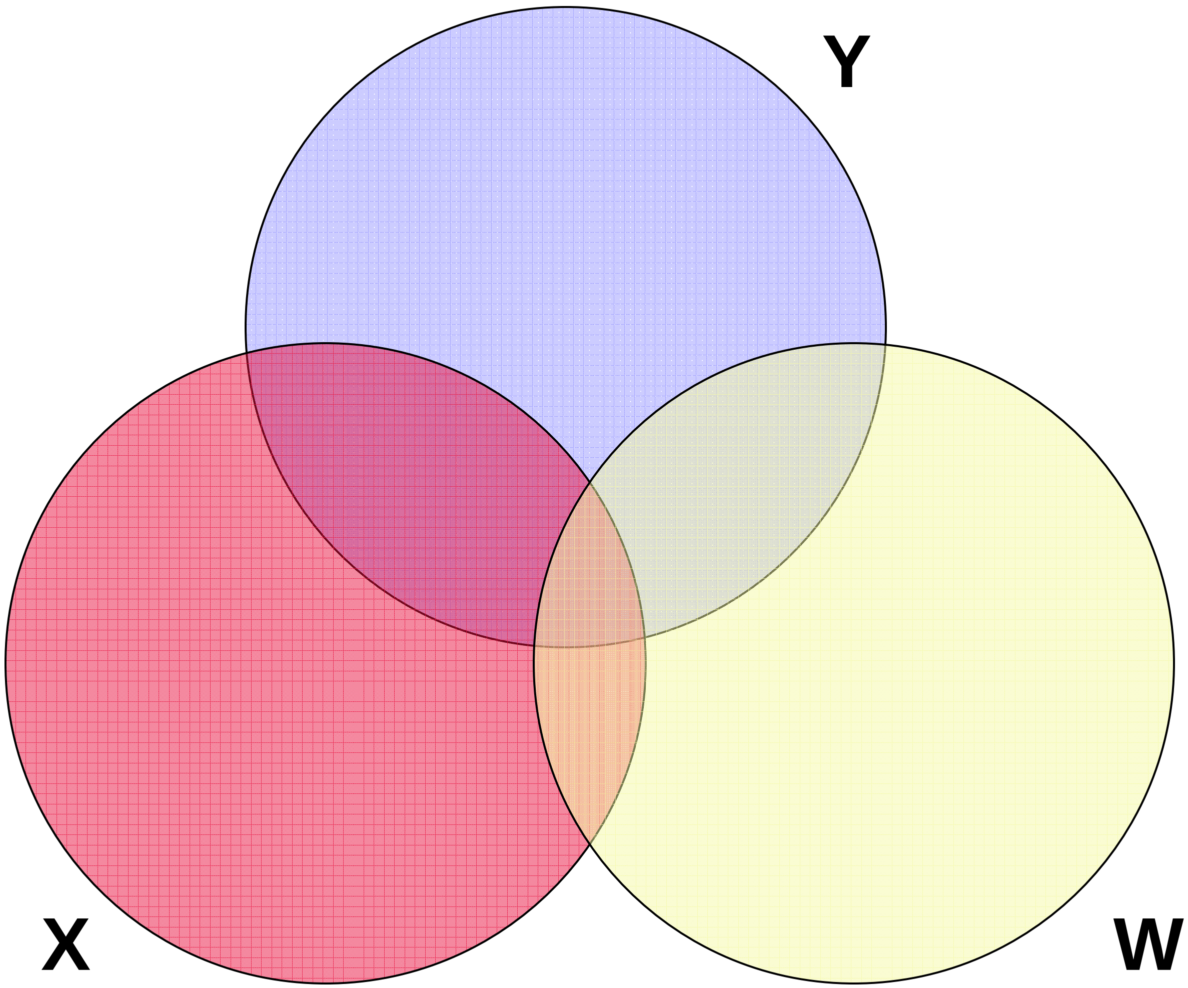
Only this information is used in the estimation of β_1

Only this information is used in the estimation of β_2



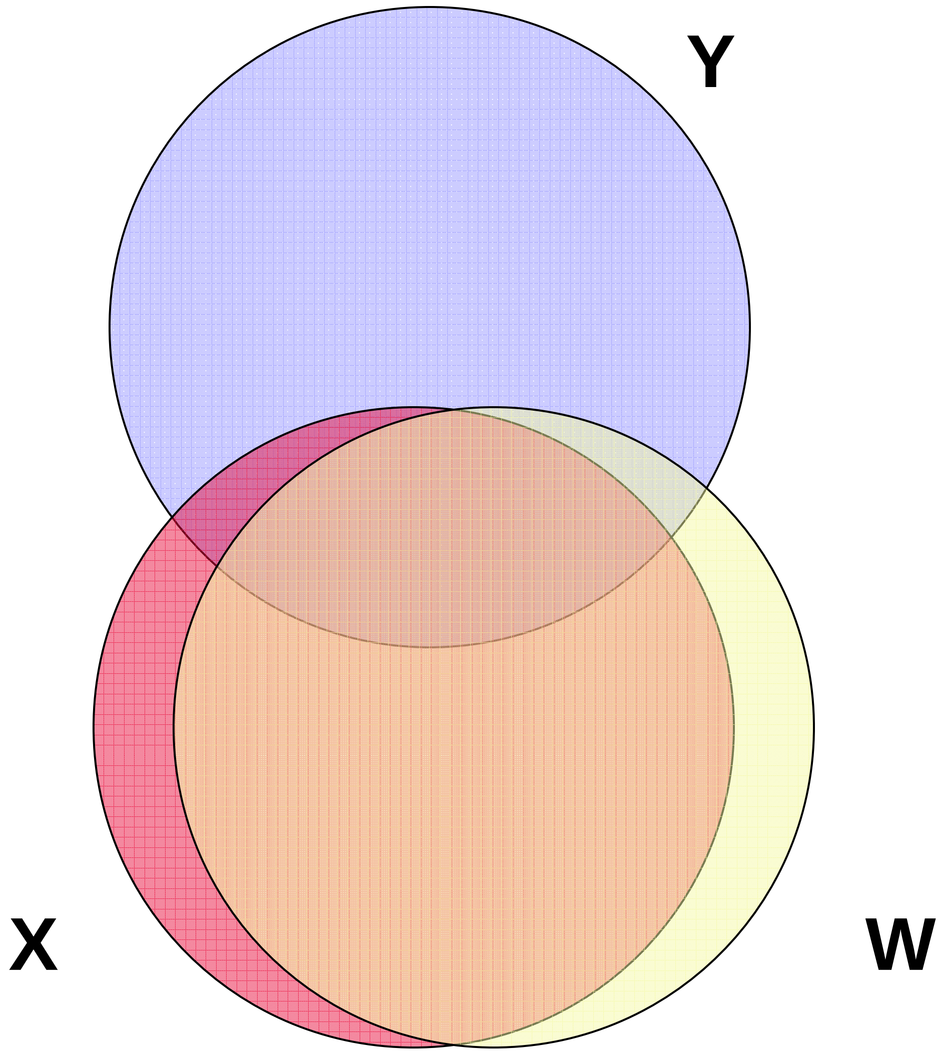
This information is NOT used in the estimation of β_1 nor β_2

Venn Diagrams and Collinearity



This is the usual situation: some overlap between regressors, but not too much.

Venn Diagrams and Collinearity

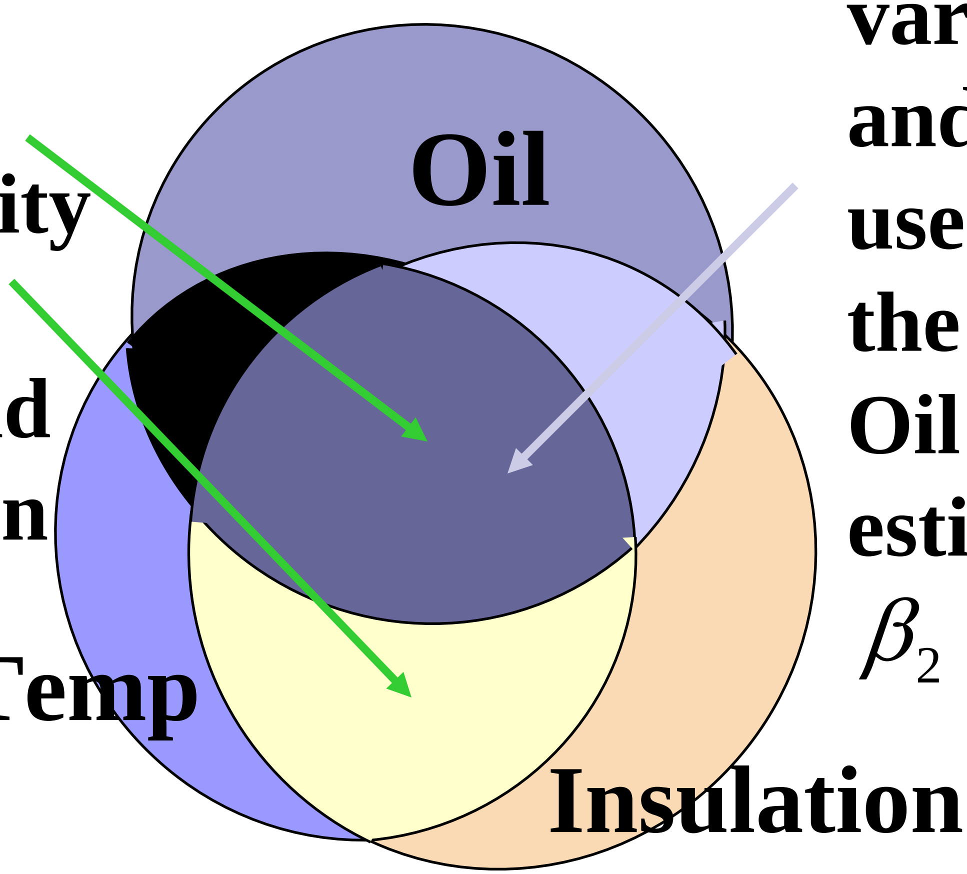


Now the overlap is so big, there's hardly any information left over to use when estimating β_x and β_w.

These variables "interfere" with each other.

Venn Diagrams and Collinearity

Large *Overlap* reflects collinearity between Temp and Insulation



Large *Overlap* in variation of Temp and Insulation is used in explaining the variation in Oil but *NOT* in estimating β_1 **and** β_2

Testing for Collinearity

“quietly” suppresses all output

```
. quietly regress api00 meals ell emer

. vif

    Variable |       VIF       1/VIF
-------------+----------------------
       meals |      2.73    0.366965
         ell |      2.51    0.398325
        emer |      1.41    0.706805
-------------+----------------------
    Mean VIF |      2.22
```

VIF = variance inflation factor
Any value over 10 is worrisome

Testing for Collinearity

“quietly” suppresses all output

```
. quietly regress api00 meals ell emer

. vif

    Variable |       VIF       1/VIF
-------------+----------------------
       meals |      2.73    0.366965
         ell |      2.51    0.398325
        emer |      1.41    0.706805
-------------+----------------------
    Mean VIF |      2.22
```

These results are not too bad

VIF = variance inflation factor
Any value over 10 is worrisome

Testing for Collinearity

Now add different regressors

```
. qui regress api00 acs_k3 avg_ed grad_sch col_grad some_col
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
      avg_ed |     43.57    0.022951
    grad_sch |     14.86    0.067274
    col_grad |     14.78    0.067664
    some_col |      4.07    0.245993
      acs_k3 |      1.03    0.971867
-------------+----------------------
    Mean VIF |     15.66
```

Testing for Collinearity

`Now add different regressors`

```
. qui regress api00 acs_k3 avg_ed grad_sch col_grad some_col
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
      avg_ed |     43.57    0.022951
    grad_sch |     14.86    0.067274
    col_grad |     14.78    0.067664
    some_col |      4.07    0.245993
      acs_k3 |      1.03    0.971867
-------------+----------------------
    Mean VIF |     15.66
```

Much worse.

Testing for Collinearity

Now add different regressors

```
. qui regress api00 acs_k3 avg_ed grad_sch col_grad some_col
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
      avg_ed |     43.57    0.022951
    grad_sch |     14.86    0.067274
    col_grad |     14.78    0.067664
    some_col |      4.07    0.245993
      acs_k3 |      1.03    0.971867
-------------+----------------------
    Mean VIF |     15.66
```

Much worse.

Problem: education variables are highly correlated

Testing for Collinearity

Now add different regressors

```
. qui regress api00 acs_k3 avg_ed grad_sch col_grad some_col
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
      avg_ed |     43.57    0.022951
    grad_sch |     14.86    0.067274
    col_grad |     14.78    0.067664
    some_col |      4.07    0.245993
      acs_k3 |      1.03    0.971867
-------------+----------------------
    Mean VIF |     15.66
```

Much worse.

Problem: education variables are highly correlated

Solution: delete collinear factors.

Testing for Collinearity

Delete average parent education

```
. qui regress api00 acs_k3 grad_sch col_grad some_col

. vif

    Variable |       VIF       1/VIF
-------------+----------------------
    col_grad |      1.28    0.782726
    grad_sch |      1.26    0.792131
    some_col |      1.03    0.966696
      acs_k3 |      1.02    0.976666
-------------+----------------------
    Mean VIF |      1.15
```

This solves the problem.

Measurement errors in x's

- Fact: least squares estimates are biased and inferences about

  $$\mu(y|x1, x2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

  can be misleading if the available data for estimating the regression are observations y, x_1, x_2*,where x_2* is an imprecise measurement of x_2 (even though it may be an unbiased measurement)
- This is an important problem to be aware of; general purpose solutions do not exist in standard statistical programs
- Exception: if the purpose of the regression is to predict future y's from future values of x_1 and x_2* then there is no need to worry about x_2* being a measurement of x_2